The Effect of Domain and Diacritics in Yorùbá–English Neural Machine Translation

David Ifeoluwa Adelani* didelani@lsv.uni-saarland.de
Spoken Language Systems Group (LSV), Saarland University, Germany & Masakhane NLP

Dana Ruiter* druiter@lsv.uni-saarland.de
Spoken Language Systems Group (LSV), Saarland University, Germany

Jesujoba O. Alabi* jalabi@mpi-inf.mpg.de
Max Planck Institute for Informatics, Saarbrücken, Germany & Masakhane NLP

Damilola Adebonojo iyayorubagidi@gmail.com
Alamoja Yoruba & Masakhane NLP

Adesina Ayeni info@yobamoodua.org
Yobamoodua Cultural Heritage (YMCH)

Mofe Adeyemi mofetoluwa@outlook.com
Defence Space Administration, Abuja, Nigeria & Masakhane NLP

Ayodele Awokoya ayodeleawokoya@gmail.com
University of Ibadan, Nigeria & Masakhane NLP

Cristina España-Bonet cristinae@dfki.de
DFKI GmbH, Saarland Informatics Campus, Saarbrücken, Germany

Abstract

Massively multilingual machine translation (MT) has shown impressive capabilities, including zero and few-shot translation between low-resource language pairs. However, these models are often evaluated on high-resource languages with the assumption that they generalize to low-resource ones. The difficulty of evaluating MT models on low-resource pairs is often due to lack of standardized evaluation datasets. In this paper, we present MENYO-20k, the first multi-domain parallel corpus with a special focus on clean orthography for Yorùbá–English with standardized train-test splits for benchmarking. We provide several neural MT benchmarks and compare them to the performance of popular pre-trained (massively multilingual) MT models both for the heterogeneous test set and its subdomains. Since these pre-trained models use huge amounts of data with uncertain quality, we also analyze the effect of diacritics, a major characteristic of Yorùbá, in the training data. We investigate how and when this training condition affects the final quality and intelligibility of a translation. Our models outperform massively multilingual models such as Google (+8.7 BLEU) and Facebook M2M (+9.1 BLEU) when translating to Yorùbá, setting a high quality benchmark for future research.

* Equal contribution to the work

1 Introduction

Neural machine translation (NMT) achieves high quality performance when large amounts of parallel sentences are available (Barrault et al., 2020). Large and freely-available parallel corpora do exist for a small number of high-resource pairs and domains. However, for low-resource languages such as Yorùbá (yo), one can only find few thousands of parallel sentences online[1]. In the best-case scenario, i.e. some amount of parallel data exists, one can use the Bible — the Bible is the most available resource for low-resource languages (Resnik et al., 1999)— and JW300 (Agić and Vulić, 2019). Notice that both corpora belong to the religious domain and they do not generalize well to popular domains such as news and daily conversations.

In this paper, we address this problem for the Yorùbá–English (yo–en) language pair by creating a multi-domain parallel dataset, MENYO-20k, which we make publicly available[2] with CC BY-NC 4.0 licence. It is a heterogeneous dataset that comprises texts obtained from news articles, TED talks, movie and radio transcripts, science and technology texts, and other short articles curated from the web and translated by professional translators. Based on the resulting train-development-test split, we provide a benchmark for the yo–en translation task for future research on this language pair. This allows us to properly evaluate the generalization of MT models trained on JW300 and the Bible on new domains. We further explore transfer learning approaches that can make use of a few thousand sentence pairs for domain adaptation. Finally, we analyze the effect of Yorùbá diacritics on the translation quality of pre-trained MT models, discussing in details how this affects the understanding of the translated text especially in the *en–yo* direction. We show the benefit of automatic diacritic restoration in addressing the problem of noisy diacritics.

2 The Yorùbá Language

The Yorùbá language is the third most spoken language in Africa, and it is native to south-western Nigeria and the Republic of Benin. It is one of the national languages in Nigeria, Benin and Togo, and spoken across the West African regions. The language belongs to the Niger-Congo family, and it is spoken by over 40 million native speakers (Eberhard et al., 2019).

Yorùbá has 25 letters without the Latin characters c, q, v, x and z, and with additional characters ẹ, gb, ṣ , ọ. Yorùbá is a tonal language with three tones: low, middle and high. These tones are represented by the grave (e.g. "à "), optional macron (e.g. "ā") and acute (e.g. "á") accents respectively. These tones are applied on vowels and syllabic nasals, but the mid tone is usually ignored in writings. The tone information and underdots are important for the correct pronunciation of words. Often, articles written online, including news articles such as BBC[3] ignore diacritics. Ignoring diacritics makes it difficult to identify or pronounce words except when they are embedded in context. For example, *èdè* (language), *edé* (crayfish), *ẹdẹ* (a town in Nigeria), *ẹ̀dẹ* (trap) and *ẹ̀dẹ̀* (balcony) will be mapped to *ede* without diacritics.

Machine translation might be able to learn to disambiguate the meaning of words and generate correct English even with un-diacriticized Yorùbá. However, one cannot generate correct Yorùbá if the training data is un-diacriticized. One of the purposes of our work is to build a corpus with correct and complete diacritization in several domains.

3 MENYO-20k

The dataset collection was motivated by the inability of machine translation models trained on JW300 to generalize to new domains (∀ et al., 2020). Although ∀ et al. (2020) evaluated this

[1] http://opus.nlpl.eu

[2] https://github.com/uds-lsv/menyo-20k_MT

[3] https://www.bbc.com/yoruba

Data name	Source	No. Sent.
source language: en-yo		
JW News	jw.org/yo/iroyin	3,508
VON News	von.gov.ng	3,048
GV News	globalvoices.org	2,932
Yorùbá Proverbs	@yoruba_proverbs	2,700
Movie Transcript	"Unsane" on YouTube	774
UDHR	ohchr.org	100
ICT localization	from Yorùbá translators	941
Short texts	from Yorùbá translators	687
source language: en		
TED talks	ted.com/talks	2,945
Out of His Mind	from the book author	2,014
Radio Broadcast	from Bond FM Radio	258
CC License	Creative Commons	193
Total		20,100

	Number of Sentences		
Domain	**Train. Set**	**Dev. Set**	**Test Set**
MENYO-20k			
News	4,995	1,391	3,102
TED Talks	507	438	2,000
Book	-	1,006	1,008
IT	356	312	273
Yorùbá Proverbs	2,200	250	250
Others	2,012	250	250
Standard (religious) corpora			
Bible	30,760	–	–
JW300	459,871	–	–
TOTAL	500,701	3,397	6,633

Table 1: **Left:** Data collection. **Right:** MENYO-20k domains and training, development and test splits (top); figures for standard corpora used in this work (bottom).

for Yorùbá with surprisingly high BLEU scores, the evaluation was done on very few examples from the COVID-19 and TED Talks domains with 39 and 80 sentences respectively. Inspired by the FLoRes dataset for Nepali and Sinhala (Guzmán et al., 2019), we create a high quality test set for Yorùbá-English with few thousands of sentences in different domains to check the quality of industry MT models, pre-trained MT models, and MT models based on popular corpora such as JW300 and the Bible.

3.1 Dataset Collection for MENYO-20k

Table 1 summarizes the texts collected, their source, the original language of the texts and the number of sentences from each source. We collected both parallel corpora freely available on the web (e.g JW News) and monolingual corpora we are interested in translating (e.g. the TED talks) to build the MENYO-20k corpus. The JW News is different from the JW300 since they contain only news reports, and we manually verified that they are not in JW300. Some few sentences were donated by professional translators such as "short texts" in Table 1. Our curation followed two steps: (1) translation of monolingual texts crawled from the web by professional translators; (2) verification of translation, orthography and diacritics for parallel texts obtained online and translated. Texts obtained from the web that were judged by native speakers being high quality were verified once, the others were verified twice. The verification of translation and diacritics was done by professional translators and volunteers who are native speakers.

Table 1 on the right (top) summarizes the figures for the MENYO-20k dataset with 20,100 parallel sentences split into 10,070 training sentences, 3,397 development sentences, and 6,633 test sentences. The test split contains 6 domains, 3 of them have more than 1000 sentences and can be used as domain test sets by themselves.

3.2 Other Corpora for Yorùbá and English

Parallel corpora For our experiments, we use two widely available parallel corpora from the religion domain, Bible and JW300 (Table 1, bottom). The parallel version of the Bible is not available, so we align the verses from the New International Version (NIV) for English and the Bible Society of Nigeria version (BSN) for Yorùbá. After aligning the verses, we obtain

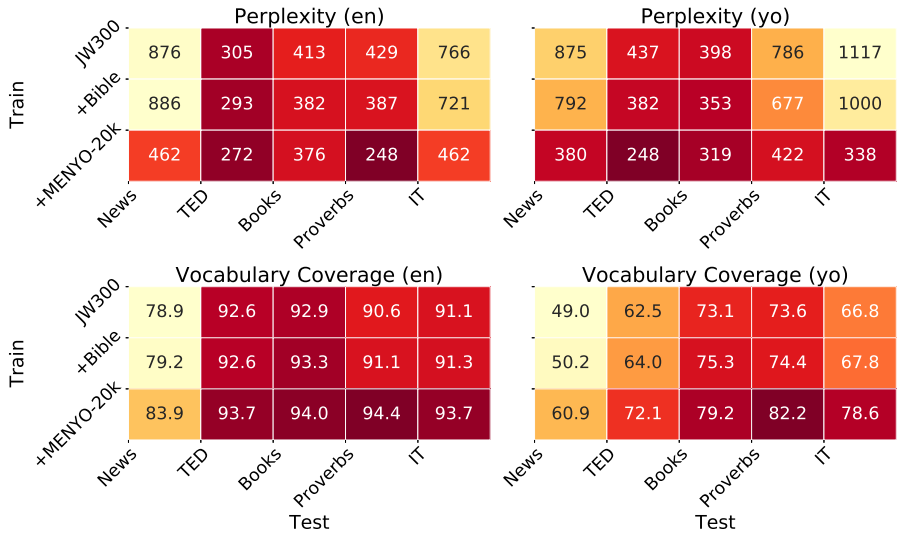

Figure 1: **Top:** Perplexities of KenLM 5-gram language model learned on different training corpora and tested on subsets of MENYO-20k for English (left) and Yorùbá (right) respectively. **Bottom:** Vocabulary coverage (%) of different subsets of the MENYO-20k test set per training sets for English (left) and Yorùbá (right).

30,760 parallel sentences. Also, we download the JW300 parallel corpus which is available for a large variety of low-resource language pairs. It has parallel corpora from English to 343 languages containing religion-related texts. From the JW300 corpus, we get $459,871$ sentence pairs already tokenized with *Polyglot*[4] (Al-Rfou, 2015).

Monolingual Corpora We make use of additional monolingual data to train the semi-supervised MT model using back-translation. The Yorùbá monolingual texts are from the Yorùbá embedding corpus (Alabi et al., 2020), one additional book ("Ojowu") with permission from the author, JW300-*yo*, and Bible-*yo*. We only use Yorùbá texts that are properly diacritized. In order to keep the topics in the Yorùbá and English monolingual corpora close, we choose two Nigerian news websites (The Punch Newspaper[5] and Voice of Nigeria [6]) for the English monolingual corpus. The news scraped covered categories such as politics, business, sports and entertainment. Overall, we gather 475,763 monolingual sentences from the website.

3.3 Dataset Domain Analysis

MENYO-20k is, on purpose, highly heterogeneous. In this section we analyze the differences and how its (sub)domains depart from the characteristics of the commonly used Yorùbá–English corpora for MT.

Characterizing the domain of a dataset is a difficult task. Some metrics previously used need either large corpora or a characteristic vocabulary of the domain (Beyer et al., 2020; España-Bonet et al., 2020). Here, we do not have these resources and we report the overlapping vocabulary between training and test sets and the perplexity observed in the test sets when a language model (LM) is trained on the MT training corpora.

[4] https://github.com/aboSamoor/polyglot
[5] https://punchng.com
[6] https://von.gov.ng

In order to estimate the perplexities, we train a language model of order 5 with KenLM (Heafield, 2011) on each of the 3 training data subsets: JW300 (named C2 for short in tables), JW300+Bible (C3), JW300+Bible+MENYO-20k (C4). Following NMT standard processing pipelines (see subsection 4.2), we perform byte-pair encoding (BPE) (Sennrich et al., 2016) on the corpora to avoid a large number of out-of-vocabulary tokens which, for small corpora, could alter the LM probabilities. For each of the resulting language models, we evaluate their average **perplexity** on the different domains of the test set to evaluate *compositional* domain differences (Figure 1, top). As expected, the average perplexity drops when adding more training data. Due to the limited domain of both JW300 and Bible, a literary style close to the Books domain, the decrease in perplexity is small when adding additional Bible data to JW300, namely -8% (*en*) and -11% (*yo*). Interestingly, both JW300 and Bible also seem to be close to the TED domain (1st and 2nd lowest perplexities for *en* and *yo* respectively), which may be due to discourse/monologue content in both training corpora. Adding the domain-diverse MENYO-20k corpus largely decreases the perplexity across all domains with a major decrease of -66% on IT (*yo*) and smallest decrease of -1% on Books (*en*). The perplexity scores correlate negatively with the resulting BLEU scores in Table 3, with a Pearson's r (r) of -0.367 (*en*) and -0.461 (*yo*), underlining that compositional domain differences between training and test subsets is the main factor of differences in translation quality.

Further, to evaluate *lexical* domain differences, we calculate the **vocabulary coverage** (tokenized, not byte-pair encoded[7]) of the different domains of the test set by each of the training subsets (Figure 1, bottom). The vocabulary coverage increases to a large extend when MENYO-20k is added. However, while vocabulary coverage and average perplexities have a strong (negative) correlation, $r = -0.756$ (*en*) and $r = -0.689$ (*yo*), a high perplexity does not necessarily mean low vocabulary coverage. E.g., the vocabulary coverage of the IT domain by JW300 is high (91% for *en*) despite leading to high perplexities (765 for *en*). In general, vocabulary coverage of the test sets is less indicative of the resulting translation performance than perplexity, showing only a weak correlation between vocabulary coverage and BLEU, with $r = 0.150$ and $r = 0.281$ for *en* and *yo* respectively.

4 Neural Machine Translation for Yorùbá–English

4.1 Systems

Supervised NMT We use the transformer-base architecture proposed by Vaswani et al. (2017) as implemented in Fairseq[8] (Ott et al., 2019). We set the drop-out at 0.3 and batch size at $10,240$ tokens. For optimization, we use *adam* (Kingma and Ba, 2015) with $\beta_1 = 0.9$ and $\beta_2 = 0.98$ and a learning rate of 0.0005. The learning rate has a warmup update of 4000, using label smoothed cross-entropy loss function with label-smoothing value of 0.1.

Semi-supervision via iterative back-translation We use the best performing supervised system to translate the monolingual corpora described in section 3 yielding to $476k$ back-translations. This data is used together with the original corpus to train a new system. The process is repeated until convergence.

Fine-tuning mT5 We examine a transfer learning approach by fine-tuning a massively multilingual model mT5 (Xue et al., 2021). mT5 had been pre-trained on 6.3T tokens originating from Common Crawl in 101 languages (including Yorùbá). The approach has already shown competitive results on other languages (Tang et al., 2020). In our experiments, we use mT5-

[7] We do not use byte-pair encoded data here, since, due to the nature of BPE, the vocabulary overlap would be close to 1 between all training and test sets.

[8] `https://github.com/pytorch/fairseq`

base, a model with 580M parameters. We transferred all the parameters of the model including the sub-word vocabulary.

Publicly Available NMT Models We further evaluate the performance of three multilingual NMT systems: OPUS-MT (Tiedemann and Thottingal, 2020), Google Multilingual NMT (GMNMT) (Arivazhagan et al., 2019) and Facebook's M2M-100 (Fan et al., 2020) with 1.2B parameters. All the three pre-trained models are trained on over 100 languages. While GMNMT and M2M-100 are a single multilingual model, OPUS-MT models are for each translation direction, e.g *yo–en*. We generate the translations of the test set using the *Google Translate* interface,[9] and OPUS-MT using *Easy-NMT*.[10] For M2M-100, we make use of *Fairseq* to translate the test set.

4.2 Experimental Settings

Data and Preprocessing For the MT experiments, we use the training part of our MENYO-20k corpus and two other parallel corpora, Bible and JW300 (section 3). For tuning the hyperparameters, we use the development split of the multi-domain data which has $3,397$ sentence pairs and for testing the test split with $6,633$ parallel sentences. To ensure that all the parallel corpora are in the same format, we convert the Yorùbá texts in the JW300 dataset to Unicode Normalization Form Composition (NFC), the format of the Yorùbá texts in the Bible and multi-domain dataset. Our preprocessing pipeline includes punctuation normalization, tokenization, and truecasing. For punctuation normalization and truecasing, we use the *Moses* toolkit (Koehn et al., 2007) while for tokenization, we use *Polyglot*, since it is the tokenizer used in JW300. We apply joint BPE, with a vocabulary threshold of 20 and 40k merge operations.

Evaluation Metrics To evaluate the models, we use tokenized BLEU (Papineni et al., 2002) score implemented in *multi-bleu.perl* and confidence intervals ($p = 95\%$) in the scoring package[11]. Since diacritics are applied on individual characters, we also use chrF, a character n-gram F1-score (Popović, 2015), for *en–yo* translations.

Automatic Diacritization In order to automatically diacritize Google MNMT and M2M-100 outputs for comparison, we train an automatic diacritization system using the supervised NMT setup. We use the Yorùbá side of MENYO-20k and JW300, which use consistent diacritization. We split the resulting corpus into train ($458k$ sentences), test (517 sentences) and development (500 sentences) portions. We apply a small BPE of 2k merge operations to the data. We apply noise on the diacritics by $i)$ randomly removing a diacritic with probability $p = 0.3$ and $ii)$ randomly replacing a diacritic with $p = 0.3$. The corrupted version of the corpus is used as the source data, and the NMT model is trained to reconstruct the original diacritics. On the test set, where the corrupted source has a BLEU (precision) of 19.0 (29.8), reconstructing the diacritics using our system lead to a BLEU (precision) of 87.0 (97.1), thus a major increase of $+68.0$ ($+67.3$) respectively.

4.3 Automatic Evaluation

Internal Comparison We train four basic NMT engines on different subsets of the training data: Bible (C1), JW300 (C2), JW300+Bible (C3) and JW300+Bible+MENYO-20k (C4). Further, we analyse the effect of fine-tuning for in-domain translation. For this, we fine-tune the converged model trained on JW300+Bible on MENYO-20k (C3+Transfer) and, similarly, we fine-tune the converged model trained on JW300+Bible+MENYO-20k on MENYO-20k (C4+Transfer). This yields six NMT models in total for *en–yo* and *yo–en* each. Their transla-

[9] https://translate.google.com/

[10] https://github.com/UKPLab/EasyNMT

[11] https://github.com/lvapeab/confidence_intervals

Model	*en–yo*		*en–yo*^{p}		*yo–en*	*yo–en*^{u}
	chrF	BLEU	chrF	BLEU	BLEU	BLEU
Internal Comparison						
C1: Bible	16.9	2.2±0.1	–	–	1.4±0.1	1.6±0.1
C2: JW300	29.1	7.5±0.2	–	–	9.6±0.3	9.3±0.3
C3: JW300+Bible	29.8	8.1±0.2	–	–	10.8±0.3	10.5±0.3
+Transfer	33.8	12.3±0.3	–	–	13.2±0.3	13.9±0.3
C4: JW300+Bible+MENYO-20k	32.5	10.9±0.3	–	–	14.0±0.3	14.0±0.3
+Transfer	34.3	**12.4**±0.3	–	–	14.6±0.3	–
+ BT	34.6	12.0±0.3	–	–	18.2±0.4	–
mT5: mT5-base+Transfer	32.9	11.5±0.3	–	–	16.3±0.4	16.3±0.4
External Comparison						
OPUS-MT	–	–	–	–	5.9±0.2	–
Google GMNMT	18.5	3.7±0.2	34.4	10.6±0.3	**22.4**±0.5	–
Facebook M2M-100	15.8	3.3±0.2	25.7	6.8±0.3	4.6±0.3	–

Table 2: Tokenized BLEU with confidence intervals ($p = 95\%$) and chrF scores over the full test for NMT models trained on different subsets of the training data C_i (top) and performance of external systems (bottom). For Yorùbá, we analyse the effect of diacritization: *en–yo*^{p} applies an in-house diacritizer on the translations obtained from **p**re-trained models and *yo–en*^{u} reports results using **u**ndiacritized Yorùbá texts as source sentences for training (see text). Top-scoring results per block are underlined and globally boldfaced.

tion performance is evaluated on the complete MENYO-20k test set (Table 2, top) and later we analyze in-domain translation in Table 3.

As expected, the BLEU scores obtained after training on Bible only (C1) are low, with BLEU 2.2 and 1.4 for *en–yo* and *yo–en* respectively, which is due to its small amount of training data. Training on the larger JW300 corpus (C2) leads to higher scores of BLEU 7.5 (*en–yo*) and 9.6 (*yo–en*), while combining it with Bible (C3) only leads to a small increase of BLEU $+0.6$ and $+1.2$ for *en–yo* and *yo–en* respectively. When further adding MENYO-20k (C4) to the training data, the translation quality increases by $+2.8$ (*en–yo*) and $+3.2$ (*yo–en*). When, instead of adding MENYO-20k to the training pool, it is used to fine-tune the converged JW300+Bible model, (C3+Transfer) the increase in BLEU over JW300+Bible is even larger for *en–yo* (BLEU $+4.2$), which results in an overall top-scoring model with BLEU 12.3. For *yo–en* fine-tuning is slightly less effective (BLEU 13.2) than simply adding MENYO-20k to the training data (BLEU 14.0). As seen in subsection 3.3, perplexities and vocabulary coverage in English are not as distant among training/test sets as in Yorùbá, so the fine-tuning step resulted less efficient.

When we use the MENYO-20k dataset to fine-tune the converged JW300+Bible+ MENYO-20k model (C4+Transfer) we observe an increase in BLEU over JW300+Bible for both translation directions: $+4.3$ for *en–yo* and $+3.8$ for *yo–en*. This is the best performing system and the one we use for back-translation. Table 2 also shows the performance of the semi-supervised system (C4+Transfer+BT). After two iterations of BT, we obtain an improvement of $+3.6$ BLEU points on *yo–en*. There is, however, no improvement in the *en–yo* direction probably because a significant portion of our monolingual data is based on JW300. Finally, fine-tuning mT5 with MENYO-20k does not improve over fine-tuning only the JW300+Bible system on *en–yo*, but it does for *yo–en*. Again, multilingual systems are stronger when used for English, and we need the contribution of back-translation to outperform the generic mT5.

External Comparison We evaluate the performance of the open source multilingual engines introduced in the previous section on the full test set (Table 2, bottom). **OPUS-MT**, while having no model available for *en–yo*, achieves a BLEU of 5.9 for *yo–en*. Thus, despite being trained on JW300 and other available *yo–en* corpora on OPUS, it is largely outperformed by our NMT model trained on JW300 only (BLEU +3.7). This may be caused by some of the noisy corpora included in OPUS (like CCaligned), which can depreciate the translation quality.

Facebook's **M2M-100**, is also largely outperformed even by our simple JW300 baseline by 5 BLEU points in both translation directions. A manual examination of the *en–yo* LASER extractions used to train M2M-100 shows that these are very noisy similar to the findings of Caswell et al. (2021), which explains the poor translation performance.

Google, on the other hand, obtains impressive results with **GMNMT** for the *yo–en* direction, with BLEU 22.4. The opposite direction *en–yo*, however, shows a significantly lower performance (BLEU 3.7), being outperformed even by our simple JW300 baseline (BLEU +3.8). The difference in performance for English can be attributed to the highly multilingual but English-centric nature of the Google MNMT model. As already noticed by Arivazhagan et al. (2019), low-resourced language pairs benefit from multilinguality when translated into English, but improvements are minor when translating into the non-English language. For the other translation direction, *en–yo*, we notice that lots of diacritics are lost in Google translations, damaging the BLEU scores. Whether this drop in BLEU scores really affects understanding or not is analyzed via a human evaluation (Section 4.4).

Diacritization Diacritics are important for Yorùbá embeddings (Alabi et al., 2020). However, they are often ignored in popular multilingual models (e.g. multilingual BERT (Devlin et al., 2019)), and not consistently available in training corpora and even test sets. In order to investigate whether the diacritics in Yorùbá MT can help to disambiguate translation choices, we additionally train *yo–en*^{u} equivalent models on **u**ndiacritized JW300, JW300+Bible and JW300+Bible+MENYO-20k (Table 2, indicated as *yo–en*^{u} in comparison to the ones with diacritics *yo–en*). Since one cannot generate correct Yorùbá text when training without diacritics, *en–yo*^{u} systems are not trained. Alternatively, we restore diacritics using our in-house diacritizer in the output of open source models that produce undiacritized text.

Results for *yo–en* are not conclusive. Diacritization is useful when only out-of-domain data is used in training (JW300, JW300+Bible[12] for testing on MENYO-20k). In this case, the domain of the training data is very different from the domain of the test set, and disambiguation is needed not to bias all the lexicon towards the religious domain. When we include in-domain data (JW300+Bible+MENYO-20k), both models perform equally well, with BLEU 14.0 for both diacritized and undiacritized versions. Diacritization is not needed when we fine-tune the model with data that shares the domain with the test set (JW300+Bible+Transfer), BLEU is 13.2 for the diacritized version vs. BLEU 13.9 for the undiacritized one.

In practice, this means that, when training data is far from the desired domain, investing work for a clean diacritized Yorùbá source input can help improve the translation performance. When more data is present, the diacritization becomes less important, since context is enough for disambiguation.

When Yorùbá is the target language, diacritization is always needed. An example is the low automatic scores GMNMT (BLEU 3.7, chrF 18.5) and M2M-100 (BLEU 3.3, chrF 15.8) reach for *en–yo* translation. Table 2-bottom (indicated as *en–yo*^{p}) show the improvements after automatically restoring the diacritics, namely $BLEU + 6.9$ points, chrF +15.9 for GMNMT; and +3.5 and +9.9 for M2M-100. Even if the diacritizer is not perfect, diacritics do not seem enough to get state-of-the-art results according to automatic metrics: fine-tuning with high

[12] We do not consider Bible alone. Due to its small data size, the BLEU scores are less indicative.

	en–yo					*yo–en*				
	Prov.	**News**	**TED**	**Book**	**IT**	**Prov.**	**News**	**TED**	**Book**	**IT**
C1	0.8	1.7	3.1	3.4	1.5	1.1	0.9	2.1	2.4	0.9
C2	2.2	6.4	9.8	9.8	4.8	2.6	8.4	13.1	9.6	7.0
C3	3.5	6.7	10.7	11.3	4.9	4.8	9.5	14.4	10.9	7.8
+Transfer	9.0	10.2	**16.1**	**15.0**	11.8	8.6	12.5	16.8	10.8	9.7
C4	7.0	10.0	12.3	11.5	10.5	8.7	13.5	16.7	11.6	12.4
+Transfer	**10.3**	10.9	15.1	13.2	**13.6**	**9.3**	14.0	17.8	11.9	13.7
+BT	7.5	11.4	12.9	14.5	9.7	7.9	**18.6**	**20.6**	**13.3**	**16.4**
mT5+Transfer	3.8	**11.2**	13.1	11.8	7.9	6.0	16.4	18.9	13.1	15.1

Table 3: Tokenized BLEU over different domains of the test set for NMT models trained on different subsets of the training data, with top-scoring results per domain in bold.

	en–yo			*yo–en*		
Task	**C4+Trf**	**C4+Trf+BT**	**GMNMT**	**mT5+Trf**	**C4+Trf+BT**	**GMNMT**
Adequacy	3.12*	3.58	**3.69**	3.42*	3.41*	**4.02**
Fluency	**4.57***	4.49*	3.74	4.39*	4.18*	**4.71**
Diacritics acc.	**4.91***	4.90*	1.74	-	-	-

Table 4: Human evaluation for *en–yo* and *yo–en* MT models (C4+Transfer (C4+Trf), C4+Trf+BT, mT5+Trf, and GMNMT) in terms of Adequacy, Fluency and Diacritics prediction accuracy. The rating that is significantly different from GMNMT is indicated by * (T-test $p < 0.05$)

.

quality data (C4+Transfer+BT, chrF 34.6) is still better than using huge but unadapted systems.

Domain Differences In order to analyze the domain-specific performance of the different NMT models, we evaluate each model on the different domain subsets of the test set (Table 3). The Proverb subset is especially difficult in both directions, as it shows the lowest translation performance across all domains, i.e. maximum BLEU of 9.04 (*en–yo*) and 8.74 (*yo–en*). This is due to the fact that proverbs often do not have literal counterparts in the target language, thus making them especially difficult to translate. The TED domain is the best performing test domain, with maximum BLEU of 16.1 (*en–yo*) and 16.8 (*yo–en*). This can be attributed to the decent base coverage of the TED domain by JW300 and Bible together (monologues) with the additional TED domain data included in the MENYO-20k training split (507 sentence pairs). Also, most BLEU results are on line with the LM perplexity results and conclusions drawn in subsection 3.3. Due to the closeness of Bible and JW300 to the book domain, we see only small improvements of BLEU on this domain, i.e. $+0.2$ (*en–yo*) and $+0.7$ (*yo–en*), when adding MENYO-20k (C4) to the JW300+Bible (C3) training data pool. On the other hand, the IT domain benefits the most from the additional MENYO-20k data, with major gains of BLEU $+5.5$ (*en–yo*) and 4.6 (*yo–en*), owing to the introduction of IT domain content in the MENYO-20k training data ($\sim 1k$ sentence pairs), which is completely lacking in JW300 and Bible.

4.4 Human Evaluation

To have a better understanding of the quality of the translation models and the intelligibility of the translations, we compare three top performing models in *en–yo* and *yo–en*. For *en–yo*, we

use **C4+Transfer**, **C4+Transfer+BT** and **GMNMT**. Although GMNMT is not the third best system according to BLEU (Table 2), we are interested in the study of diacritics in translation quality and intelligibility. For the *yo–en*, we choose **C4+Transfer+BT**, **mT5+Transfer** and **GMNMT** being the 3 models with the highest BLEU scores on Table 2.

We ask 7 native speakers of Yorùbá that are fluent in English to rate the adequacy, fluency and diacritic accuracy in a subset of test sentences. Four of them rated the *en–yo* translation direction and the others rated the opposite direction *yo–en*. We randomly select 100 sentences within the outputs of the six systems and duplicate 5 of them to check the intra-agreement consistency of our raters. Each annotator is then asked to rate 105 sentences per system on a $1-5$ Likert scale for each of the features (for English, diacritic accuracy cannot be evaluated). We calculate the agreement among raters using Krippendorff's α. The inter-agreement per task is 0.44 (adequacy), 0.40 (fluency) and 0.87 (diacritics) for Yorùbá, and 0.71 (adequacy), 0.55 (fluency) for English language. We observe that a lot of raters often rate the fluency score for many sentences with the same values (e.g 4 or 5), which results to a lower Krippendorff's α for fluency. The intra-agreement for the four Yorùbá raters are 0.75, 0.91, 0.66, and 0.87, while the intra-agreement for the three English raters across all evaluation tasks are 0.92, 0.71, and 0.81.

For *yo–en*, our evaluators rated on average GMNMT to be the best in terms of adequacy (4.02 out of 5) and fluency (4.71), followed by mT5+Transfer, which shows that fine-tuning massively multilingual models also benefits low resource languages MT especially in terms of fluency (4.39). This contradicts the results of the automatic evaluation which prefers C4+Transfer+BT over mT5+Transfer.

For *en–yo*, GMNMT is still the best in terms of adequacy (3.69) followed by C4+Transfer+BT, but performs the worst in terms of fluency and diacritics prediction accuracy. So, the bad quality of the diacritics affects fluency and drastically penalises automatic metrics such as BLEU, but does not interfere with the intelligibility of the translations as shown by the good average adequacy rating. Automatic diacritic restoration for Yorùbá (Orife, 2018; Orife et al., 2020) can therefore be very useful to improve translation quality. C4+Transfer and C4+Transfer+BT perform similarly with high scores in terms of fluency and near perfect score in diacritics prediction accuracy (4.91 ± 0.1) as a result of being trained on cleaned corpora.

5 Related Work

In order to make MT available for a broader range of linguistic communities, recent years have seen an effort in creating new **parallel corpora** for low-resource language pairs. Recently, Guzmán et al. (2019) provided novel supervised, semi-supervised and unsupervised benchmarks for Indo-Aryan languages {Sinhala,Nepali}–English on an evaluation set of professionally translated sentences sourced from the Sinhala, Nepali and English Wikipedias.

Novel parallel corpora focusing on **African languages** cover South African languages ({Afrikaans, isiZulu, Northern Sotho, Setswana, Xitsonga}–English) (Groenewald and Fourie, 2009) with MT benchmarks evaluated in Martinus and Abbott (2019), as well as multidomain (News, Wikipedia, Twitter, Conversational) Amharic–English (Hadgu et al., 2020) and multidomain (Government, Wikipedia, News etc.) Igbo–English (Ezeani et al., 2020). Further, the LORELEI project (Strassel and Tracey, 2016) has created parallel corpora for a variety of low-resource language pairs, including a number of Niger-Congo languages such as {isiZulu, Twi, Wolof, Yorùbá }–English. However, these are not open-access. On the contrary, Masakhane (∀ et al., 2020) is an ongoing participatory project focusing on creating new freely-available parallel corpora and MT benchmark models for a large variety of African languages.

While creating parallel resources for low-resource language pairs is one approach to increase the number of linguistic communities covered by MT, this does not scale to the sheer amount of possible language combinations. Another research line focuses on **low-resource**

MT from the modeling side, developing methods which allow a MT system to learn the translation task with smaller amounts of supervisory signals. This is done by exploiting the weaker supervisory signals in larger amounts of available monolingual data, e.g. by identifying additional parallel data in monolingual corpora (Artetxe and Schwenk, 2019; Schwenk et al., 2021, 2020), comparable corpora (Ruiter et al., 2019, 2021), or by including auto-encoding (Currey et al., 2017) or language modeling tasks (Gulcehre et al., 2015; Ramachandran et al., 2017) during training. Low-resource language pairs can benefit from high-resource languages through transfer learning (Zoph et al., 2016), e.g. in a zero-shot setting (Johnson et al., 2017), by using pre-trained language models (Lample and Conneau, 2019), or finding an optimal path of pivoting through related languages (Leng et al., 2019). By adapting the model hyperparameters to the low-resource scenario, Sennrich and Zhang (2019) were able to achieve impressive improvements over a standard NMT system.

6 Conclusion

We present MENYO-20k, a novel *en–yo* multi-domain parallel corpus for machine translation and domain adaptation. By defining a standardized train-development-test split of this corpus, we provide several NMT benchmarks for future research on the *en–yo* MT task. Further, we analyze the domain differences on the MENYO-20k corpus and the translation performance of NMT models trained on religion corpora, such as JW300 and Bible, across the different domains. We show that, despite consisting of only $10k$ parallel sentences, adding the MENYO-20k corpus train split to JW300 and Bible largely improves the translation performance over all domains. Further, we train a variety of supervised, semi-supervised and fine-tuned MT benchmarks on available *en–yo* corpora, creating a high quality baseline that outperforms current massively multilingual models, e.g. Google MNMT by BLEU $+18.8$ (*en–yo*). This shows the positive impact of using smaller amounts of high-quality data (e.g. C4+Transfer, BLEU 12.4) that takes into account language-specific characteristics, i.e. diacritics, over massive amounts of noisy data (Facebook M2M-100, BLEU 3.3). Apart from having low BLEU scores, our human evaluation reveals that models trained on low-quality diacritics (Google MNMT) suffer especially in fluency, while still being intelligible to the reader. While correctly diacritized data is vital for translating *en–yo*, it only has an impact on the quality of *yo–en* translation quality when there is a domain mismatch between training and testing data.

Acknowledgements

We would like to thank Adebayo O. Adeojo, Babunde O. Popoola, Olumide Awokoya, Modupe Olaniyi, Princess Folasade, Akinade Idris, Tolulope Adelani, Oluyemisi Olaose, and Benjamin Ajibade for their support in translating English sentences to Yorùbá, verification of Yorùbá diacritics, and human evaluation. We thank Bayo Adebowale and 'Dele 'Adelani for donating their books ("Out of His Mind", and "Ojowu"). We thank Iroro Orife for providing the Bible corpus and Yorùbá Proverbs corpus. We thank Marine Carpuat, Mathias Müller, and the entire Masakhane NLP community for their feedback. We are also thankful to Damyana Gateva for evaluations with open-source models. This project was funded by the AI4D language dataset fellowship (Siminyu et al., 2021)[13]. DIA acknowledges the support of the EU-funded H2020 project COMPRISE under grant agreement No. 3081705. CEB is partially funded by the German Federal Ministry of Education and Research under the funding code 01IW20010. The authors are responsible for the content of this publication.

[13] https://www.k4all.org/project/language-dataset-fellowship/

References

Agić, Ž. and Vulić, I. (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.

Al-Rfou, R. (2015). *Polyglot: A massive multilingual natural language processing pipeline*. PhD thesis, Stony Brook University.

Alabi, J., Amponsah-Kaakyire, K., Adelani, D., and España-Bonet, C. (2020). Massive vs. curated embeddings for low-resourced languages: the case of Yorùbá and Twi. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2754–2762, Marseille, France. European Language Resources Association.

Arivazhagan, N., Bapna, A., Firat, O., Lepikhin, D., Johnson, M., Krikun, M., Chen, M. X., Cao, Y., Foster, G. F., Cherry, C., Macherey, W., Chen, Z., and Wu, Y. (2019). Massively multilingual neural machine translation in the wild: Findings and challenges. *arXiv e-prints 1907.05019*.

Artetxe, M. and Schwenk, H. (2019). Margin-based parallel corpus mining with multilingual sentence embeddings. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3197–3203, Florence, Italy. Association for Computational Linguistics.

Barrault, L., Bojar, O., Bougares, F., Chatterjee, R., Costa-jussà, M. R., Federmann, C., Fishel, M., Fraser, A., Graham, Y., Guzman, P., Haddow, B., Huck, M., Yepes, A. J., Koehn, P., Martins, A., Morishita, M., Monz, C., Nagata, M., Nakazawa, T., and Negri, M., editors (2020). *Proceedings of the Fifth Conference on Machine Translation*. Association for Computational Linguistics, Online.

Beyer, A., Kauermann, G., and Schütze, H. (2020). Embedding space correlation as a measure of domain similarity. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2431–2439, Marseille, France. European Language Resources Association.

Caswell, I., Kreutzer, J., Wang, L., Wahab, A., van Esch, D., Ulzii-Orshikh, N., Tapo, A., Subramani, N., Sokolov, A., Sikasote, C., Setyawan, M., Sarin, S., Samb, S., Sagot, B., Rivera, C., Rios, A., Papadimitriou, I., Osei, S., Ortiz Suárez, P. J., Orife, I., Ogueji, K., Niyongabo, R. A., Nguyen, T. Q., Müller, M., Müller, A., Hassan Muhammad, S., Muhammad, N., Mnyakeni, A., Mirzakhalov, J., Matangira, T., Leong, C., Lawson, N., Kudugunta, S., Jernite, Y., Jenny, M., Firat, O., Dossou, B. F. P., Dlamini, S., de Silva, N., Çabuk Ballı, S., Biderman, S., Battisti, A., Baruwa, A., Bapna, A., Baljekar, P., Abebe Azime, I., Awokoya, A., Ataman, D., Ahia, O., Ahia, O., Agrawal, S., and Adeyemi, M. (2021). Quality at a Glance: An Audit of Web-Crawled Multilingual Datasets. *arXiv e-prints*, page arXiv:2103.12028.

Currey, A., Miceli Barone, A. V., and Heafield, K. (2017). Copied monolingual data improves low-resource neural machine translation. In *Proceedings of the Second Conference on Machine Translation*, pages 148–156, Copenhagen, Denmark. Association for Computational Linguistics.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Eberhard, D. M., Simons, G. F., and (eds.), C. D. F. (2019). Ethnologue: Languages of the world. twenty-second edition.

España-Bonet, C., Barrón-Cedeño, A., and Màrquez, L. (2020). Tailoring and Evaluating the Wikipedia for in-Domain Comparable Corpora Extraction. *arXiv e-prints 2005.01177*, pages 1–26.

Ezeani, I., Rayson, P., Onyenwe, I., Chinedu, U., and Hepple, M. (2020). Igbo-english machine translation: An evaluation benchmark. In *Eighth International Conference on Learning Representations: ICLR 2020*.

Fan, A., Bhosale, S., Schwenk, H., Ma, Z., El-Kishky, A., Goyal, S., Baines, M., Celebi, O., Wenzek, G., Chaudhary, V., Goyal, N., Birch, T., Liptchinsky, V., Edunov, S., Grave, E., Auli, M., and Joulin, A. (2020). Beyond english–centric multilingual machine translation. *arXiv e-prints 2010.11125*.

∀, Nekoto, W., Marivate, V., Matsila, T., Fasubaa, T., Fagbohungbe, T., Akinola, S. O., Muhammad, S., Kabongo Kabenamualu, S., Osei, S., Sackey, F., Niyongabo, R. A., ..., Ogueji, K., Siminyu, K., Kreutzer, J., .., and Bashir, A. (2020). Participatory research for low-resourced machine translation: A case study in african languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2144–2160, Online. Association for Computational Linguistics.

Groenewald, H. J. and Fourie, W. (2009). Introducing the autshumato integrated translation environment. In *Proceedings of the 13th Annual conference of the European Association for Machine Translation*, pages 190–196, Barcelona, Spain. European Association for Machine Translation.

Gulcehre, C., Firat, O., Xu, K., Cho, K., Barrault, L., Lin, H.-C., Bougares, F., Schwenk, H., and Bengio, Y. (2015). On using monolingual corpora in neural machine translation. *arXiv preprint arXiv:1503.03535*.

Guzmán, F., Chen, P.-J., Ott, M., Pino, J., Lample, G., Koehn, P., Chaudhary, V., and Ranzato, M. (2019). The FLORES evaluation datasets for low-resource machine translation: Nepali–English and Sinhala–English. 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)*, pages 6098–6111, Hong Kong, China. Association for Computational Linguistics.

Hadgu, A. T., Beaudoin, A., and Aregawi, A. (2020). Evaluating Amharic Machine Translation. *arXiv e-prints 2003.14386*.

Heafield, K. (2011). KenLM: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland. Association for Computational Linguistics.

Johnson, M., Schuster, M., Le, Q. V., Krikun, M., Wu, Y., Chen, Z., Thorat, N., Viégas, F. B., Wattenberg, M., Corrado, G., Hughes, M., and Dean, J. (2017). Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation. *Transactions of the Association for Computational Linguist*, 5:339–351.

Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. In *International Conference for Learning Representations (ICLR)*.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Lample, G. and Conneau, A. (2019). Cross-lingual language model pretraining. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 32, pages 7059–7069. Curran Associates, Inc.

Leng, Y., Tan, X., Qin, T., Li, X.-Y., and Liu, T.-Y. (2019). Unsupervised pivot translation for distant languages. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 175–183.

Martinus, L. and Abbott, J. Z. (2019). A focus on neural machine translation for african languages. *arXiv e-prints 1906.05685*.

Orife, I., Adelani, D., Fasubaa, T. E., Williamson, V., Oyewusi, W. F., Wahab, O., and Túbosún, K. (2020). Improving Yorùbá Diacritic Restoration. *ArXiv*, abs/2003.10564.

Orife, I. F. d. (2018). Sequence-to-Sequence Learning for Automatic Yorùbá Diacritic Restoration. In *Proceedings of the Interspeech*, pages 27–35.

Ott, M., Edunov, S., Baevski, A., Fan, A., Gross, S., Ng, N., Grangier, D., and Auli, M. (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.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (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.

Popović, M. (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.

Ramachandran, P., Liu, P., and Le, Q. (2017). Unsupervised pretraining for sequence to sequence learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 383–391, Copenhagen, Denmark. Association for Computational Linguistics.

Resnik, P., Olsen, M. B., and Diab, M. T. (1999). The Bible as a Parallel Corpus: Annotating the 'Book of 2000 Tongues'. *Computers and the Humanities*, 33:129–153.

Ruiter, D., España-Bonet, C., and van Genabith, J. (2019). Self-supervised neural machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1828–1834, Florence, Italy. Association for Computational Linguistics.

Ruiter, D., Klakow, D., van Genabith, J., and España-Bonet, C. (2021). Integrating Unsupervised Data Generation into Self-Supervised Neural Machine Translation for Low-Resource Languages. In *Proceedings of Machine Translation Summit (Research Track)*. European Association for Machine Translation.

Schwenk, H., Chaudhary, V., Sun, S., Gong, H., and Guzmán, F. (2021). WikiMatrix: Mining 135M Parallel Sentences in 1620 Language Pairs from Wikipedia. In Merlo, P., Tiedemann, J., and Tsarfaty, R., editors, *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pages 1351–1361. Association for Computational Linguistics.

Schwenk, H., Wenzek, G., Edunov, S., Grave, E., and Joulin, A. (2020). Ccmatrix: Mining billions of high-quality parallel sentences on the web. *arXiv e-prints arXiv:1911.04944*.

Sennrich, R., Haddow, B., and Birch, A. (2016). Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725.

Sennrich, R. and Zhang, B. (2019). Revisiting low-resource neural machine translation: A case study. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 211–221, Florence, Italy. Association for Computational Linguistics.

Siminyu, K., Kalipe, G., Orlic, D., Abbott, J., Marivate, V., Freshia, S., Sibal, P., Neupane, B., Adelani, D., Taylor, A., Ali, J. T., Degila, K., Balogoun, M., Diop, T. I., David, D., Fourati, C., Haddad, H., and Naski, M. (2021). Ai4d - african language program. *ArXiv*, abs/2104.02516.

Strassel, S. and Tracey, J. (2016). LORELEI language packs: Data, tools, and resources for technology development in low resource languages. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 3273–3280, Portorož, Slovenia. European Language Resources Association (ELRA).

Tang, Y., Tran, C., Li, X., Chen, P., Goyal, N., Chaudhary, V., Gu, J., and Fan, A. (2020). Multilingual translation with extensible multilingual pretraining and finetuning. *ArXiv*, abs/2008.00401.

Tiedemann, J. and Thottingal, S. (2020). OPUS-MT – building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Lisboa, Portugal. European Association for Machine Translation.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., and Raffel, C. (2021). 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.

Zoph, B., Yuret, D., May, J., and Knight, K. (2016). Transfer learning for low-resource neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1575, Austin, Texas. Association for Computational Linguistics.