A Few Thousand Translations Go A Long Way!
Leveraging Pre-trained Models for African News Translation

David Ifeoluwa Adelani\textsuperscript{1*}, Jesujoba Oluwadara Alabi\textsuperscript{2*}, Angela Fan\textsuperscript{3*}, Julia Kreutzer\textsuperscript{4*}, Xiaoyu Shen\textsuperscript{5}, Machiel Reid\textsuperscript{6*}, Dana Ruiter\textsuperscript{1}, Dietrich Klakow\textsuperscript{1}, Peter Nabende\textsuperscript{7*}, Ernie Chang\textsuperscript{1*}, Tajuudeen R. Gwadabe\textsuperscript{8*}, Freslia Sackey\textsuperscript{9}, Bonaventure F. P. Dossou\textsuperscript{10*}, Chris Chinemye Emezie\textsuperscript{11*}, Colin Leong\textsuperscript{12*}, Michael Beukman\textsuperscript{13*}, Shamsuddeen H. Muhammad\textsuperscript{14*}, Guyo D. Jarso\textsuperscript{*}, Oreen Yousuf\textsuperscript{15*}, Andre N. Rubungo\textsuperscript{16*}, Gilles Hacheme\textsuperscript{17*}, Eric P. Wairagala\textsuperscript{7*}, Muhammad U. Nasir\textsuperscript{18*}, Benjamin A. Ajibade\textsuperscript{*}, Tunde Oluwaseyi Ajayi\textsuperscript{*}, Yonne Wambui Gitau\textsuperscript{*}, Jade Abbott\textsuperscript{*}, Mohamed Ahmed\textsuperscript{19*}, Millicent Ochieng\textsuperscript{19*}, Anuoluwapo Aremu\textsuperscript{*}, Perez Ogayo\textsuperscript{20*}, Jonathan Mukiibi\textsuperscript{2*}, Fatoumata Ouoba Kabore\textsuperscript{*}, Godson Koffi Kalipe\textsuperscript{*}, Derguenoe Mbaye\textsuperscript{21*}, Allahsere Auguste Tapo\textsuperscript{22*}, Victoire M. Koagne\textsuperscript{*}, Edwin Munkoh-Buabeng\textsuperscript{*}, Valencia Wagner\textsuperscript{23*}, Idris Abdulmumin\textsuperscript{24*}, Ayodele Awokoya\textsuperscript{25*}, Happy Buzaaba\textsuperscript{*}, Blessing Sibanda\textsuperscript{26*}, Andiswa Bukula\textsuperscript{27*}, Sam Manthulu\textsuperscript{28*}

\textsuperscript{*}Masakhane NLP, \textsuperscript{1}Saarland University, Germany, \textsuperscript{2}Inria, France, \textsuperscript{3}Meta AI, \textsuperscript{4}Google Research, \textsuperscript{5}Amazon Alexa AI, \textsuperscript{6}The University of Tokyo, Japan, \textsuperscript{7}Makerere University, Kampala, Uganda, \textsuperscript{8}UCAS, China, \textsuperscript{9}JUAT, Kenya, \textsuperscript{10} Jacobs University, Germany, \textsuperscript{11}TUM, Germany, \textsuperscript{12}University of Dayton, USA, \textsuperscript{13}University of the Witwatersrand, South Africa, \textsuperscript{14}LIAAD-INESC TEC, Portugal, Portugal, \textsuperscript{15}Uppsala University, Sweden, \textsuperscript{16}UPC, Spain, \textsuperscript{17}i4Innov\textsuperscript{18}Ominor AI, \textsuperscript{19}Microsoft Africa Research Institute, Kenya \textsuperscript{20}CMU, USA, \textsuperscript{21}Baamtu, \textsuperscript{22}RIT, USA, \textsuperscript{23}SPU, South Africa, \textsuperscript{24}ABU, Nigeria, \textsuperscript{25}UI Ibadan, Nigeria, \textsuperscript{26}NUST, Namibia, \textsuperscript{27}SADiLaR, South Africa, \textsuperscript{28}University 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 monolingual and parallel data from the web (El-Kishky et al., 2020; Schwenk et al., 2021a; 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; \textsuperscript{∀} et al., 2020) and further have little to no training data available (§3). These languages provide an ideal testbed for two challenging knowledge 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).

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; Emuze and Dossou, 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; Ezeeani 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.,

\[\text{All data, models and code are publicly available on } \text{https://github.com/masakhane-io/lafand-mt under academic license.}\]
languages: Afro-Asiatic (e.g. Hausa), Nilo-Saharan (e.g. Luo), English Creole (e.g. Nigerian, English), and agglutinative. 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 (Agić 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. Éwé, 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

2Spoken by an estimated 1.1M people in Cameroon

\[ \text{languages are tonal, and about nine make use of diacritics. Many African languages are morphologically rich. For example, all Bantu languages are agglutinative. Fon, Mossi, and Yorùbá are highly isolating. All languages follow the Subject-Verb-Object sentence structure like English and French. Table C provides more details.} \]

### 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 Ghomálá (e.g. Hausa), Nilo-Saharan (e.g. Luo), English Creole (e.g. Nigerian, Pidgin/Naïja) 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. “ç”, “ð”, “ñ”, “ž”) and additions (e.g. “é”, “è”, “ì”, “ò”, “ù”, “ñ”, “ú”, “ü”, “ç”, “é”, “í”, “ó”, “ú”) including digraphs like “gh”, “kp”, “gb”, 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 agglutinative. 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 (Agić 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. Éwé, 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

2Spoken by an estimated 1.1M people in Cameroon

\[ \text{languages are tonal, and about nine make use of diacritics. Many African languages are morphologically rich. For example, all Bantu languages are agglutinative. Fon, Mossi, and Yorùbá are highly isolating. All languages follow the Subject-Verb-Object sentence structure like English and French. Table C provides more details.} \]

### 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 Ghomálá (e.g. Hausa), Nilo-Saharan (e.g. Luo), English Creole (e.g. Nigerian, Pidgin/Naïja) 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. “ç”, “ð”, “ñ”, “ž”) and additions (e.g. “é”, “è”, “ì”, “ò”, “ù”, “ñ”, “ú”, “ü”, “ç”, “é”, “í”, “ó”, “ú”) including digraphs like “gh”, “kp”, “gb”, 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 agglutinative. 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 (Agić 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. Éwé, 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

2Spoken by an estimated 1.1M people in Cameroon

\[ \text{languages are tonal, and about nine make use of diacritics. Many African languages are morphologically rich. For example, all Bantu languages are agglutinative. Fon, Mossi, and Yorùbá are highly isolating. All languages follow the Subject-Verb-Object sentence structure like English and French. Table C provides more details.} \]
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). Ghomalá 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: Hausa, Igbo (Ezeani et al., 2020), Swahili, Yoruba (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 corpus (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

---

3Some languages like Luo and Luganda are covered by JW300 but are no longer available at the time of paper writing.
4Crawled/downloaded from https://ebible.org/, except for Bambara that we obtained from https://live.bible.is/ and Ghomalá from www.beblia.com

5https://www.statmt.org/wmt21/translation-task.html
6https://sw.globalvoices.org/
We describe two methods for adding new languages under study are included. To adapt the existing PLMs to our languages corpora for each language.

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

| Pre-trained Model (PM) | PM Size | # African Lang. | Focus languages covered |
|------------------------|---------|-----------------|------------------------|
| MT5/ByT5               | 580M    | 13              | hau, ibo, swa, yor, zul |
| Afri[*T5]              | 580M    | 17              | hau, ibo, pcm, swa, yor, zul |
| mBART50                | 610M    | 2               | swa                     |
| AfriMBART              | 610M    | 17              | hau, ibo, pcm, swa, yor, zul |
| M2M-100                | 418M    | 17              | hau, ibo, lug, swa, tsn, wol, yor, zul |

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

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.

### 5.1 Adaptation to the Focus Languages

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 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 pre-train to specialize models to African languages further improves performance.

---

8 Changing 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.

---

7 https://github.com/google/sentencepiece
Table 3: Results adding African Languages to Pre-Trained Models, en/fr-xx. We calculate BLEU and CHRFF on the news domain when training on only NEWS data from MAFAND-MT.

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 bitemporal text during pre-training (2.4M sentences in CCAigned (El-Kishky et al., 2020)). Other African languages except for Afrikaans have less than 600K sentences in CCAigned, and are also of a lower quality (Kreutz 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, mBART50 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/xx) 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 bbj, 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/ff 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/ff).

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-bbm improved by 1.9 BLEU over ByT5 and AfriMT5 on en-tsn improved by 3.6 BLEU over MT5. On average, AfriMT5 im-
proven over MT5 by 1.3 BLEU in en/fr-xx and 2.4 BLEU in the xx-en direction. The improvement for AfriByT5 was much smaller: 0.6 and 0.9 BLEU in en/fr-xx and xx-en 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 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→REL. In general, the improvement depends on the size of REL corpus. For languages trained on the Bible such as bbj, bam, lug, luo, 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 domain.
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-ibo) and relatively small (eng-swa) 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-
Chichewa, Kinyarwanda, Shona, and isiXhosa, in-cluding an expansion of the Hausa corpus, they will be released under MAFAND-MT dataset name. Except for Luo which is not supported.

Table 8: spBLEU on Wikipedia domain (FLORES) and REL for M2M-100 before (✗) and after (✓) fine-tuning on NEWS.

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: COMPRIScE (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

David Adelani, Dana Ruiter, Jesujoba Alabi, Damilola Adebonojo, Adesina Ayeni, Mofe Adeyemi, Ayo-dele Esther Awokoya, and Cristina España-Bonet. 2021a. The effect of domain and diacritics in Yoruba–English neural machine translation. In Proceedings of Machine Translation Summit XVIII: Research Track, pages 61–75, Virtual. Association for Machine Translation in the Americas.

David Ifeoluwa Adelani, Jade Abbott, Graham Newbig, Daniel D’souza, Julia Kreutzer, Constantine Lignos, Chester Palen-Michel, Happy Buzzaa, Shruti Rijhwani, Sebastian Ruder, Stephen Mayhew, Israel Abebe Azime, Shamsuudeen H. Muhammad, Chris Chinenye Emezu, Joyce Nakambata-Nabende, Perez Ogayo, Aremu Anuoluwapo, Catherine GITau, Derguene Mbaye, Jesujoba Alabi, Seid Muhie Yimam, Tajuudeen Rabiu Gwadabe, Ignatius Ezeani, Rubungo Andre Niyongabo, Jonathan Mukibi, Ver-rah Otieno, Iroko Orife, Davis David, Samba Ngom, Tosin Adewumi, Paul Rayson, Mofetoluwada Adeyemi, Gerald Muriuki, Emmanuel Anebi, Chiamaka Chukwuneka, Nkiru Odu, Eric Wairagala, Samuel Oyerinde, Clemencia Siro, Tobius Saul Bateesa, Temilola Oloyede, Yvonne Wambui, Victor Akinode, Deborah Nabagereka, Maurice Katusiime, Ayodele Awokoya, Mouhamadou MBoup, Dibora Georgyohannes, Henok Tilye, Kelechi Nwaike, De-gaga Wolde, Abdoulaye Faye, Blessing Sibanda, Ore-vaaghene Ahia, Bonaventure F. P. Dossou, Kelechi Ogueji, Thiero Ibrahim DIOP, Abdoulaye Diyallo, Adewale Akinfaderin, Tendai Marengerere, and Salomey Osei. 2021b. MasakhaNER: Named entity recognition for African languages. Transactions of the Association for Computational Linguistics, 9:1116–1131.

Željko Agić and Ivan Vulić. 2019. JW300: A wide-coverage parallel corpus for low-resource languages.

In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 3204–3210, Florence, Italy. Association for Computational Linguistics.

Rooe Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively multilingual neural machine translation. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 3874–3884, Minneapolis, Minnesota. Association for Computational Linguistics.

Felermino M. D. A. Ali, Andrew Caines, and Jaimi-tro L. A. Malavi. 2021. Towards a parallel corpus of portuguese and the bantu language emakhuwa of mozambique. ArXiv, abs/2104.05753.

Antonios Anastasopoulos, Alessandro Cattelan, Zi-Yi Dou, Marcello Federico, Christian Federmann, Dmitriy Genzel, Francisco Guzmán, Junjie Hu, Mac-duff Hughes, Philipp Koehn, Rosie Lazar, Will Lewis, Graham Newbig, Mengmeng Niu, Alp Öktem, Eric Paquin, Grace Tang, and Sylvia Tur. 2020. TICO-19: the translation initiative for COVID-19. In Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020, Online. Association for Computational Linguistics.

Paul Azunre, Salomey Osei, Salomey Addo, Lawrence Asamoah Adu-Gyamfi, Stephen Moore, Bernard Adabankah, Bernard Opoku, Clara Asare-Nyarko, Samuel Nyarko, Cynthia Amoaba, Esther Dansoo Appiah, Felix Akwer, Richard Nii Lante Lawson, Joel Budu, Emmanuel Debrah, Nana Boateng, Wisdom Ofori, Edwin Buabeng-Munkoh, Franklin Adjei, Isaac Kojo Essel Ampomah, Joseph Otoo, Reindorf Borkor, Standylove Birago Mensah, Lucien Mensah, Mark Amoako Marcel, Anoyke Acheampong Amponsah, and James Ben Hayfron-Acquah. 2021a. Nlp for Ghanaian languages. AfricaNLP Workshop, abs/2103.15475.

Paul Azunre, Salomey Osei, Salomey Addo, Lawrence Asamoah Adu-Gyamfi, Stephen Moore, Bernard Adabankah, Bernard Opoku, Clara Asare-Nyarko, Samuel Nyarko, Cynthia Amoaba, Esther Dansoo Appiah, Felix Akwer, Richard Nii Lante Lawson, Joel Budu, Emmanuel Debrah, Nana Adowaa Boateng, Wisdom Ofori, Edwin Buabeng-Munkoh, Franklin Adjei, Isaac Kojo Essel Ampomah, Joseph Otoo, Reindorf Borkor, Standylove Birago Mensah, Lucien Mensah, Mark Amoako Marcel, Anoyke Acheampong Amponsah, and James B. Hayfron-Acquah. 2021b. English-twi parallel corpus for machine translation. ArXiv, abs/2103.15625.

Alexandra Birch, Barry Haddow, Antonio Valerio Miceli Barone, Jindrich Helcl, Jonas Waldendorf, Felipe Sánchez Martínez, Mikel Forcada, Víctor Sánchez Cartagena, Juan Antonio Pérez-Ortiz, Miquel Esplà-Gomis, Wilker Aziz, Lina Murady,
Sevi Sariisik, Peggy van der Kreeft, and Kay Macquarrie. 2021. Surprise language challenge: Developing a neural machine translation system between Pashto and English in two months. In Proceedings of Machine Translation Summit XVIII: Research Track, pages 92–102, Virtual. Association for Machine Translation in the Americas.

Steven Bird. 2020. Decolonising speech and language technology. In Proceedings of the 28th International Conference on Computational Linguistics, pages 3504–3519, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Samuel Cahyawijaya, Genta Indra Winata, Bryan Wilie, Karissa Vincentio, Xiaohong Li, Adhiguna Kunoro, Sebastian Ruder, Zhi Yuan Lim, Syafri Bahar, Masayu Khodra, Ayu Purwarianti, and Pascale Fung. 2021. IndoNLP: Benchmark and resources for evaluating Indonesian natural language generation. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 8875–8898, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Josh Schroeder, and Cameron Shaw Fordyce, editors. 2008. Proceedings of the Third Workshop on Statistical Machine Translation. Association for Computational Linguistics, Columbus, Ohio.

Ahmed El-Kishky, Vishrav Chaudhary, Francisco Guzmán, and Philipp Koehn. 2020. CCAligned: A massive collection of cross-lingual web-document pairs. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020), pages 5960–5969, Online. Association for Computational Linguistics.

Chris Chinenyeye Emezue and Bonaventure F. P. Dossou. 2021. MMTAfrica: Multilingual machine translation for African languages. In Proceedings of the Sixth Conference on Machine Translation, pages 398–411, Online. Association for Computational Linguistics.

Chris Chinenyeye Emezue and Femi Pancrace Bonaventure Dossou. 2020. FFR v1.1: Fon-French neural machine translation. In Proceedings of the The Fourth Widening Natural Language Processing Workshop, pages 83–87, Seattle, USA. Association for Computational Linguistics.

Ignatius Ezeani, Paul Rayson, I. Onyenwe, C. Uchechukwu, and M. Hepple. 2020. Igbo-english machine translation: An evaluation benchmark. ArXiv, abs/2004.00648.

Angela Fan, Shrutih Bhasale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Michael Auli, and Armand Joulin. 2021a. Beyond english-centric multilingual machine translation. Journal of Machine Learning Research, 22(107):1–48.

Angela Fan, Shrutih Bhasale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, et al. 2021b. Beyond english-centric multilingual machine translation. Journal of Machine Learning Research, 22(107):1–48.

Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. 2016. Multi-way, multilingual neural machine translation with a shared attention mechanism. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 866–875, San Diego, California. Association for Computational Linguistics.

∀ Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Taiwo Fagbogbunge, Solomon Oluwole Akinola, Shamsudeen Muhammad, Salomon Kabongo Kabenamuula, Salomey Osei, Freshia Sackey, Rubungo Andre Niyongabo, Ricky Macharm, Perez Ogayo, Orevaoghena Ahia, Music Meressa Berhe, Mofetoluwa Adeyemi, Masabata Mokgesi-Selinga, Lawrence Okegbemi, Laura Martinus, Kolawole Tajudeen, Kevin Degila, Kelechi Oguej, Kathleen Siminyu, Julia Kreutzer, Jason Webster, Jamil Toure Ali, Jade Abbott, Iroro Orife, Ignatius Ezeani, Idris Abdulkadir Dangana, Herman Kamper, Hady Elsahar, Goodness Duru, Ghollah Kioko, Murhabazi Espoir, Elan van Biljon, Daniel Whiteenack, Christopher Onyefuluchi, Chris Chinenyeye Emezue, Bonaventure F. P. Dossou, Blessing Sibanda, Blessing Bassey, Ayodele Olabiyi, Arshath Ramkilowan, Alp Öktem, Adewale Akinfaderin, and Abdallah Bashir. 2020. Participatory research for low-resourced machine translation: A case study in African languages. In Findings of the Association for Computational Linguistics: EMNLP 2020, Online.

Andargachew Mekonnen Gezmu, A. Nürnberger, and Tesfaye Bayu Bati. 2021. Extended parallel corpus for amharic-english machine translation. ArXiv, abs/2104.03543.

Thamme Gowda, Zhao Zhang, Chris Mattmann, and Jonathan May. 2021. Many-to-English machine translation tools, data, and pretrained models. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations, pages 306–316, Online. Association for Computational Linguistics.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjan Krishnan, Marc’Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2021. The flores-101 evaluation benchmark for low-resource and multilingual machine translation. ArXiv, abs/2106.03193.

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
Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 8342–8360.

Barry Haddow, Rachel Bawden, Antonio Valerio Miceli Barone, Jindrich Helcl, and Alexandra Birch. 2021. Survey of low-resource machine translation. ArXiv, abs/2109.00486.

Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, Shujie Liu, Tie-Yan Liu, Renqian Luo, Arul Menezes, Tao Qin, Frank Seide, Xu Tan, Fei Tian, Lijun Wu, Shuangzhi Wu, Yingce Xia, Dongdong Zhang, Zhuirui Zhang, and Ming Zhou. 2018. Achieving human parity on automatic chinese to english news translation. CoRR, abs/1803.05567.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernando Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google’s multilingual neural machine translation system: Enabling zero-shot translation. Transactions of the Association for Computational Linguistics, 5:339–351.

Pratik Joshi, Sebasstin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 6282–6293, Online. Association for Computational Linguistics.

Wei-Jen Ko, Ahmed El-Kishky, Adithya Renduchintala, Vishrav Chaudhary, Naman Goyal, Francisco Guzmán, Pascale Fung, Philipp Koehn, and Mona Diab. 2021. Adapting high-resource NMT models to translate low-resource related languages without parallel data. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 802–812, Online. Association for Computational Linguistics.

Julia Kreutzer, Jasmin Bastings, and Stefan Riezler. 2019. Joey NMT: A minimalist NMT toolkit for novices. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations, pages 109–114, Hong Kong, China. Association for Computational Linguistics.

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulziig-Orshikh, Alahsara Auguste Tape, Nishant Subramani, Artem Sokolov, Claytong Sikasote, Monang Setyawana, Supheakmunkol Sarin, Sokhamb Samb, Benoît Sagot, Clara Rivera, Annette Rios Gonzales, Isabel Padidimitriou, Salomey Osei, Pedro Ortiz Suarez, Iroro Orife, Kelechi Ogueji, Rubungo Andre Niyongabo, Toan Q. Nguyen, Mathias Muller, André Muller, Shamsuddeen Hassan Muhammad, Nanda Muhammad, Ayanda Mnyakeni, Jamshedek Mirzakhahov, Tapiwanashe Matangira, Colin Leong, Nze Lawson, Sneha Kadugunta, Yacine Jernite, M. Jenny, Orhan Firat, Bonaventure F. F. Dossou, Sakhiile Dlamini, Nansana de Silva, Sakine cCebuk Balli, Stella Rose Biderman, Alessia Battisti, Ahmed Baruwa, Ankur Bapna, Pallavi N. Baljekar, Israel Abebe Azime, Ayodele Awokoya, Duygu Ataman, Oreevaghene Ahia, Oghenefego Ahia, Ñwetá Agrawal, and Mofetoluwa Adeyemi. 2021. Quality at a glance: An audit of web-crawled multilingual datasets. ArXiv, abs/2103.12028.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers, pages 66–75, Association for Computational Linguistics.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc’Aurelio Ranzato. 2018. Unsupervised machine translation using monolingual corpora only. In International Conference on Learning Representations.

Samuel Läubli, Rico Sennrich, and Martin Volk. 2018. Has machine translation achieved human parity? a case for document-level evaluation. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 4791–4796, Brussels, Belgium. Association for Computational Linguistics.

En-Shiun Annie Lee, Sarubi Thillainathan, Shrvan Nayak, Surangika Ranathunga, David Ifeoluwa Adelani, Ruisi Su, and Arya D. McCarthy. 2022. Pre-trained multilingual sequence-to-sequence models: A hope for low-resource language translation? In Findings of ACL 2022, abs/2203.08850.

Zihan Liu, Genta Indra Winata, and Pascale Fung. 2021. Continual mixed-language pre-training for extremely low-resource neural machine translation. In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 2706–2718, Online. Association for Computational Linguistics.

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

Lovish Madaan, Soumya Sharma, and Parag Singla. 2020. Transfer learning for related languages: Submissions to the WMT20 similar language translation task. In Proceedings of the Fifth Conference on Machine Translation, pages 402–408, Online. Association for Computational Linguistics.

Manuel Mager, Arturo Oncevay, Abteen Ebrahimi, John Ortega, Annette Rios, Angela Fan, Ximena Gutierrez-Vasques, Luis Chiruzzo, Gustavo Gimenez-Lugo, Ricardo Ramos, Ivan Vladimir Meza Ruiz, Rolando
Coto-Solano, Alexis Palmer, Elisabeth Mager-Hois, Vishrav Chaudhary, Graham Neubig, Ngoc Thang Vu, and Katharina Kann. 2021. Findings of the AmericasNLP 2021 shared task on open machine translation for indigenous languages of the Americas. In Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas, pages 202–217, Online. Association for Computational Linguistics.

Kelly Marchisio, Kevin Duh, and Philipp Koehn. 2020. When does unsupervised machine translation work? In Proceedings of the Fifth Conference on Machine Translation, pages 571–583, Online. Association for Computational Linguistics.

Arya D. McCarthy, Rachel Wicks, Dylan Lewis, Aaron Mueller, Winston Wu, Oliver Adams, Garrett Nicolai, Matt Post, and David Yarowsky. 2020. The Johns Hopkins University Bible corpus: 1600+ tongues for typological exploration. In Proceedings of the 12th Language Resources and Evaluation Conference, pages 2884–2892, Marseille, France. European Language Resources Association.

Rubungo Andre Niyongabo, Qu Hong, Julia Kreutzer, and Li Huang. 2020. KINNEWS and KIRNEWS: Benchmarking cross-lingual text classification for Kinyarwanda and Kirundi. In Proceedings of the 28th International Conference on Computational Linguistics, pages 5507–5521, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Evander Nyoni and Bruce A. Bassett. 2021. Low-resource neural machine translation for southern african languages. ArXiv, abs/2104.00366.

Kelechi Ogueji, Yuxin Zhu, and Jimmy Lin. 2021. Small data? no problem! exploring the viability of pretrained multilingual language models for low-resourced languages. In Proceedings of the 1st Workshop on Multilingual Representation Learning, pages 116–126, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Alp Öktem, Eric DeLuca, Rodrigue Bashizi, Eric Paquin, and Grace Tang. 2021. Congolese swahili machine translation for humanitarian response. AfricaNLP Workshop.

Alp Öktem, Mirok Pllit, and Grace Tang. 2020. Tigrinya neural machine translation with transfer learning for humanitarian response. AfricaNLP Workshop.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations), pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Amandalynne Paullada. 2020. How does machine translation shift power? Resistance in AI Workshop.

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, Lisbon, Portugal. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In Proceedings of the Third Conference on Machine Translation: Research Papers, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Machel Reid, Junjie Hu, Graham Neubig, and Yutaka Matsuo. 2021. AfroMT: Pretraining strategies and reproducible benchmarks for translation of 8 African languages. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 1306–1320, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2021a. WikiMatrix: Mining 135M parallel sentences in 1620 language pairs from Wikipedia. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, pages 1351–1361, Online. Association for Computational Linguistics.

Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, Armand Joulin, and Angela Fan. 2021b. CCMatrix: Mining billions of high-quality parallel sentences on the web. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 6490–6500, Online. Association for Computational Linguistics.

Kathleen Siminyu, Godson Kalipe, Davor Orlic, Jade Z. Abbott, Yukosi Marivate, Sackey Freshia, Prateek Sibal, Bhanu Neupane, David Iteoluwa Adelani, Amelia Taylor, Jamil Toure Ali, Kevin Degila, Momboladj Balogoun, Thierno Ibrahima Diop, Davis David, Chayma Fourati, Hutm Haddad, and Malek Naski. 2021. A4d - african language program. ArXiv, abs/2104.02516.

Y. Tang, C. Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning. ArXiv, abs/2008.00401.
B Available Parallel Corpora

We found Five African languages with publicly available parallel texts in the news domain: Hausa, Igbo, Swahili, Yorùbá, and isiZulu. Table 1 provides news source, the TRAIN, DEV and TEST splits.

Hausa The Hausa Khamenei\(^{11}\) 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\(^{12}\) 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,
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.

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 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.
Model Name | HuggingFace Model name | Remark
--- | --- | ---
AfriMT5 | masakhane/afri-mt5-base | mT5-base adaptation to 17 African languages, English, French and Arabic.
AfriByT5 | masakhane/afri-byt5-base | ByT5-base adaptation to 17 African languages, English, French and Arabic.
AfriMBART | masakhane/afri-mbart50 | mBART50 adaptation to 17 African languages, English, French and Arabic.

MT5 fine-tuned on (src)-[tgt] direction using parallel NEWS corpus.
AfriMT5 fine-tuned on (src)-[tgt] direction using parallel NEWS corpus.
AfriByT5 fine-tuned on (src)-[tgt] direction using parallel NEWS corpus.
AfriMBART fine-tuned on (src)-[tgt] direction using parallel NEWS corpus.

Table 11: Model names on HuggingFace Model Hub. For bilingual models, supply the correct translations of the test set illustrate the advantage generally better. On the other hand, in the evaluation are still needed to show that spBLEU is BLEU especially in the direction of BLEU metric for the domain transfer experiments. 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 AfriMT5, AfriByT5, and AfriMBART).

E BLEU vs spBLEU

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.

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 Alimanyidoko kpôdôssa Etchelkoun kpô ô, sôn azàn mokpàn dyê Ô, ye dô wôlv ñe we twauñ dô agbaza më, có ye kà tuun fî é azon nê lër gozin ê ô à.

- **Reference** (fr): Les faits Louis Guy Alimagniidoko et Issa Etchelkoun ont depuis quelques jours de multiples douleurs, ignorant l’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, fr→fon): Sûn azàn ñ’ywywëwe ñê dyê fôkôwë wë nû è kan Louis Guy Alimagniidoko kpôdô Issa Etchelkoun kàñmê këp we è ñe dô xô Ø wë dô dô wëlvë gege wë, ye kà tuun nû è wu wûwë yeton ñê ô à.

The translation of the bilingual Transformer model using Fairseq (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 13. 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.
is very poor and far from the Fon source, highlighting how poorly the model generalized from the few thousand training sentences. The M2M-100 model gives a more meaningful and adequate translation. M2M-100 makes a surprising but beautiful move, switching se plaignent depuis quelques jours de multiples douleurs (sin az´an m˚kpan dye ð, ye ð vôw s´e w´e tawin ð o agbaza mr) to ont depuis plusieurs jours souffert d’une maladie grave. The BLEU score here might be low but the meaning is conserved and even more detailed than the French reference. In fact, in this source context, vôw means souffrir, souffrance (suffer, suffering); the French reference made use of se plaignent (complaining) which makes less sense than souffert used in the M2M-100 prediction. M2M-100 also learned the style of the sentence: cố ye ká tuun fi é az˘an nr ler gosin (but they do know the origin of their sufferings) é ða (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 malgr´e 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 Etch-lekon ont depuis plusieurs jours souffert d’une maladie grave malgr´e (dont) ils ne connaissent pas les cons´equences (causes/raisons) de cette maladie qu’ils ne connaissent pas."

In the opposite translation direction, fr→fon, M2M-100 (REL+NEWS→NEWS) still preserved some sense of logical reasoning and predicted the last part right ye ká tuun nˇu ò wˇu vôw õ yet˘an (they do know why they are suffering) dˇe ða (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 sin az˘an + number means since xxx days but yrywfr is not a number, and do not have any meaning in this context.

**G Limitations and Risks**

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 resourcedness, 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.
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.