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

Anonymous ACL submission

Abstract

Recent advances in the pre-training of language models leverages 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 additional languages and additional domains is to leverage small quantities of high-quality translation data to fine-tune large pre-trained models.

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., 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 knowledge transfer tasks: **(1)** How can pre-trained models create translations for languages unseen at training time? and **(2)** Since training data may exist only in single domain (i.e. religious texts), so 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 all have millions of native speakers and a massive need for translation technologies. For example, in most of Sub-Saharan Africa news concerning the African continent, they are almost exclusively published in English, French, or Arabic, and thereby inaccessible for speakers of only native 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). Further, 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 these 16 native

African languages and English or French (§4).

This allows us to compare three approaches to leveraging large-scale multilingual models for translating 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 high quality bitext is remarkably effective, and further augment with continual pre-training on African languages and fine-tuning on news domain data (§6). Our contributions are the following:[1]

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 measuring 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 Dossou, 2020; Azunre et al., 2021; 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., 2021; Mabuya et al., 2021; Ezeani et al., 2020) but ours covers 11 African languages (see §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; 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., 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 models often outperforms bilingual models (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., 2021). 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

[1] All data, models and code will be made publicly available upon publication of this paper.

Target Language	Family	African Region	No. of Speakers	Source Lang.	NEWS Source	NEWS Split Sizes	REL Source	REL Total Size
Bambara (bam)	NC / Manding	West	14M	French	Maliweb.net	3302/ 1484/ 1600	Bible	28K
Ghomálá' (bbj)	NC / Grassfields	Central	1M	French	Cameroun Web	2232/ 1133/ 1430	Bible	8K
Éwé (ewe)	NC / Kwa	West	7M	French	Benin Web TV	2026/ 1414/ 1563	JW300	618K
Fon (fon)	NC / Volta-Niger	West	2M	French	ORTB, Nation, Héraut, Matin Libre, LB Libéré, LE Précis, Visages.	2637/ 1227/ 1579	JW300	32K
Hausa (hau)	Afro-Asiatic / Chadic	West	63M	English	WMT2021: Khamenei.v1	3098/ 1300/ 1500	JW300	236K
Igbo (ibo)	NC / Volta-Niger	West	27M	English	(Ezeani et al., 2020)	6998/ 1500/ 1500	JW300	415K
Luganda (lug)	NC / Bantu	East	7M	English	Independent Uganda	4075/ 1500/ 1500	Bible	31K
Luo (luo)	Nilo-Saharan	East	4M	English	Lolwe, Standard Media	4262/ 1500/ 1500	Bible	31K
Mossi (mos)	NC / Gur	West	8M	French	Burkina24, Lefaso	2287/ 1478/ 1574	JW300	216K
Naija (pcm)	English-Creole	West	75M	English	Daily Trust Nigeria	4790/ 1484/ 1564	JW300	23K
Swahili (swa)	NC / Bantu	East & Central	98M	English	Global Voices, OPUS	30782/ 1791/ 1835	JW300	872K
Setswana (tsn)	NC / Bantu	South	14M	English	SABC News	2100/ 1340/ 1500	JW300	870K
Akan/Twi (twi)	NC / Kwa	West	9M	English	StarrFM, Citi News	3337/ 1284/ 1500	JW300	601K
Wolof (wol)	NC / Senegambia	West	5M	French	Seneweb, Jotna, Yerim Post, Socialnetlink	3360/ 1506/ 1500	Bible	22K
Yorùbá (yor)	NC / Volta-Niger	West	42M	English	(Adelani et al., 2021)	5253/ 1391/ 3102	JW300	460K
isiZulu (zul)	NC / Bantu	South	27M	English	(Mabuya et al., 2021)	3500/ 1239/ 998	JW300	667K

Table 1: **Languages and Data Details for FAAND-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 FAAND-MT.

data (Joshi et al., 2020), including moderately low-resource languages such as Swahili and Hausa, and very low-resource languages such as Ghomálá' [2] 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. "ɛ", "ɔ", "ŋ", "ọ", including 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 agglutinative. Fon, Mossi, and Yorùbá are highly isolating. All languages follow the Subject-Verb-Object sentence structure like English and French. Appendix A provides more details.

Existing Parallel Corpora. We curate publicly available parallel data for our focus languages, which consists primarily of religious domain text (REL). 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 Naija have around 30K sentences. For the remaining five languages that are not in the JW300 corpus,[3] we make use of the Bible.[4] We aligned the sentences automatically by the verses (around 31k in total). Ghomálá' 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 FAAND-MT African News Corpus

4.1 Data Collection Process

We introduce our newly translated news corpus; FAAND-MT — **F**ranco & **A**nglo **A**frica **N**ews **D**ataset for **M**achine **T**ranslation. Table 1 gives the news source and data splits for 11 African languages which includes six languages (bam, bbj, ewe, fon, mos, and wol) spoken predominantly

[2] Spoken by an estimated 1.1M people in Cameroon

[3] Some languages like Luo and Luganda are covered by JW300 but are no longer available at the time of paper writing.

[4] Crawled/downloaded from https://ebible.org/, except for Bambara that we obtained from https://live.bible.is/ and Ghomálá' from www.beblia.com

in Francophone Africa and five languages (lug, luo, pcm, tsn, and twi) spoken predominantly in Anglophone Africa. The FAAND-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 of the languages were crawled from either 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. Duplicates and missing translations were removed.

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 publicly available parallel texts in the news domain: Hausa[5], Igbo (Ezeani et al., 2020), Swahili[6], Yorùbá (Adelani et al., 2021), and isiZulu (Mabuya et al., 2021). Table 1 provides news source, the TRAIN, DEV and TEST splits. Appendix B provides more details on the preprocessing of the available news corpora.

[5]https://www.statmt.org/wmt21/translation-task.html

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.

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 the more detail about 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 on pre-trained models that are approximately 500M 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[7] 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.

[6]https://sw.globalvoices.org/

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

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.

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[8] 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 AfriBART.

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)[9] and ChrF (Popović, 2015) in Appendix E.

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

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 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

[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.

[9] "intl" tokenizer, all data comes untokenized.

Model	*fr-xx* bam	bbj	ewe	fon	mos	wol	*en-xx* hau	ibo	lug	luo	pcm	swa	tsn	twi	yor	zul	AVG	MED
M2M-100 0-shot	–	–	–	–	–	1.3	0.4	2.6	–	–	–	20.0	1.0	–	1.9	4.3	–	
MT5	0.9	0.7	1.8	1.1	0.3	1.1	2.4	14.1	3.5	3.2	33.5	23.2	3.3	1.6	2.2	8.8	6.3	2.3
AfriMT5	1.9	0.9	3.1	1.6	0.3	1.8	4.5	15.4	5.9	4.5	**34.5**	26.7	6.9	2.5	4.7	9.8	7.8	4.5
ByT5	8.7	1.6	4.4	2.3	0.4	5.7	8.8	18.6	11.3	8.8	32.4	26.6	15.5	6.2	6.2	12.2	10.6	8.8
AfriByT5	10.6	1.9	4.2	2.3	0.7	6.2	9.8	19.3	12.2	9.0	32.4	**27.5**	18.0	6.3	7.1	13.4	11.3	9.4
mBART50	15.8	2.7	3.8	4.7	2.7	8.7	11.8	14.8	9.7	9.6	33.9	22.1	17.2	7.3	7.5	**17.3**	11.9	9.7
AfriMBART	13.1	2.3	4.7	3.3	2.1	7.9	9.5	18.1	8.9	9.3	29.5	25.9	12.8	6.1	7.9	16.6	11.1	9.1
M2M-100	**20.6**	**3.5**	**5.9**	**5.6**	**3.3**	**11.1**	**14.4**	**20.3**	**13.0**	**10.8**	33.2	27.0	**24.8**	**8.5**	**9.6**	16.5	**14.3**	**12.1**
M2M-100-EN/FR	16.1	2.6	5.8	3.3	2.7	9.6	6.5	18.0	8.6	9.6	34.4	26.3	19.5	6.6	6.8	10.9	11.7	9.1

Table 3: **Results adding African Languages to Pre-Trained Models, en/fr-xx**. We calculate BLEU on the news domain when training on only NEWS data from FAAND-MT.

Model	*xx-fr* bam	bbj	ewe	fon	mos	wol	*xx-en* hau	ibo	lug	luo	pcm	swa	tsn	twi	yor	zul	AVG	MED
M2M-100 0-shot	–	–	–	–	–	0.9	2.4	6.3	–	–	–	25.3	3.1	–	2.7	13.7	–	
MT5	2.8	1.0	1.3	2.3	1.2	1.4	5.9	18.0	11.5	6.7	42.2	29.0	9.4	4.2	7.9	22.2	10.4	6.3
AfriMT5	5.8	2.3	2.4	4.0	2.5	2.8	10.7	19.1	14.8	9.4	44.7	**30.7**	15.9	8.1	11.5	23.4	13.0	10.1
ByT5	9.6	2.6	4.1	4.0	2.7	6.8	14.0	**20.8**	19.3	11.9	43.4	28.8	19.0	10.5	9.6	25.2	14.5	11.2
AfriByT5	13.1	4.2	4.5	5.2	4.6	8.5	14.7	20.5	**20.6**	12.4	43.4	29.0	**20.2**	11.2	10.4	26.5	15.6	12.8
mBART50	12.7	0.9	3.4	0.5	2.2	5.9	12.3	16.4	14.1	10.2	44.4	29.2	2.0	2.0	9.8	24.6	11.9	10.0
AfriMBART	7.1	2.2	2.9	3.7	3.3	2.8	10.2	15.2	10.5	7.7	43.4	28.3	7.2	6.2	8.4	32.8	12.0	7.5
M2M-100	**20.5**	**5.0**	7.1	7.7	**6.2**	9.5	**17.2**	18.5	19.4	12.8	44.7	29.9	19.8	10.5	13.5	**36.6**	**17.4**	**15.4**
M2M-100-EN/FR	20.3	4.9	**7.9**	**8.1**	6.0	**9.7**	11.6	19.8	19.2	**13.4**	**45.2**	30.2	21.1	**11.4**	**14.4**	9.1	15.8	12.5

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

with average BLEU of 6.3, despite being pre-trained on 101 languages. ByT5 outperforms MT5 by over 4 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 for 10 out of 16 languages for en/fr-xx and 8 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. 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 (see Appendix E). 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.

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 3.5 BLEU over ByT5 and AfriMT5 on *en-tsn* improved by 6.5 BLEU over MT5. On average, AfriMT5 improved over MT5 by 1.5 BLEU in en/fr-xx and 2.6 BLEU in the *xx-en/fr*. The improvement for AfriByT5 was much smaller: 0.7 and 1.1 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 ibo (3.3 BLEU) improved in *en/fr-xx* direction. However, in the *xx-en/fr* direction, fon, tsn, twi, and zul improved by $3.2 - 8.2$ 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

Model	*fr-xx* bam	bbj	ewe	fon	mos	wol	hau	ibo	lug	luo	*en-xx* pcm	swa	tsn	twi	yor	zul	AVG	MED
Transformer																		
REL+NEWS	2.9	0.3	4.1	1.4	1.6	0.7	8.9	15.7	0.1	1.9	11.9	15.9	25.2	6.6	6.0	7.3	6.9	5.1
M2M-100																		
REL+NEWS	17.0	3.5	7.3	5.0	3.5	10.0	12.0	20.4	12.9	10.8	**34.5**	26.6	29.1	9.2	10.6	14.2	14.2	11.4
REL→NEWS	**21.2**	**3.9**	7.7	5.6	**4.1**	10.7	14.1	**21.8**	13.7	11.3	34.4	27.5	31.5	9.4	10.7	18.8	15.4	12.5
REL+NEWS→NEWS	20.1	3.8	**8.6**	**5.8**	3.8	**11.4**	**14.4**	21.5	**14.2**	**11.3**	33.4	**28.3**	**31.9**	**9.8**	**11.5**	**18.9**	**15.5**	**12.9**
	xx-fr										*xx-en*							
Transformer																		
REL+NEWS	2.2	0.1	4.7	1.3	1.4	0.3	9.6	14.5	0.8	1.5	12.2	20.7	19.8	6.8	7.5	21.4	7.8	5.8
M2M-100																		
REL+NEWS	**20.7**	5.8	10.5	**7.8**	6.1	9.3	15.7	20.5	20.4	12.9	**44.7**	30.8	26.7	12.2	15.6	29.7	18.1	15.7
REL→NEWS	18.6	5.8	10.9	**7.8**	**6.9**	9.7	17.4	**21.3**	19.9	13.4	44.3	30.7	27.5	**12.9**	15.0	35.1	18.6	16.2
REL+NEWS→NEWS	**20.7**	**6.1**	**11.2**	7.6	6.5	**10.5**	**18.1**	20.9	**21.7**	**14.0**	44.0	**32.2**	**27.5**	**12.9**	**16.4**	**37.2**	**19.2**	**17.3**

Table 5: **Results adapting to Domain Shift**. We calculate BLEU on the news domain when training on different combinations of REL and NEWS.

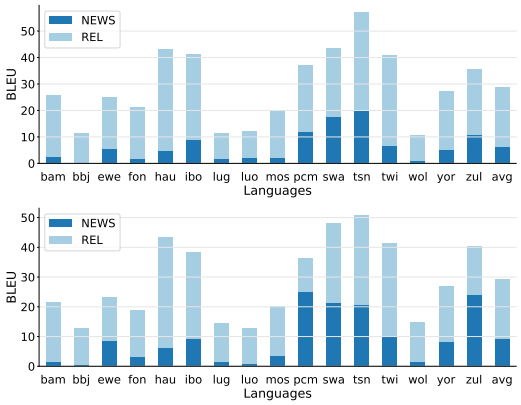


Figure 1: **Domain shift** of M2M-100 Transformer models trained on en/fr-xx (top) or xx-en/fr (bottom) REL domain and tested on the NEWS vs. REL domains.

bam-fr	
SRC	Ani k'a fɔu ye ko cɛmancɛ fanga bɛ sigi ntuloma saba kan.
TGT	Et leur dire que la transition se repose sur trois piliers.
REL	Et qu'on leur dise que la puissance du milieu est sur trois sauterelles;
R+N→N	Et de leur dire que la force de la transition repose sur trois piliers.
lug-en	
SRC	Murasaki Shikibu yawandiika ekitabo ekijjuvu akaasookera ddala mu nsi yonna.
TGT	Murasaki Shikibu wrote the world's first full novel.
REL	And Murshach Shikib writes a full scroll of the first in all the earth.
R+N→N	Murasaki Shikibu wrote a complete book first in the world.

Table 6: **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.

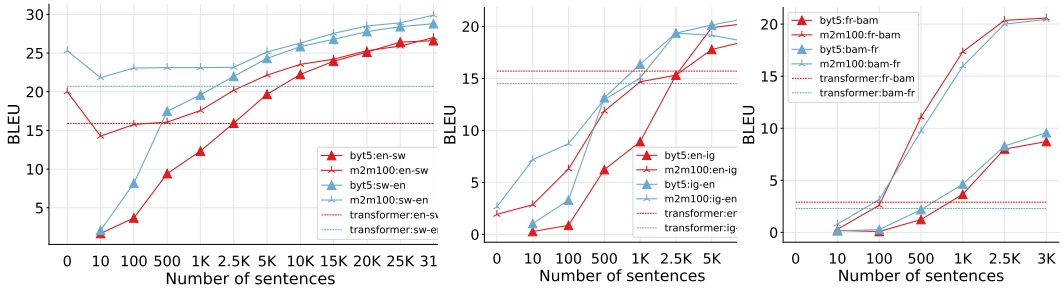
depends on the size of REL corpus. For languages trained on the Bible such as bbj, bam, lug, luo, and wol, the improvement is minimal. For M2M-100, the REL+NEWS performance does not improve over NEWS despite the larger quantity of training data. This demonstrates that increasing the size in the target 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 1.2 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 1.8 and 1.1 BLEU, respectively. Appendix F provides similar findings for the PLMs.

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. 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



(a) eng-swa (b) eng-ibo (c) fr-bam

Figure 2: **Number of fine-tuning sentences** needed to exceed the performance of a bilingual Transformer model.

(eng-ibo) and relatively small (eng-swa) domain gap, the quality of target domain transtions 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 6. 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 7 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 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. In the REL domain, the zero-shot transfer is around (1.7 – 14.6 BLEU) in both translation directions, evaluating on the Bible gave slightly worse result than JW300 since the latter contains other non-religious topics. For the Wikipedia domain, the transfer is much better than REL. For the *xx-en/fr* direction, the performance is between (7.5 – 25.8) and (4.0 – 27.1) in the *en/fr-xx* direction. This finding shows that expanding African news corpora and developing better MT models for news can be easily adapted to other domains of interest.

[10]except for Luo which is not supported

Domains	hau	ibo	lug	luo	swa	wol	yor	zul
en/fr-xx								
Baseline	2.6	2.8	0.8	–	20.9	0.6	1.5	3.3
REL	3.7	10.3	3.3	5.4	14.6	6.7	10.2	13.0
FLORES	4.0	19.9	7.6	13.7	27.1	8.2	10.4	19.2
NEWS	20.2	31.6	22.6	16.4	31.4	19.9	23.3	27.6
xx-en/fr								
Baseline	8.0	7.2	3.7	–	26.9	3.0	3.8	11.9
REL	3.8	6.0	1.7	2.5	13.9	1.7	5.5	12.5
FLORES	16.3	12.0	7.7	11.8	25.8	7.5	7.6	19.2
NEWS	17.6	22.8	24.4	15.8	32.0	12.3	16.3	39.0

Table 7: **spBLEU on Wikipedia domain** (FLORES), REL, and NEWS for M2M-100 fine-tuned on NEWS. Zero-shot evaluation was performed on domains in gray. "Baseline" was evaluated on the pre-trained M2M-100.

7 Conclusion

We have created FAAND-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.

References

David Adelani, Dana Ruiter, Jesujoba Alabi, Damilola Adebonojo, Adesina Ayeni, Mofe Adeyemi, Ayodele Esther Awokoya, and Cristina España-Bonet. 2021. 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.

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.

Roee 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 Jaimito 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, Franscisco Guzmán, Junjie Hu, Macduff Hughes, Philipp Koehn, Rosie Lazar, Will Lewis, Graham Neubig, Mengmeng Niu, Alp Öktem, Eric Paquin, Grace Tang, and Sylwia 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 Dansoa Appiah, Felix Akwerh, 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 Nartey Borkor, Standylove Birago Mensah, Lucien Mensah, Mark Amoako Marcel, Anokye Acheampong Amponsah, and James B. Hayfron-Acquah. 2021. 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.

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 Chinenye 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 Chinenye 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, Shruti Bhosale, 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, Shruti Bhosale, 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 Fagbohungbe, Solomon Oluwole Akinola, Shamsuddeen Muhammad, Salomon Kabongo Kabenamualu, Salomey Osei, Freshia Sackey, Rubungo Andre Niyongabo, Ricky Macharm, Perez Ogayo, Orevaoghene Ahia, Musie Meressa Berhe, Mofetoluwa Adeyemi, Masabata Mokgesi-Selinga, Lawrence Okegbemi, Laura Martinus, Kolawole Tajudeen, Kevin Degila, Kelechi Ogueji, Kathleen Siminyu, Julia Kreutzer, Jason Webster, Jamiil 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 Whitenack, Christopher Onyefuluchi, Chris Chinenye 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, Jindvrich 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, Zhirui 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, Fernanda 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, Sebastin 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, Jasmijn 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 Ulzii-Orshikh, Allahsera Auguste Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios Gonzales, Isabel Papadimitriou, Salomey Osei, Pedro Ortiz Suarez, Iroro Orife, Kelechi Ogueji, Rubungo Andre Niyongabo, Toan Q. Nguyen, Mathias Muller, Andr'e Muller, Shamsuddeen Hassan Muhammad, Nanda Muhammad, Ayanda Mnyakeni, Jamshidbek Mirzakhalov, Tapiwanashe Matangira, Colin Leong, Nze Lawson, Sneha Kudugunta, Yacine Jernite, M. Jenny, Orhan Firat, Bonaventure F. P. Dossou, Sakhile Dlamini, Nisansa de Silva, Sakine cCabuk Balli, Stella Rose Biderman, Alessia Battisti, Ahmed Baruwa, Ankur Bapna, Pallavi N. Baljekar, Israel Abebe Azime, Ayodele Awokoya, Duygu Ataman, Orevaoghene Ahia, Oghenefego Ahia, Sweta 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.

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.

Rooweither 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, Luis Chiruzzo, Gustavo Giménez-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, Mirko Plitt, 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, Vukosi Marivate, Sackey Freshia, Prateek Sibal, Bhanu Neupane, David Ifeoluwa Adelani, Amelia Taylor, Jamiil Toure Ali, Kevin Degila, Momboladji Balogoun, Thierno Ibrahima Diop, Davis David, Chayma Fourati, Hatem Haddad, and Malek Naski. 2021. Ai4d - 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.

Antonio Toral, Sheila Castilho, Ke Hu, and Andy Way. 2018. Attaining the unattainable? reassessing claims of human parity in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 113–123, Brussels, Belgium. Association for Computational Linguistics.

Chau Tran, Shruti Bhosale, James Cross, Philipp Koehn, Sergey Edunov, and Angela Fan. 2021. Facebook ai wmt21 news translation task submission. *arXiv preprint arXiv:2108.03265*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Zihan Wang, Karthikeyan K, Stephen Mayhew, and Dan Roth. 2020. Extending multilingual BERT to low-resource languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2649–2656, Online. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. 2021a. Byt5: Towards a token-free future with pre-trained byte-to-byte models. *ArXiv*, abs/2105.13626.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021b. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Jian Yang, Shuming Ma, Haoyang Huang, Dongdong Zhang, Li Dong, Shaohan Huang, Alexandre Muzio, Saksham Singhal, Hany Hassan Awadalla, Xia Song, and Furu Wei. 2021. Multilingual machine translation systems from microsoft for wmt21 shared task. *ArXiv*, abs/2111.02086.

A Language Characteristics

Table 8 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, 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., 2021) 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 9 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, Shona, Somali, Sesotho, Swahili, isiXhosa, Yorùbá, and isiZulu.

D Model Hyper-parameters for Reproducibility

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. All the experiments were performed on a single GPU (Nvidia V100).

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 of $5e-5$ for $50,000$ steps using Fairseq (Ott et al., 2019) without modifying the mBART50 vocabulary.

E ChrF Evaluation

Table 11 and Table 12 provides the evaluation of the Transformer and M2M-100 models using the ChrF metric. We find the ChrF to be better for morphologically rich African languages like bbj, lug, swa, tsn, and zul. For example, fine-tuning M2M-100 on NEWS and evaluating on zul has a BLEU of 16.5 in *en/fr-xx*, and BLEU of 36.6 in the *xx-en/fr* showing a large gap in performance in both directions. However, with the ChrF, we find a smaller performance gap for zul in both translation directions (51.2 in *en/fr-xx* and 55.5 in the *xx-en/fr*. This shows the limitation of BLEU for evaluating the performance of morphologically rich languages. On the other hand, for isolating languages like Fon, Mossi, and Yorùbá, we did not find so much additional benefit. Although, comparing Table 7 and Table 10, we find spBLEU to produce higher evaluation scores for both isolating and morphologically rich languages. The best metric for low-resource languages is still subject to further research. We leave this for future work.

F Baseline Results for All PLMs

Table 13 and Table 14 provides baseline results for PLMs (MT5, ByT5 and mBART50) on all exper-

[11] https://www.statmt.org/wmt21/translation-task.html

[12] https://sw.globalvoices.org/

Language	No. of Letters	Latin Letters Omitted	Letters added	Tonality	diacritics	sentence morphology	structure
Bambara (bam)	27	q,v,x	ɛ, ɔ, ɲ, ŋ	yes, 2 tones	yes	isolating	SVO & SOV
Ghomálá' (bbj)	40	q, w, x, y	bv, dz, ə, aə, ɛ, gh, ny, nt, ŋ, ŋk, ɔ, pf, mpf, sh, ts, ʉ, zh, '	yes, 5 tones	yes	agglutinative	SVO
Éwé (ewe)	35	c, j, q	ɖ, dz, ɛ, ƒ, gb, ɣ, kp, ny, ŋ, ɔ, ts, ʋ	yes, 3 tones	yes	isolating	SVO
Fon (fon)	33	q	ɖ, ɛ,gb, hw, kp, ny, ɔ, xw	yes, 3 tones	yes	isolating	SVO
Hausa (hau)	44	p,q,v,x	ɓ, ɗ, ƙ, ʸ, kw, ƙw, gw, ky, ƙy, gy, sh, ts	yes, 2 tones	no	agglutinative	SVO
Igbo (ibo)	34	c, q, x	ch, gb, gh, gw, kp, kw, nw, ny, ọ, ò, sh, ụ	yes, 2 tones	yes	agglutinative	SVO
Luganda (lug)	24	h, q, x	ŋ, ny	yes, 3 tones	no	agglutinative	SVO
Luo (luo)	31	c, q, x, v, z	ch, dh, mb, nd, ng', ng, ny, nj, th, sh	yes, 4 tones	no	agglutinative	SVO
Mossi (mos)	26	c, j, q, x	', ɛ, ɩ, ʋ	yes, 2 tones	yes	isolating	SVO
Naija (pcm)	26	–	–	no	no	mostly analytic	SVO
Swahili (swa)	33	x, q	ch, dh, gh, kh, ng', ny, sh, th, ts	no	yes	agglutinative	SVO
Setswana (tsn)	36	c, q, v, x, z	ê, kg, kh, ng, ny, ô, ph, š, th, tl, tlh, ts, tsh, tš, tšh	yes, 2 tones	no	agglutinative	SVO
Akan/Twi (twi)	22	c,j,q,v,x,z	ɛ, ɔ	yes, 5 tones	no	isolating	SVO
Wolof (wol)	29	h,v,z	ŋ, à, é, ë, ó, ñ	no	yes	agglutinative	SVO
Yorùbá (yor)	25	c, q, v, x, z	ẹ, gb, ṣ , ọ	yes, 3 tones	yes	isolating	SVO
isiZulu (zul)	55	–	nx, ts, nq, ph, hh, ny, gq, hl, bh, nj, ch, ngc, ngq, th, ngx, kl, ntsh, sh, kh, tsh, ng, nk, gx, xh, gc, mb, dl, nc, qh	yes, 3 tones	no	agglutinative	SVO

Table 8: Linguistic Characteristics of the Languages

imental settings (REL, REL+NEWS, REL→NEWS, and REL+NEWS→NEWS).

G 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 Alimanyiɖokpo kpódÍssa Etchlekoun kpó ɔ, sín azǎn mɔkpán ɖye ɔ, ye ɖò wǔvɛ sè wɛ tawun ɖò agbaza mɛ, có ye ká tuun fí é azɔn nɛ lɛɛ gosin é ɔ ǎ.

- **Reference** (fr): Les faits Louis Guy Alimagnidokpo et Issa Etchlekoun se plaignent 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, fon→fr): Louis Guy Alimanyion et Issa Etchlekoun ont depuis plusieurs jours souffert d'une maladie grave malgré les conséquences de cette maladie qu'ils ne connaissent pas.

- **M2M-100** (REL+NEWS→NEWS, fr→fon): Sín azǎn yɔywɛywɛ ɖé ɖye ɖokpóo wɛ́ nǔ è kàn Louis Guy Alimagnidokpo kpódó Issa Etchlɛkén kpán ɖè ɔ́ ɖò xó ɖɔ wé ɖɔ́ wǔvɛ́ gege wɛ́, ye ká tuun nǔ è wú wǔvɛ́ yetɔn ɖè ɔ́ ǎ.

The translation of the bilingual Transformer model 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 (sín azǎn mɔkpán ɖye ɔ, ye ɖò wǔvɛ sè wɛ tawun ɖò agbaza mɛ)* 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, *wǔvɛ* 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: *có ye ká tuun fí é azɔn nɛ lɛɛ gosin (but they do know the origin of their sufferings) é ɔ́ǎ (**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 *malgré* 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 Etchlekoun ont depuis plusieurs jours souffert d'une maladie grave ~~malgré~~ (dont) *ils ne connaissent pas* les ~~conséquences~~ (causes/raisons) ~~de cette maladie~~

Language	Source	Size (MB)	No. of sentences
Afrikaans (af)	mC4 (subset) (Xue et al., 2021b)	752.2MB	3,697,430
Amharic (am)	mC4 (subset), and VOA	1,300MB	2,913,801
Arabic (ar)	mC4 (subset)	1,300MB	3,939,375
English (en)	mC4 (subset), and VOA	2,200MB	8,626,571
French (fr)	mC4 (subset), and VOA	960MB	4,731,196
Hausa (ha)	mC4 (all), and VOA	594.1MB	3,290,382
Igbo (ig)	mC4 (all), and AfriBERTa Corpus (Ogueji et al., 2021)	287.5MB	1,534,825
Malagasy (mg)	mC4 (all)	639.6MB	3,304,459
Chichewa (ny)	mC4 (all), Chichewa News Corpus (Siminyu et al., 2021)	373.8MB	2,203,040
Oromo (om)	AfriBERTa Corpus, and VOA	67.3MB	490,399
Naija (pcm)	AfriBERTa Corpus, and VOA	54.8MB	166,842
Rwanda-Rundi (rw/rn)	AfriBERTa Corpus, KINNEWS & KIRNEWS (Niyongabo et al., 2020), and VOA	84MB	303,838
Shona (sn)	mC4 (all), and VOA	545.2MB	2,693,028
Somali (so)	mC4 (all), and VOA	1,000MB	3,480,960
Sesotho (st)	mC4 (all)	234MB	1,107,565
Swahili (sw)	mC4 (all)	823.5MB	4,220,346
isiXhosa (xh)	mC4 (all), and Isolezwe Newspaper	178.4MB	832,954
Yorùbá (yo)	mC4 (all), Alaroye News, Asejere News, Awikonko News, BBC, and VON	179.3MB	897,299
isiZulu (zu)	mC4 (all), and Isolezwe Newspaper	700.7MB	3,252,035

Table 9: Monolingual Corpora (after pre-processing – we followed AfriBERTa (Ogueji et al., 2021) approach) , their sources and size (MB), and number of sentences.

Domains	hau	ibo	lug	luo	swa	wol	yor	zul
en/fr-xx								
Baseline	2.4	2.0	0.9	–	19.6	0.4	1.0	1.9
REL	6.7	9.4	1.1	2.4	17.4	2.7	8.6	8.3
FLORES	2.9	12.3	4.9	8.8	22.5	4.2	4.0	8.4
NEWS	14.4	20.3	13.0	10.8	27.0	11.1	9.6	16.5
xx-en/fr								
Baseline	6.6	6.0	2.6	–	26.2	2.1	2.7	10.5
REL	7.7	10.7	1.8	2.6	20.5	1.7	9.2	12.9
FLORES	5.4	11.8	6.9	10.3	25.4	6.6	6.3	18.1
NEWS	17.2	18.5	19.4	12.8	29.9	9.5	13.5	36.6

Table 10: BLEU on the Wikipedia domain (FLORES), REL and NEWS for M2M-100 fine-tuned on NEWS.

~~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 nǔ è wú wǔvé yetɔn (they do know why they are suffering) ɖè ɔ́ ǎ (**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 *sín azǎn + number* means *since xxx days* but *yɛywɛ* is not a number, and do not have any meaning in this context.

	fr-xx						*en-xx*										
Model	**bam**	**bbj**	**ewe**	**fon**	**mos**	**wol**	**hau**	**ibo**	**lug**	**luo**	**pcm**	**swa**	**tsn**	**twi**	**yor**	**zul**	**AVG**
Transformer																	
REL+NEWS	25.6	13.0	32.2	11.8	20.5	18.3	35.9	47.3	27.4	25.5	40.7	46.7	55.0	36.0	30.8	50.4	32.3
M2M-100																	
REL	21.6	12.2	34.9	19.7	15.1	16.4	30.2	40.6	21.5	26.3	35.2	48.6	51.7	36.8	26.8	49.7	30.5
NEWS	48.2	23.1	30.9	27.6	16.7	35.7	43.2	50.0	45.5	39.0	64.0	56.4	52.0	38.2	34.5	51.2	41.0
REL+NEWS	46.8	22.1	36.7	26.2	16.0	33.5	38.4	50.1	44.5	38.1	64.7	53.0	57.2	39.7	35.7	53.1	41.0
REL→NEWS	44.1	22.6	34.1	27.7	16.8	34.7	41.3	51.3	45.6	38.6	64.7	57.2	59.3	40.6	35.9	56.3	41.9
REL+NEWS→NEWS	49.9	23.5	37.5	28.5	16.8	35.8	42.1	51.3	46.9	39.4	64.2	57.0	59.5	40.8	37.3	56.3	42.9

Table 11: ChrF on the news domain comparing Transformer model and M2M-100: en/fr-xx.

	xx-fr						*xx-en*										
Model	**bam**	**bbj**	**ewe**	**fon**	**mos**	**wol**	**hau**	**ibo**	**lug**	**luo**	**pcm**	**swa**	**tsn**	**twi**	**yor**	**zul**	**AVG**
Transformer																	
REL+NEWS	24.7	12.1	30.9	10.2	18.3	18.9	30.4	42.6	25.1	22.4	42.7	47.7	49.6	34.3	33.5	49.5	30.8
M2M-100																	
REL	20.7	18.1	33.6	19.4	15.3	21.1	10.3	13.2	22.4	20.3	54.0	49.0	10.9	35.5	31.6	49.0	26.5
NEWS	46.0	26.5	30.9	27.5	17.7	33.8	38.8	46.1	46.4	36.7	68.6	54.8	45.2	35.1	36.4	55.5	40.4
REL+NEWS	47.1	27.5	36.4	27.9	16.6	34.0	36.8	47.5	47.2	37.3	68.9	54.7	53.0	38.4	39.9	53.3	41.7
REL→NEWS	44.5	27.7	37.0	28.2	16.8	34.4	39.6	48.0	47.0	38.0	68.7	55.8	53.6	38.7	39.2	56.4	42.1
REL+NEWS→NEWS	49.0	28.5	37.2	28.9	17.2	35.3	40.2	47.9	48.5	38.3	68.6	55.7	54.0	38.7	40.7	57.7	42.9

Table 12: ChrF on the news domain comparing Transformer model and M2M-100: xx-en/fr

	fr-xx						*en-xx*										
Model	**bam**	**bbj**	**ewe**	**fon**	**mos**	**wol**	**hau**	**ibo**	**lug**	**luo**	**pcm**	**swa**	**tsn**	**twi**	**yor**	**zul**	**AVG**
REL+NEWS																	
Transformer	2.9	0.3	4.1	1.3	1.6	0.7	8.6	15.7	0.1	1.9	11.9	15.9	25.2	6.6	6.0	7.3	6.9
M2M-100 0-shot	–	–	–	–	–	1.3	0.4	2.6	–	–	–	20.0	1.0	–	1.9	4.3	–
NEWS																	
MT5	0.9	0.7	1.8	1.1	0.3	1.1	2.4	14.1	3.5	3.2	33.5	23.2	3.3	1.6	2.2	8.8	6.3
ByT5	8.7	1.6	4.4	2.3	0.4	5.7	8.8	18.6	11.3	8.8	32.4	26.6	15.5	6.2	6.2	12.2	10.6
mBART50	15.8	2.7	3.8	4.7	2.7	8.7	11.8	14.8	9.7	9.6	33.9	22.1	17.2	7.3	7.5	17.3	11.9
M2M-100	20.6	3.5	5.9	5.6	3.3	11.1	**14.4**	20.3	13.0	10.8	33.2	27.0	24.8	8.5	9.6	16.5	14.3
REL+NEWS																	
MT5	1.3	0.4	6.3	0.9	2.0	1.4	7.1	15.2	2.5	3.0	32.9	23.8	23.1	7.4	7.8	11.3	9.2
ByT5	8.7	1.4	7.3	2.6	2.8	4.9	9.0	19.4	11.1	8.3	31.9	25.2	24.6	7.9	9.7	13.8	11.8
mBART50	12.5	3.2	6.7	5.9	**4.1**	9.5	12.5	21.5	12.0	11.0	34.2	21.9	30.3	10.2	11.0	15.9	13.9
M2M-100	17.0	3.5	7.3	5.0	3.5	10.0	12.0	20.4	12.9	10.8	**34.5**	26.6	29.1	9.2	10.6	14.2	14.2
REL+NEWS→NEWS																	
MT5	3.6	0.5	8.3	1.2	2.3	1.5	9.7	19.1	4.9	4.6	34.0	27.6	28.2	9.0	9.3	15.3	11.2
ByT5	12.8	2.3	**8.8**	3.0	2.8	7.0	12.8	**22.4**	12.5	9.8	32.5	28.5	**32.0**	**10.4**	11.0	19.6	14.3
mBART50	15.6	3.2	7.3	**6.0**	3.7	10.2	14.1	21.1	12.2	10.7	33.1	23.4	31.1	9.8	10.9	**20.2**	14.5
M2M-100	20.1	3.8	8.6	5.8	3.8	**11.4**	**14.4**	21.5	**14.2**	**11.3**	33.4	**28.3**	31.9	9.8	11.5	18.9	**15.5**
REL→NEWS																	
M2M-100	**21.2**	**3.9**	7.7	5.6	**4.1**	10.7	14.1	21.8	13.7	11.3	34.4	27.5	31.5	9.4	10.7	18.8	15.4
NEWS: Additional pre-training on Monolingual texts.																	
AfriMT5	1.9	0.9	3.1	1.6	0.3	1.8	4.5	15.4	5.9	4.5	34.5	26.7	6.9	2.5	4.7	9.8	7.8
AfriByT5	10.6	1.9	4.2	2.3	0.7	6.2	9.8	19.3	12.2	9.0	32.4	27.5	18.0	6.3	7.1	13.4	11.3
AfriMBART	13.1	2.3	4.7	3.3	2.1	7.9	9.5	18.1	8.9	9.3	29.5	25.9	12.8	6.1	7.9	16.6	11.1
NEWS: Many-to-Many Multilingual MT model																	
M2M-100-EN/FR	16.1	2.6	5.8	3.3	2.7	9.6	6.5	18.0	8.6	9.6	34.4	26.3	19.5	6.6	6.8	10.9	11.7

Table 13: BLEU score on News domain : fr/en-xx

	xx-fr						*xx-en*										
Model	**bam**	**bbj**	**ewe**	**fon**	**mos**	**wol**	**hau**	**ibo**	**lug**	**luo**	**pcm**	**swa**	**tsn**	**twi**	**yor**	**zul**	**AVG**
REL+NEWS																	
Transformer	2.2	0.1	4.7	0.4	1.4	0.3	9.5	14.5	0.8	1.5	12.2	20.7	19.8	6.8	7.5	21.4	7.8
M2M-100 0-shot	–	–	–	–	–	0.9	2.4	6.3	–	–	–	25.3	3.1	–	2.7	13.7	–
NEWS																	
MT5	2.8	1.0	1.3	2.3	1.2	1.4	5.9	18.0	11.5	6.7	42.2	29.0	9.4	4.2	7.9	22.2	10.4
ByT5	9.6	2.6	4.1	4.0	2.7	6.8	14.0	20.8	19.3	11.9	43.4	28.8	19.0	10.5	9.6	25.2	14.5
mBART50	12.7	0.9	3.4	0.5	2.2	5.9	12.3	16.4	14.1	10.2	44.4	29.2	2.0	2.0	9.8	24.6	11.9
M2M-100	20.5	5.0	7.1	7.7	6.2	9.5	17.2	18.5	19.4	12.8	44.7	29.9	19.8	10.5	13.5	36.6	17.4
REL+NEWS																	
MT5	2.5	0.7	8.7	3.7	4.1	1.0	10.8	18.5	10.3	6.4	41.1	29.5	23.8	11.9	14.1	27.9	13.4
ByT5	9.3	2.2	11.4	5.2	5.8	5.0	13.1	21.3	17.7	11.1	42.1	24.1	25.0	13.2	12.3	29.3	15.5
mBART50	13.5	3.6	10.7	7.2	6.1	1.0	16.5	20.1	17.4	12.3	44.3	31.4	**28.3**	13.2	**16.5**	33.6	17.2
M2M-100	20.7	5.8	10.5	**7.8**	6.1	9.3	15.7	20.5	20.4	12.9	44.7	30.8	26.7	12.2	15.6	29.7	18.1
REL+NEWS→NEWS																	
MT5	5.2	0.9	10.3	3.8	5.5	1.7	14.3	20.8	14.2	9.1	42.7	31.4	26.6	13.3	15.0	30.8	15.4
ByT5	13.7	3.5	**12.5**	5.4	**7.2**	7.4	16.3	**22.9**	20.4	13.2	43.4	27.8	27.3	**14.6**	13.0	32.7	17.6
mBART50	16.3	3.6	10.4	6.7	6.0	3.3	16.7	19.2	17.1	12.5	43.5	31.4	27.2	12.6	15.4	**40.2**	17.6
M2M-100	20.7	**6.1**	11.2	7.6	6.5	**10.5**	**18.1**	20.9	**21.7**	**14.0**	44.0	**32.2**	27.5	12.9	16.4	37.2	**19.2**
REL→NEWS																	
M2M-100	18.6	5.8	10.9	7.8	6.9	9.7	17.4	21.3	19.9	13.4	44.3	30.7	27.5	12.9	15.0	35.1	18.6
NEWS: Additional pre-training on Monolingual texts.																	
AfriMT5	5.8	2.3	2.4	4.0	2.5	2.8	10.7	19.1	14.8	9.4	44.7	30.7	15.9	8.1	11.5	23.4	13.0
AfriByT5	13.1	4.2	4.5	5.2	4.6	8.5	14.7	20.5	20.6	12.4	43.4	29.0	20.2	11.2	10.4	26.5	15.6
AfriMBART	7.1	2.2	2.9	3.7	3.3	2.8	10.2	15.2	10.5	7.7	43.4	28.3	7.2	6.2	8.4	32.8	12.0
NEWS: Many-to-Many Multilingual MT model																	
M2M-100-EN/FR	20.3	4.9	7.9	**8.1**	6.0	9.7	11.6	19.8	19.2	13.4	**45.2**	30.2	21.1	11.4	14.4	9.1	15.8

Table 14: BLEU score on News domain : xx-fr/en

H 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 G 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.