The Flores-101 Evaluation Benchmark for Low-Resource and Multilingual Machine Translation

Naman Goyal[1], Cynthia Gao[1], Vishrav Chaudhary[1], Peng-Jen Chen[1],
Guillaume Wenzek[2], Da Ju[1], Sanjana Krishnan[1], Marc'Aurelio Ranzato[1],
Francisco Guzmán[1], Angela Fan[2,3]
[1]Facebook AI Research, USA, [2]Facebook AI Research, France, [3]LORIA
flores@fb.com

Abstract

One of the biggest challenges hindering progress in low-resource and multilingual machine translation is the lack of good evaluation benchmarks. Current evaluation benchmarks either lack good coverage of low-resource languages, consider only restricted domains, or are low quality because they are constructed using semi-automatic procedures. In this work, we introduce the Flores-101 evaluation benchmark, consisting of 3001 sentences extracted from English Wikipedia and covering a variety of different topics and domains. These sentences have been translated in 101 languages by professional translators through a carefully controlled process. The resulting dataset enables better assessment of model quality on the long tail of low-resource languages, including the evaluation of many-to-many multilingual translation systems, as all translations are fully aligned. By publicly releasing such a high-quality and high-coverage dataset, we hope to foster progress in the machine translation community and beyond.

1 Introduction

Machine translation (MT) is one of the most successful applications in natural language processing, as exemplified by its numerous practical applications and the number of contributions on this topic at major machine learning and natural language processing venues. Despite recent advances in translation quality for a handful of language pairs and domains, MT systems still perform poorly on *low-resource languages*, that is, languages without a lot of training data. In fact, many low-resource languages are not even supported by most popular translation engines. Yet, much of the world's population speak low-resource languages and would benefit from improvements in translation quality on their native languages. As a result, the field has been increasing focus towards low-resource languages.

At present, there are very few benchmarks on low-resource languages. These often have very low coverage of low-resource languages (Riza et al., 2016; Thu et al., 2016; Guzmán et al., 2019; Barrault et al., 2020a; Nekoto et al., 2020; Ebrahimi et al., 2021; Kuwanto et al., 2021), limiting our understanding of how well methods generalize and scale to a larger number of languages with a diversity of linguistic features. There are some benchmarks that have high coverage, but these are often in specific domains, like COVID-19 (Anastasopoulos et al., 2020) or religious texts (Christodouloupoulos and Steedman, 2015; Malaviya et al., 2017; Tiedemann, 2018; Agić and Vulić, 2019); or have low quality because they are built using automatic approaches (Zhang et al., 2020; Schwenk et al., 2019, 2021). As a result, it is difficult to draw firm conclusions about research efforts on low-resource MT. In particular, there are even fewer benchmarks that are suitable for evaluation of many-to-many multilingual translation, as these require multi-lingual alignment (i.e., having the translation of the same sentence in multiple languages), which hampers the progress of the field despite all the recent excitement on this research direction.

We present the Flores-101 benchmark, consisting of 3001 sentences sampled from various topics in English Wikipedia and professionally translated in 101 languages. With this dataset, we make several contributions. First, we provide the community with a high-quality benchmark that has much larger breadth of topics and coverage of low resource languages than any other existing dataset (§4). Second, Flores-101 is suitable for many-to-many evaluation, meaning that it enables seamless evaluation of 10,100 language pairs. This enables the evaluation of popular multilingual MT systems as well as the evaluation of regionally-relevant language pairs like Spanish-Aymara and Vietnamese-Thai, for example. Third, we

Transactions of the Association for Computational Linguistics, vol. 10, pp. 522–538, 2022. https://doi.org/10.1162/tacl_a_00474
Action Editor: Colin Cherry. Submission batch: 7/2021; Revision batch: 11/2021; Published 5/2022.

thoroughly document the annotation process we followed (§3), helping the community build institutional knowledge about how to construct MT datasets. Fourth, we release not only sentences with their translation but also rich meta-data that enables other kinds of evaluations and tasks, such as document level translation, multimodal translation, and text classification. Fifth, we propose to use the BLEU metric based on sentence piece tokenization (Kudo and Richardson, 2018) (§5) to enable evaluation of all languages in the set in a unified and extensible framework, while preserving the familiarity of BLEU. Finally, we publicly release both data and baselines used in our experiments (§6), to foster research in low-resource machine translation and related areas.

2 Related Work

A major challenge in machine translation, particularly as the field shifts its focus to low-resource languages, is the lack of availability of evaluation benchmarks. Much recent work has focused on the creation of training corpora (Auguste Tapo et al., 2021; Ali et al., 2021; Adelani et al., 2021; Gezmu et al., 2021; Nyoni and Bassett, 2021; Chauhan et al., 2021) and development of models (Koneru et al., 2021; Nagoudi et al., 2021; Aulamo et al., 2021), but evaluation is critical to being able to assess and improve translation quality.

Traditionally, the yearly Workshop on Machine Translation (WMT) and its associated shared tasks have provided standardized benchmarks and metrics to the community, fostering progress by providing means of fair comparison among various approaches. Over recent years, the main translation task at WMT has challenged participants with low-resource languages, but the evaluation has been limited to a handful of languages—for example, Latvian in 2017 (Bojar et al., 2017), Kazakh in 2018 (Bojar et al., 2018), Gujarati and Lithuanian in 2019 (Barrault et al., 2019), and Inuktitut, Khmer, Pashto, and Tamil in 2020 (Barrault et al., 2020b). Moreover, these tasks have considered translation to and from English only, while the field has been recently focusing on large-scale multilingual models (Johnson et al., 2016; Aharoni et al., 2019; Freitag and Firat, 2020; Fan et al., 2020).

There are other datasets for evaluation purposes, such as Flores v1.0 (Guzmán et al., 2019), LORELEI (Strassel and Tracey, 2016), ALT (Thu et al., 2016; Riza et al., 2016; Ding et al., 2016), and TICO-19 (Anastasopoulos et al., 2020), as well as datasets for specific languages such as Igbo (Ezeani et al., 2020) and Fon (Dossou and Emezue, 2020). These are similar to Flores-101 because they focus on low-resource languages. However, the language coverage of these datasets is much smaller. Among these, only TICO-19 is suitable for multilingual machine translation, but its content is centered around COVID-19, unlike the much broader coverage of topics offered by Flores-101. The Tatoeba corpus (Tiedemann, 2020) covers a large number of languages and translation directions, but the low volume of sentences for many directions makes the evaluation less reliable. Further, many of the Tatoeba corpus sentences are very short and straightforward, making the dataset not as generalizable to a more diverse set of content.

3 Dataset Construction

We describe how Flores-101 was constructed, first noting where the sentences originated from and subsequently describing the carefully designed translation process.

3.1 Sourcing Sentences

Original Source. All source sentences were extracted from multiple Wikimedia sources, as this is a repository of text that is public and freely available under permissive licensing, and covers a broad range of topics. Although Wikipedia currently supports more than 260 languages,[1] several low-resource languages have relatively few articles containing well structured sentences. Moreover, translating a few hundred sentences for several thousand different language pairs would be infeasible, at the very least because of the lack of qualified translators that can read both the source and target.

Instead, we opted to source all sentences from three locations in English Wikimedia, while considering a broad set of topics that could be of general interest regardless of the native language of the reader. In particular, we collected a third of the sentences from *Wikinews*,[2] which is a collection of international news articles, a third

[1] https://en.wikipedia.org/wiki/Wikipedia:Multilingual_statistics.

[2] https://en.wikinews.org/wiki/Main_Page.

from *Wikijunior*,[3] which is a collection of age-appropriate nonfiction books for children from birth to age 12, and a third from *WikiVoyage*,[4] which is a travel guide with a collection of articles about travel tips, food, and destinations around the globe. By translating the same set of English sentences in more than hundred languages, we enable evaluation of multilingual MT with the caveat that *source* sentences not in English are produced by human translators. While translationese (or overly literal or awkward translations) has known idiosyncrasies (Zhang and Toral, 2019), we conjecture that these effects are rather marginal when evaluating models in low-resource languages, where current MT systems produce many severe mistakes. Another downside to this English-centric translation approach is a possible artificial increase of the differences between two dialects of the same language. For example, Catalan translations of Spanish sentences are very close to the original source sentences. However, translating from English to Spanish and English to Catalan can produce Spanish and Catalan sentences that are no longer so similar. We believe the benefits of many-to-many evaluation, which supports the measurement of traditionally neglected regionally relevant pairs such as Xhosa-Zulu, Vietnamese-Thai, and Spanish-Aymara, largely outsize the risk of evaluating translationese.

Sentence Selection. The sentence process consisted of selecting an article at random from each domain, and then selecting between 3 and 5 contiguous sentences, avoiding segments with very short or malformed sentences. We selected one paragraph per document, from either the beginning, middle, or end of the article. For each sentence, we extracted the URL, topic, and noted Boolean flags to indicate whether the sentence contained entities linked to other Wikipedia pages and images.

Several contiguous sentences are extracted from the same article and we also provide the corresponding URL. Additional document-level context can be accessed through this provided metadata when translating each sentence. On average, we select 3.5 contiguous sentences per article, providing assessment possibilities beyond single sentences. However, we note that compared to the document-level evaluation datasets in WMT 2018 to 2020 for Russian and German, FLORES-101 does not contain translations of full documents. On the other hand, FLORES-101 covers a much wider array of domains and topics, facilitated by translating a far greater number of articles and fewer sentences per article. Further, the metadata for each sentence in FLORES-101 is provided, and thus the full English document could be used in studies of document-level translation. Overall, we find this a reasonable compromise between evaluating beyond sentence-level context while creating a diverse evaluation dataset.

Finally, with the additional meta-data provided in FLORES-101, we also enable evaluation of multimodal machine translation as users can access images through the metadata. Around two-thirds of all articles chosen for translation contain images (see Table 3), allowing the incorporation of both text and image content for evaluation.

3.2 Translation Guidelines

We describe how translators were selected and detail the guidelines provided to translators.

Translator Selection Translators are required to be native speakers and educated in the target language. They must also have a high level of fluency (C1-C2) in English. Translators are required to have at least two to three years of translation experience in the relevant language pair if they have an academic degree in translation or linguistics and three to five years of translation experience if they do not have any relevant academic qualification. Translators are also required to to be an experienced, generalist translator and/or familiar with United States and international news, current affairs, politics, sports, and so forth. Translators also undergo a translation test every 18 months to assess their translation quality. In addition to having the necessary translation skills, FLORES translators must be able to communicate effectively in English.

Instructions for Translation Translators were instructed to translate source data as informative, neutral, and standardized content. Assistance from any machine translation was strictly prohibited and translators were advised to translate localized in the target language as appropriate for content in the informative domain. Particular guidance was made on translating named entities, in which proper nouns were to be translated

[3] https://en.wikibooks.org/wiki/Wikijunior.

[4] https://en.wikivoyage.org/wiki/Main_Page.

ISO 639-3	Language	Family	Subgrouping	Script	Bitext w/ En	Mono Data
afr	**Afrikaans**	Indo-European	Germanic	Latin	570K	26.1M
amh	**Amharic**	Afro-Asiatic	Afro-Asiatic	Ge’ez	339K	3.02M
ara	**Arabic**	Afro-Asiatic	Afro-Asiatic	Arabic	25.2M	126M
hye	**Armenian**	Indo-European	Other IE	Armenian	977K	25.4M
asm	**Assamese**	Indo-European	Indo-Aryan	Bengali	43.7K	738K
ast	**Asturian**	Indo-European	Romance	Latin	124K	—
azj	**Azerbaijani**	Turkic	Turkic	Latin	867K	41.4M
bel	**Belarusian**	Indo-European	Balto-Slavic	Cyrillic	42.4K	24M
ben	**Bengali**	Indo-European	Indo-Aryan	Bengali	2.16M	57.9M
bos	**Bosnian**	Indo-European	Balto-Slavic	Latin	187K	15.9M
bul	**Bulgarian**	Indo-European	Balto-Slavic	Cyrillic	10.3M	235M
mya	**Burmese**	Sino-Tibetan	Sino-Tibetan+Kra-Dai	Myanmar	283K	2.66M
cat	**Catalan**	Indo-European	Romance	Latin	5.77M	77.7M
ceb	**Cebuano**	Austronesian	Austronesian	Latin	484K	4.11M
zho	**Chinese (Simpl)**	Sino-Tibetan	Sino-Tibetan+Kra-Dai	Han	37.9M	209M
zho	**Chinese (Trad)**	Sino-Tibetan	Sino-Tibetan+Kra-Dai	Han	37.9M	85.2M
hrv	**Croatian**	Indo-European	Balto-Slavic	Latin	42.2K	144M
ces	**Czech**	Indo-European	Balto-Slavic	Latin	23.2M	124M
dan	**Danish**	Indo-European	Germanic	Latin	10.6M	344M
nld	**Dutch**	Indo-European	Germanic	Latin	82.4M	230M
est	**Estonian**	Uralic	Uralic	Latin	4.82M	46M
tgl	**Filipino (Tagalog)**	Austronesian	Austronesian	Latin	70.6K	107M
fin	**Finnish**	Uralic	Uralic	Latin	15.2M	377M
fra	**French**	Indo-European	Romance	Latin	289M	428M
ful	**Fula**	Atlantic-Congo	Nilotic+Other AC	Latin	71K	531K
glg	**Galician**	Indo-European	Romance	Latin	1.13M	4.22M
lug	**Ganda**	Atlantic-Congo	Bantu	Latin	14.4K	537K
kat	**Georgian**	Kartvelian	Other	Georgian	1.23M	31.7M
deu	**German**	Indo-European	Germanic	Latin	216M	417M
ell	**Greek**	Indo-European	Other IE	Greek	23.7M	201M
guj	**Gujarati**	Indo-European	Indo-Aryan	Gujarati	160K	9.41M
hau	**Hausa**	Afro-Asiatic	Afro-Asiatic	Latin	335K	5.87M
heb	**Hebrew**	Afro-Asiatic	Afro-Asiatic	Hebrew	6.64M	208M
hin	**Hindi**	Indo-European	Indo-Aryan	Devanagari	3.3M	104M
hun	**Hungarian**	Uralic	Uralic	Latin	16.3M	385M
isl	**Icelandic**	Indo-European	Germanic	Latin	1.17M	37.5M
ibo	**Igbo**	Atlantic-Congo	Nilotic+Other AC	Latin	145K	693K
ind	**Indonesian**	Austronesian	Austronesian	Latin	39.1M	1.05B
gle	**Irish**	Indo-European	Other IE	Latin	329K	1.54M
ita	**Italian**	Indo-European	Romance	Latin	116M	179M
jpn	**Japanese**	Japonic	Other	Han, Hiragana, Katakana	23.2M	458M
jav	**Javanese**	Austronesian	Austronesian	Latin	1.49M	24.4M
kea	**Kabuverdianu**	Indo-European	Romance	Latin	5.46K	178K
kam	**Kamba**	Atlantic-Congo	Bantu	Latin	50K	181K
kan	**Kannada**	Dravidian	Dravidian	Telugu-Kannada	155K	13.1M
kaz	**Kazakh**	Turkic	Turkic	Cyrillic	701K	35.6M
khm	**Khmer**	Austro-Asiatic	Austro-Asiatic	Khmer	398K	8.87M
kor	**Korean**	Koreanic	Other	Hangul	7.46M	390M
kir	**Kyrgyz**	Turkic	Turkic	Cyrillic	566K	2.02M
lao	**Lao**	Kra-Dai	Sino-Tibetan+Kra-Dai	Lao	153K	2.47M
lav	**Latvian**	Indo-European	Balto-Slavic	Latin	4.8M	68.4M
lin	**Lingala**	Atlantic-Congo	Bantu	Latin	21.1K	336K
lit	**Lithuanian**	Indo-European	Balto-Slavic	Latin	6.69M	111M
luo	**Luo**	Nilo-Saharan	Nilotic+Other AC	Latin	142K	239K
ltz	**Luxembourgish**	Indo-European	Germanic	Latin	3.41M	—
mkd	**Macedonian**	Indo-European	Balto-Slavic	Cyrillic	1.13M	28.8M
msa	**Malay**	Austronesian	Austronesian	Latin	968K	77.5M
mal	**Malayalam**	Dravidian	Dravidian	Malayalam	497K	24.8M
mlt	**Maltese**	Afro-Asiatic	Afro-Asiatic	Latin	5.82M	—
mri	**Māori**	Austronesian	Austronesian	Latin	196K	—

ISO 639-3	Language	Family	Subgrouping	Script	Bitext w/ En	Mono Data
mar	**Marathi**	Indo-European	Indo-Aryan	Devanagari	109K	14.4M
mon	**Mongolian**	Mongolic	Other	Cyrillic	555K	20.4M
npi	**Nepali**	Indo-European	Indo-Aryan	Devanagari	19.6K	17.9M
nso	**Northern Sotho**	Atlantic-Congo	Bantu	Latin	13.8K	612K
nob	**Norwegian**	Indo-European	Germanic	Latin	10.9M	338M
nya	**Nyanja**	Atlantic-Congo	Bantu	Latin	932K	—
oci	**Occitan**	Indo-European	Romance	Latin	5.11K	—
ory	**Oriya**	Indo-European	Indo-Aryan	Oriya	5K	2.47M
orm	**Oromo**	Afro-Asiatic	Afro-Asiatic	Latin	162K	752K
pus	**Pashto**	Indo-European	Indo-Aryan	Perso-Arabic	293K	12M
fas	**Persian**	Indo-European	Indo-Aryan	Perso-Arabic	6.63M	611M
pol	**Polish**	Indo-European	Balto-Slavic	Latin	40.9M	256M
por	**Portuguese (Brazil)**	Indo-European	Romance	Latin	137M	340M
pan	**Punjabi**	Indo-European	Indo-Aryan	Gurmukhi	142K	5.02M
ron	**Romanian**	Indo-European	Romance	Latin	31.9M	391M
rus	**Russian**	Indo-European	Balto-Slavic	Cyrillic	127M	849M
srp	**Serbian**	Indo-European	Balto-Slavic	Cyrillic	7.01M	35.7M
sna	**Shona**	Atlantic-Congo	Bantu	Latin	877K	—
snd	**Sindhi**	Indo-European	Indo-Aryan	Perso-Arabic	21.8K	314K
slk	**Slovak**	Indo-European	Balto-Slavic	Latin	10.5M	174M
slv	**Slovenian**	Indo-European	Balto-Slavic	Latin	5.42M	74.7M
som	**Somali**	Afro-Asiatic	Afro-Asiatic	Latin	358K	14.1M
ckb	**Sorani Kurdish**	Indo-European	Indo-Aryan	Arabic	305K	7.98M
spa	**Spanish (Latin America)**	Indo-European	Romance	Latin	315M	379M
swh	**Swahili**	Atlantic-Congo	Bantu	Latin	349K	35.8M
swe	**Swedish**	Indo-European	Germanic	Latin	54.8M	580M
tgk	**Tajik**	Indo-European	Indo-Aryan	Cyrillic	544K	—
tam	**Tamil**	Dravidian	Dravidian	Tamil	992K	68.2M
tel	**Telugu**	Dravidian	Dravidian	Telugu-Kannada	381K	17.2M
tha	**Thai**	Kra-Dai	Sino-Tibetan+Kra-Dai	Thai	10.6M	319M
tur	**Turkish**	Turkic	Turkic	Latin	41.2M	128M
ukr	**Ukrainian**	Indo-European	Balto-Slavic	Cyrillic	5.44M	357M
umb	**Umbundu**	Atlantic-Congo	Bantu	Latin	217K	142K
urd	**Urdu**	Indo-European	Indo-Aryan	Perso-Arabic	630K	28M
uzb	**Uzbek**	Turkic	Turkic	Latin	—	7.54M
vie	**Vietnamese**	Austro-Asiatic	Austro-Asiatic	Latin	32.1M	992M
cym	**Welsh**	Indo-European	Other IE	Latin	826K	12.7M
wol	**Wolof**	Atlantic-Congo	Nilotic+Other AC	Latin	86.9K	676K
xho	**Xhosa**	Atlantic-Congo	Bantu	Latin	130K	995K
yor	**Yoruba**	Atlantic-Congo	Nilotic+Other AC	Latin	171K	1.59M
zul	**Zulu**	Atlantic-Congo	Bantu	Latin	123K	994K

Table 1: **101 Languages in Flores-101.** We include the ISO 639-3 code, the language family, and script. Next to each language family, we include more fine-grained subgrouping information. We also include the amount of resources available in OPUS (for bitext with English) and cc100 (for monolingual data) at the time this report was written. The parallel datasets were used to train the baseline described in §5, the monolingual datasets were only used to calculate SentencePiece, see Section §5.

in the most commonly used form in the target language and when such equivalent terms did not exist, transliteration in the target language was advised. Translators were also advised to translate abbreviations and idiomatic expressions to their best knowledge for how these terms and phrases usually appear in the target language, finding equivalents rather than literal word-for-word translations. Gender neutral pronouns were also advised to be used when 3rd person pronouns are ambiguous in the source text.

3.3 Translation and Evaluation

Obtaining high translation quality in low-resource languages is difficult because the translation job relies on the skill of a small set of translators. If one translator is not perfectly fluent or uses a different local convention for that language, this could

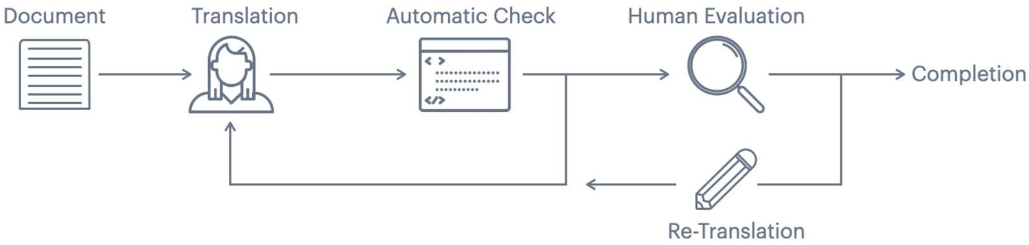

Figure 1: **Depiction of Overall Translation Workflow.**

render the quality of the dataset insufficient or inconsistent for that language. Here, we describe the process we followed with our Language Service Providers (LSPs) for translation and evaluation.

Translation Quality Score. How do we know if the translations are good enough to include in FLORES-101, and how do we know when a language has completed translation? Twenty percent of the dataset is sampled and reviewed, this is the same set of sentences across all languages, which allows us to compare quality. Each sentence-translation pair in the sampled data is assessed by a language-specific reviewer. We assess quality through a *Translation Quality Score* per language on a 0 to 100 scale, determined based on the number of identified errors by the evaluation LSPs. The following errors are examined: grammar, punctuation, spelling, capitalization, addition or omission of information, mistranslation, unnatural translation, untranslated text, and register. Each error is also associated with a severity level:minor, major or critical. Based on pilots, we set the acceptable score to 90%.

Translation Workflow. The overall translation workflow is depicted in Figure 1. For each language, all source sentences are sent to a certain translation LSP. Once sentences are translated, the data is sent to different translators within the LSP for editing and then moves on to automated quality control steps. If any of the checks fail, the LSP has to re-translate until all verification is passed. Afterwards, translations are sent to an evaluation LSP that performs quality assessment, providing a translation quality score and constructive linguistic feedback both on the sentence and language levels. If the score is below 90%, translations together with the assessment report are sent back to the translation LSP for re-translation. Languages scoring above our quality threshold of 90% have an average 15% of the reviewed sample with major or critical errors, 2% of the reviewed sample with critical errors, and 3% with unnatural translation errors. We summarize in Table 4 the overall statistics around the translation process.

4 FLORES-101 At a Glance

In this section, we analyze FLORES-101. We provide a high-level comparison of FLORES-101 with existing benchmarks, then discuss the sentences, languages, and translation quality in detail.

4.1 Comparison with Existing Benchmarks

We compare FLORES-101 with several existing benchmarks, summarized in Table 2. FLORES-101 combines large language coverage with topic diversity, support for many-to-many evaluation, and high quality human translations (e.g., produced with no automatic alignment). Further, FLORES-101 adds document-level evaluation and support multimodal translation evaluation through provided metadata.

4.2 Sentences in FLORES-101

Table 3 provides an overview of FLORES-101. The total dataset translates 3001 sentences into 101 languages. On average, sentences contain around 20 words. These sentences originate from 1,175 different articles in three domains: WikiNews, WikiJunior, and WikiVoyage. On average, 3 sentences are selected from each document, and then documents are divided into dev, devtest, and test sets. The articles are rich in metadata: 40% of articles contain hyperlinks to other pages, and 66% of articles contain images. We manually classify the content of the sentences into one of 10 broader topics, and display the distribution. Overall, most sentences are about world travel (sourced from WikiVoyage), though there are also a large number of sentences about science, politics, and crime.

	# Languages	Diverse Topics	Many to Many	Manual Alignments	Document Level	Multi modal
FLORES v1 (Guzmán et al., 2019)	2	✓	✗	✓	✗	✗
AmericasNLI (Ebrahimi et al., 2021)	10	✓	✓	✓	✗	✗
ALT (Riza et al., 2016)	13	✓	✓	✓	✗	✗
Europarl (Koehn, 2005)	21	✗	✓	✗	✓	✗
TICO-19 (Anastasopoulos et al., 2020)	36	✗	✓	✓	✗	✗
OPUS-100 (Zhang et al., 2020)	100	✓	✓	✗	✗	✗
M2M (Fan et al., 2020)	100	✗	✓	✓✗	✗	✗
FLORES-101	101	✓	✓	✓	✓	✓

Table 2: **Comparison of Various Evaluation Benchmarks**. We compare FLORES-101 to a variety of popular, existing translation benchmarks, indicating language coverage, topic diversity, whether many-to-many translation is supported, if the translations are manually aligned by humans, and if the tasks of document-level translation or multimodal translation are supported.

Number of Sentences		3001
Average Words per Sentence		21
Number of Articles		842
Average Number of Sentences per Article		3.5
% of Articles with Hyperlinked Entities		40
% of Articles with Images		66
Evaluation Split	**# Articles**	**# Sentences**
dev	281	997
devtest	281	1012
test	280	992
Domain	**# Articles**	**# Sentences**
WikiNews	309	993
WikiJunior	284	1006
WikiVoyage	249	1002
Sub-Topic	**# Articles**	**# Sentences**
Crime	155	313
Disasters	27	65
Entertainment	28	68
Geography	36	86
Health	27	67
Nature	17	45
Politics	171	341
Science	154	325
Sports	154	162
Travel	505	1529

Table 3: **Statistics of FLORES-101.**

# of Languages requiring Re-translation	45
Avg # of Re-translations	1
Max # of Re-translations	3
Avg # of Days to Translate 1 language	26
Avg # of Days to Re-Translate	35
Avg # of Days for 1 language	61
Shortest Turnaround (days) for 1 language	31
Longest Turnaround (days) for 1 language	89

Table 4: **Statistics of FLORES-101 Translation Workflow.**

4.3 Languages in FLORES-101

We summarize all languages in FLORES-101 in Table 1. Our selected languages cover a large percentage of people globally, with a large diversity of scripts and families. Many languages are spoken by millions, despite being considered low-resource in the research community. In Table 1, we estimate resource level by reporting the amount of data available in OPUS, a public repository for multilingual data. The majority of languages have bilingual data through English and monolingual data, though a number of languages have less than 100K sentences through English. Many have no monolingual data available, making them truly low-resource.

4.4 Translation Quality

The translation quality score across all languages is depicted in Figure 2. All 101 languages in FLORES-101 meet our threshold of 90% quality. Overall, about 50% of languages have fairly high quality (above 95%), with few near the 90% threshold boundary. Even low-resource languages like Lao and Zulu can score well on the quality metric. The largest error category across all languages was *mistranslation*, a broad category that generally notes that the source text was not translated faithfully and the translation has rendered an incorrect meaning in the target language. Error categories with few errors include register, grammar, and punctuation.

5 Multilingual Evaluation

Automatic evaluation of translation quality is an active research field. Each year, the WMT

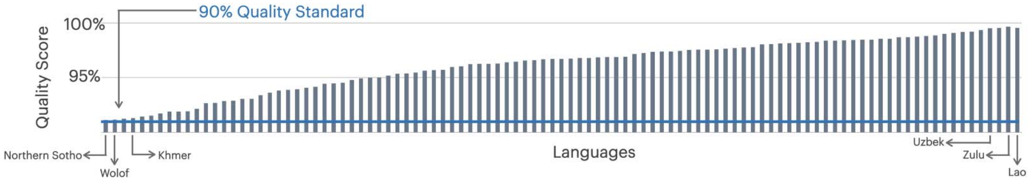

Figure 2: **Translation Quality Score across Languages.** We require the final translation quality score to be above 90% before the translation is of sufficient quality to include in FLORES-101.

Metrics shared task seeks to determine the metrics that better correlate with human evaluations (Mathur et al., 2020). While many metrics have been proposed through the years, most have not gained traction in the community.[5] In fact, 99% of MT publications in the last decade still report BLEU, and 74% do so exclusively (Marie et al., 2021). Through the years, researchers continue to use BLEU to compare different models. As a result, the community has developed strong intuitions on the significance of the effects when looking at BLEU. Unfortunately, using word-level BLEU *as-is* is suboptimal in a multilingual context, as *n*-gram overlap heavily depends on the particular tokenization used, which is not well defined in many low-resource languages.

5.1 The Challenge of Multilingual Evaluation

Making BLEU comparable by using equivalent tokenization schemes has been challenging for the translation community. It has been partially addressed by `sacrebleu` (Post, 2018) which allows specifying evaluation *signatures* that take tokenization into account. For example, `sacre-bleu` uses the standardized NIST tokenizer[6] as a default.

However, the picture is not so simple when looking into multilingual evaluation. For instance, some languages like Hindi and Japanese already have custom tokenizers that are used when computing BLEU, although these appear scattered in various publication footnotes. For many other languages, tokenizers do not exist and English rules are applied as a default. While English tokenization rules might operate reasonably well for European languages, they do not extend to global support. For example, white-space tokenization is insufficient for some languages like Burmese or Khmer, which do not segment words with white space. Other languages like Arabic are morphologically rich, and encode more meaning into a single word through the use of clitics.[7] In short, there is no standard for universal tokenization and developing tokenizers for each language of interest is a challenging effort (Dossou and Emezue, 2021; Li et al., 2021) that is difficult to scale.

Ideally, we would like an automatic evaluation process that is robust, simple, and applicable to any language without the need to specify any particular tokenizer, as this will make it easier for researchers to compare against each other. We would like our automatic evaluation to also support future languages—as translation quality continues to improve, the community will naturally produce models for more and more languages.

5.2 SentencePiece BLEU

Towards this goal, we propose to use BLEU over text tokenized with a single language-agnostic and publicly available fixed *SentencePiece* subword model. We call this evaluation method spBLEU, for brevity. It has the benefit of continuing to use a metric that the community is familiar with, while addressing the proliferation of tokenizers.

For this, we have trained a SentencePiece (SPM) tokenizer (Kudo and Richardson, 2018) with 256,000 tokens using monolingual data (Conneau et al., 2020; Wenzek et al., 2019) from all the FLORES-101 languages. SPM is a system that learns subword units based on training data, and does not require tokenization. The logic is not dependent on language, as the system treats all sentences as sequences of Unicode. Given the large amount of multilingual data and the large number of languages, this essentially provides a *universal* tokenizer, that can operate on any language.

[5] In the past 10 years, only RIBES and chrF++ have been used more than twice in MT publications (Marie et al., 2021).

[6] `https://bit.ly/3CoGWma`.

[7] This has incentivized the creation of BLEU variants (Bouamor et al., 2014) that rely on morphological analyzers.

Training SPM. One challenge is that the amount of monolingual data available for different languages is not the same—an effect that is extreme when considering low-resource languages. Languages with small quantities of data may not have the same level of coverage in subword units, or an insufficient quantity of sentences to represent a diverse enough set of content.

To address the low resource languages, first we extend monolingual data of lowest 80 resource languages to 60 Common Crawl snapshots. We then perform temperature sampling (Arivazhagan et al., 2019) with temperature = 5.0 so that low-resource languages are well represented. The SPM model is trained on a combined total of 100M sampled sentences based on the temperature sampling probability mentioned above. We use a character coverage value of 0.9995 following Fan et al. (2020) to have sufficient representation of character-based languages. For Flores-101 languages, the max unknown token rate with our SPM model is 3.8% for Tagalog, with all other languages below 1%, indicating good coverage for low resource languages from the trained tokenizer. In the future if a new language is added to Flores-101 and this tokenizer does not support its script, we can add new tokens to encode it as desired.

Computing spBLEU. Given this SPM-tokenizer, we compute BLEU by tokenizing the system output and the reference, and then calculate BLEU in the space of sentence-pieces. SpBLEU is integrated into `sacrebleu` for ease of use[8] as the `spm` tokenizer.

5.3 Experiments and Analysis

In this section we evaluate spBLEU to understand its properties. Particularly, we want to verify that it preserves the intuitions that researchers have built over years using BLEU. To do so, we contrast its results to languages where `mosestokenizer` is used as the default; and when custom tokenizers are used. Secondly, we verify that spBLEU offers similar results to other metrics that are tokenizer-independent such as chrF++ (Popović, 2017).

Note that a rigorous assessment of any automatic metric requires the measurement of correlation with respect to human evaluations. However, to date, such annotations are only available for a handful of languages, most of which are high-resource. Obtaining human evaluations for a large proportion of languages covered in FLORES101 is costly and time-consuming. Moreover, the conclusions of a partial study focused on high-resource languages might not generalize well to other languages. Therefore, we defer the in-depth evaluation of the best metric for multilingual evaluation to future work, once the adequate data is available.

Lang	Correlation spBLEU v. BLEU	Correlation spBLEU v. chrF++
French	0.99	0.98
Italian	0.99	0.98
Spanish	0.99	0.98
Hindi	0.99	0.98
Tamil	0.41	0.94
Chinese	0.99	0.98

Table 5: **Spearman Correlation of spBLEU, BLEU, and chrF++.** We evaluate on three sets of languages (En-XX). Models evaluated are derived from our baselines (discussed in Section 6). In the top section, we evaluate languages that often use the standard `mosestokenizer`. In the bottom section, we evaluate languages that have their own custom tokenization.

spBLEU Correlates with BLEU. First, we examine the correlation between spBLEU and BLEU across various languages where the `mosestokenizer` is widely used. We examine Spanish, Italian, and French. As shown in Table 5 (top), spBLEU correlates well (0.99) with BLEU on these languages. In the bottom section, we show the correlation for Chinese, Hindi, and Tamil where custom tokenizers are needed.[9] We observe a high correlation (0.99) for Hindi and Chinese, and a weaker correlation (0.41) for Tamil.[10]

spBLEU Has a Strong Correlation with chrF++. Next, we compare how well spBLEU tracks against another well vetted tokenizer-independent metric: chrF++. In Table 5 we observe that both metrics are highly correlated across target languages ranging from 0.94 to 0.98. This is consistent across all the 101×100 language pairs

[8] `https://github.com/mjpost/sacrebleu`.

[9] Chinese is supported by `mosetokenizer` with special rules. Hindi and Tamil have a popularly used tokenizer in the community from IndicNLP (`https://anoopkunchukuttan.github.io/indic_nlp_library/`).

[10] We hypothesize that a weaker correlation for Tamil might be due to the agglutinative nature of the language, which spBLEU handles more similarly to chrF++.

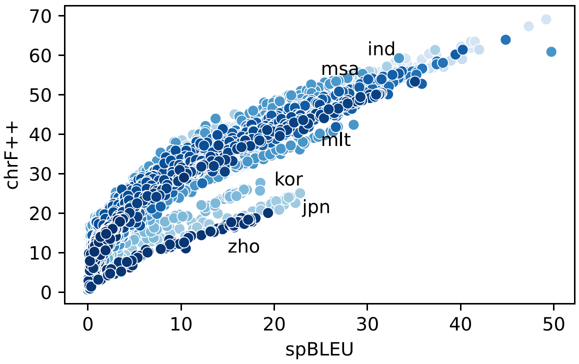

Figure 3: **Scatterplot of spBLEU against chrF++ for 101×100 language pairs in Flores-101 devtest**. Each point represents the score of the translation for a given language pair. To illustrate the behavior of evaluation into the same target language (e.g., XX-> zho), we use different color shades for each target language. Target language groupings can be observed as streaks that extend from left to right. We can observe a high-degree of global correlation (0.94) between the chrF++ and metrics, although different trends with strong *local* correlation can be observed for individual languages like Chinese (zho), Japanese (jpn), and Korean (kor).

supported by Flores-101. In Figure 3, we plot chrF++ vs. spBLEU scores resulting from translating between all languages in FLORES. We observe that there is a strong, linear relationship between the two metrics (Pearson correlation of 0.94). Notably, target languages like Chinese, Korean, and Japanese behave differently than the rest, yet the relationship between the two metrics remains locally strong.

Takeaway. Overall, we conclude that spBLEU functions fairly similarly to BLEU, especially on languages that usually default to the `moses-tokenizer`. Moreover, spBLEU exhibits a strong correlation to the tokenization-independent chrF++, across a myriad of language pairs, yet has the advantage of keeping the familiarity of BLEU. In short, spBLEU combines the familiarity of BLEU with the generalizability of chrF++. For the vast majority of languages without custom tokenizers, spBLEU provides the ability to quantify performance in the same way, with a single tokenization model. In the rest of the work, we use spBLEU to evaluate model performance.

6 Evaluating Baselines on Flores-101

In this section, we present the evaluation of three baseline models on Flores-101.

6.1 Data Splits

Flores-101 is divided into dev, devtest, and test. The dev set is meant to be used for hyper-parameter tuning. The devtest is meant to be used for testing purposes during the development phase. The test set will not be released, but is available via a publicly available evaluation server,[11] while the dev and devtest are publicly downloadable. The primary motivation for keeping the test set available only through an evaluation server is to guarantee equivalent assessment of models and reduce overfitting to the test set. Further, as the dataset is many-to-many, if the source sentences are released, the target sentences would also be released.

6.2 Baselines

We evaluate three baselines:

- **M2M-124**: Fan et al. (2020) created a Many-to-Many translation model, but did not have full coverage of Flores-101. We extended their model by supplementing OPUS data. We trained two different sizes of models with 615M and 175M parameters.
- **OPUS-100**: Zhang et al. (2020) trained multilingual machine translation models on an English-centric OPUS dataset with language-aware layers and random online backtranslation (RoBT). We evaluate the 24-layer model with backtranslation with 254M parameters.
- **Masakhane**: The Maskhane Participatory Research effort, focusing on African languages, has developed and open-sourced for the community various machine translation models (Nekoto et al., 2020; Abbott and Martinus, 2019).

6.3 Generation

We generate from all models with beam size 5, setting the max generation length to 200. Given the large number of directions covered in Flores-101, we do not tune the beam size, length penalty, or minimum/maximum generation length.

6.4 Results

We compare the performance of various models across several axes, analyzing the effect of languages, data, and domain.

[11] `https://dynabench.org/flores`.

	Afro-Asiatic	Austronesian	Balto-Slavic	Bantu	Dravidian	Germanic	Indo-Aryan	Nilotic+Other AC	Romance	Sino-Tib+Kra-Dai	Turkic	Avg
Num Languages:	7	6	14	10	4	9	14	5	10	4	5	
Afro-Asiatic	4.20	6.82	10.93	2.31	1.21	11.95	3.43	0.93	11.70	3.66	2.73	5.44
Austronesian	6.39	11.50	13.78	3.48	2.08	15.53	4.69	**1.45**	14.95	5.78	4.13	7.61
Balto-Slavic	8.32	12.29	**22.81**	3.48	3.25	21.67	6.82	0.89	21.75	7.31	5.87	10.41
Bantu	3.28	5.70	6.29	2.37	1.16	7.40	2.16	1.37	7.16	2.77	1.95	3.78
Dravidian	3.04	4.56	7.21	1.44	2.34	7.66	3.62	0.39	7.31	2.73	2.04	3.85
Germanic	**9.48**	**14.25**	22.56	**4.17**	**3.26**	**26.09**	**6.89**	1.40	23.53	**7.98**	**6.24**	**11.44**
Indo-Aryan	3.64	5.27	8.56	1.60	2.01	8.81	3.70	0.42	8.66	3.26	2.36	4.39
Nilotic+Other AC	1.60	3.10	2.76	1.45	0.43	3.48	0.79	0.95	3.48	1.29	0.81	1.83
Romance	8.25	12.74	20.70	3.22	2.41	22.43	6.04	1.15	**24.44**	6.96	5.46	10.35
Sino-Tib+Kra-Dai	4.68	7.45	10.58	2.29	2.29	10.84	4.10	0.67	11.05	5.10	3.20	5.66
Turkic	3.55	5.24	9.35	1.61	1.24	8.81	2.96	0.58	9.14	3.13	2.38	4.36
Avg	5.13	8.08	12.32	2.49	1.97	13.15	4.11	0.93	13.01	4.54	3.38	

Table 6: **Many-to-Many Performance on Family Groups**. We display the spBLEU on the devtest of Flores-101 for the M2M-124 615M parameter model. Each cell represents the average performance for translating from all the languages in the source group (row) into the each language of the target group (column). We highlight in gray the cells that correspond to within-group evaluation. In bold we show the best performance per target group and underline the best performance per source group.

	Very Low	Low	Medium	High
	< 100K	(100K, 1M)	(1M, 100M)	>100M
# Langs	15	40	38	6
Very Low	1.6	2.3	7.0	9.1
Low	2.0	2.74	8.5	10.3
Medium	3.8	5.4	19.1	23.4
High	4.3	5.8	21.7	27.3
Avg	2.9	4.1	14.1	17.6

Table 7: **Many-to-Many Performance by available Bitext through English**. We show spBLEU on devtest for M2M-124 615M parameter model. spBLEU is worse for low-resource languages compared to high resource languages, and translating into low-resource languages is harder than translating out of them.

Num Sentences	News 993	Junior 1006	Voyage 1002	Avg
English ←	20.64	20.67	19.41	20.24
English →	16.85	16.67	15.48	16.33
Chinese ←	11.57	9.66	9.55	10.26
Chinese →	10.02	9.93	9.57	9.84
Spanish ←	14.91	13.80	13.23	13.98
Spanish →	11.67	10.96	10.37	11.00
Hindi ←	14.33	14.15	13.84	14.11
Hindi →	10.88	10.86	10.11	10.62
Arabic ←	8.39	8.23	7.74	8.12
Arabic →	9.81	10.31	9.54	9.88
Many-to-Many	8.56	7.97	7.59	

Table 8: **Many-to-Many Performance by Domain.** We show spBLEU on three partitions of the Flores-101 devtest according to the originating domains. We compute the corpus spBLEU for each language in each domain, and then average across languages.

6.4.1 Findings From All Directions

English-Centric Translation. Performance of translation *into* English is strong, with only a few languages with spBLEU below 10. Performance *out of* English is worse. Performance is heavily correlated with amount of training data, which we discuss in greater detail later. We note that the construction of Flores-101 is English-centric in and of itself, as the same set of sentences are translated from English into all languages. This can affect the performance of non-English-centric directions, because languages that are similar to each other may have been translated differently if they were not translated out of English. For example, the sentence construction of Lao when translating from Thai to Lao may look different compared to English to Thai and English to Lao.

Many-to-Many Translation. Across non-English-centric directions, performance requires improvement—translation in and out of most African

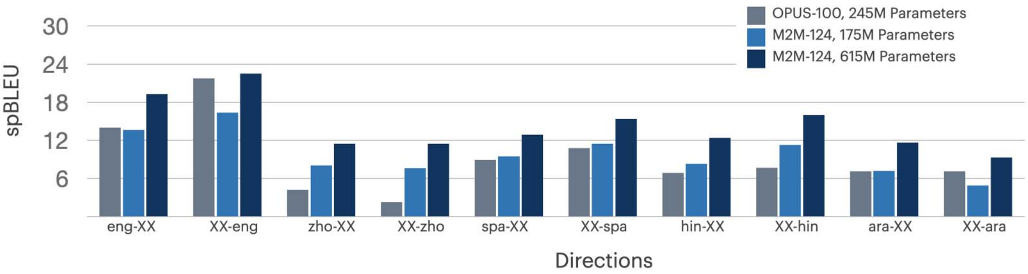

Figure 4: **Comparison between OPUS-100 and M2M-124** on several one-to-many and many-to-one translation tasks using five languages: English, Chinese, Spanish, Hindi, and Arabic. Because the open-source OPUS-100 model covers only 80 languages of Flores-101, we restrict the evaluation to only these languages.

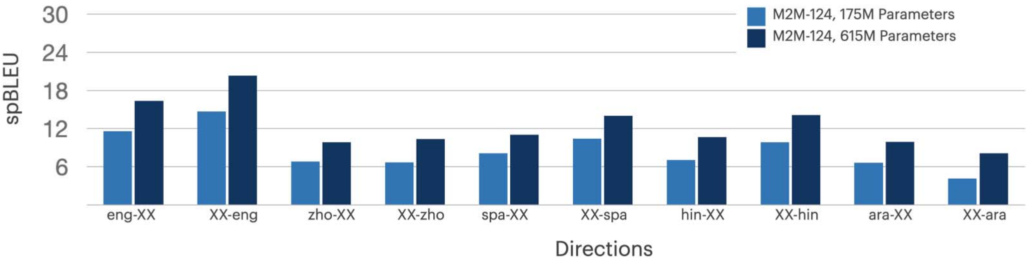

Figure 5: **Full results of M2M-124 Models** on several one-to-many and many-to-one translation tasks using five languages: English, Chinese, Spanish, Hindi, and Arabic.

languages, for example, struggles to reach 5 spBLEU. In contrast, translation into many European languages, even low-resource languages such as Occitan, has much better performance (over 10 spBLEU for many directions). This result highlights the importance of both the amount of data and transfer learning from related languages. For instance, translation to and from Occitan can naturally borrow from related high-resource languages like French, Italian, and Spanish. However, the same cannot be said about most African languages, for which related languages are also low resource and difficult to translate.

Performance by Language Family. We group languages into eleven general families and report in Table 6 the average spBLEU for translating from and into each family. Our results indicate that Bantu, Dravidian, Indo-Aryan, and Nilotic are the language families where M2M-124 struggles the most, attaining an average spBLEU below 5 points. Even translation within the language family (see values in the diagonal) is poor. In general, Germanic, Romance, and Balto-Slavic are the language families that yield the largest spBLEU scores (above 10 spBLEU points in average). Overall, many-to-many translation requires improvement.

Performance by Resource Level. Performance is often closely tied to the amount (and quality) of training data. Certain language families have much less data. For example, almost every single African language is considered a low-resource translation direction. We classify languages into four bins based on resource level of bitext with English: *high*-resource languages, with more than 100M sentences, *mid*-resource with between 1M and 100M sentences, *low*-resource with between 100K and 1M sentences, and finally *very low*-resource with less than 100K sentences. Our results are summarized in Table 7. As hypothesized, performance increases with greater quantity of training data, in a clear pattern. spBLEU increases moving from left to right, as well as from top to bottom. Even translation between high-resource and low-resource languages is still quite low, indicating that lack of training data strongly limits performance.

Performance by Domain. We analyze whether certain domains are more difficult to translate than others. We report results of translating in and out of five languages, namely, English, Chinese, Spanish, Hindi, and Arabic, as well as the average across all of the 10,000 possible directions. The results in Table 8 demonstrate that the factor

	Masakhane	M2M-124
English → Yoruba	2.04	2.17
English → Zulu	11.85	3.89
English → Swahili	22.09	26.95
English → Shona	8.19	11.74
English → Nyanja	2.19	12.9
English → Luo	5.33	3.37

Table 9: **spBLEU of Masakhane-MT**. We evaluate models on translating from English to six different African languages. We compare against the M2M-124 615M parameter model.

that affects quality the most is the language we translate in and out of rather than domain. Overall, WikiNews is the easiest with slightly higher spBLEU, and WikiVoyage is the hardest domain.

6.4.2 Comparison of Various Systems

Comparison with OPUS-100. We evaluate OPUS-100 (Zhang et al., 2020) with 254M parameters and the two versions of M2M-124 (Fan et al., 2020) with 175 and 615M parameters. We calculate spBLEU in and out of five languages: English, Chinese, Spanish, Hindi, and Arabic. Results are shown in Figure 4. Note that OPUS-100 only covers 80 languages in Flores-101, so this figure is on the subset of 80 languages covered by all models, for comparability. We see a consistent trend across models and directions: The larger M2M-124 has the best performance, followed by the smaller M2M-124 and OPUS-100.

We display results of M2M-124 175M parameters and 615M parameters on the full set of Flores-101 languages (see Figure 5). Comparing results with Figure 4, it is evident that the average performance in these language groupings has decreased, indicating that the additional languages in Flores-101 are likely very difficult. We see the same consistent trend that the larger M2M-124 model has stronger performance.

Comparison with Masakhane. The comparison with OPUS-100 compares M2M-124 with another multilingual model. However, various researchers in the low-resource translation community have developed models for specific languages, which could produce specialized models with higher quality. We evaluate models from English to the following languages: Yoruba, Zulu, Swahili, Nyanja, Shona, and Luo. Results are shown in Table 9. We observe that for two languages—Zulu and Luo—Masakhane's open sourced models have stronger performance on Flores-101 than the M2M-124 model. The remaining languages we assess have similar or worse performance than M2M-124.

7 Conclusion

The potential to develop translation for languages globally is hindered by lack of reliable, high-quality evaluation. We create and open-source Flores-101, a benchmark with 101 languages. It supports many-to-many evaluation, meaning all 10,100 directions can be evaluated. Unlike many other datasets, Flores-101 is professionally translated, including human evaluation during dataset creation. Beyond translation, Flores-101 can be used to evaluate tasks such as sentence classification, language identification, and domain adaptation.

References

Jade Abbott and Laura Martinus. 2019. Benchmarking neural machine translation for Southern African languages. In *Proceedings of the 2019 Workshop on Widening NLP*, pages 98–101, Florence, Italy. Association for Computational Linguistics.

David I. Adelani, Dana Ruiter, Jesujoba O. Alabi, Damilola Adebonojo, Adesina Ayeni, Mofe Adeyemi, Ayodele Awokoya, and Cristina España-Bonet. 2021. Menyo-20k: A multi-domain english-yor\ub\'a corpus for machine translation and domain adaptation. *arXiv preprint arXiv:2103.08647*.

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. https://doi.org/10.18653/v1/P19-1310

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 D. M. A. Ali, Andrew Caines, and Jaimito L. A. Malavi. 2021. Towards a parallel corpus of Portuguese and the Bantu language Emakhuwa of Mozambique. *arXiv preprint arXiv: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 *EMNLP Workshop on NLP-COVID*. https://doi.org/10.18653/v1/2020.nlpcovid19-2.5

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, Melvin Johnson, Maxim Krikun, Mia Xu Chen, Yuan Cao, George Foster, Colin Cherry, Wolfgang Macherey, Zhifeng Chen, and Yonghui Wu. 2019. Massively multilingual neural machine translation in the wild: Findings and challenges. *arXiv preprint arXiv:1907.05019*.

Allahsera Auguste Tapo, Michael Leventhal, Sarah Luger, Christopher M. Homan, and Marcos Zampieri. 2021. Domain-specific MT for low-resource languages: The case of Bambara-French. *arXiv e-prints*, arXiv–2104.

Mikko Aulamo, Sami Virpioja, Yves Scherrer, and Jörg Tiedemann. 2021. Boosting neural machine translation from finnish to northern Sámi with rule-based backtranslation. *NoDaLiDa 2021*, page 351.

Loïc Barrault, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo, Nikola Ljubešic, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post, and Marcos Zampieri. 2020a. Findings of the 2020 conference on machine translation (wmt20). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1–55, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-5301

Loïc Barrault, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo, Nikola Ljubešić, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post, and Marcos Zampieri. 2020b. Findings of the 2020 conference on machine translation (wmt20). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1–55. https://doi.org/10.18653/v1/W19-5301

Loïc Barrault, Ondřej Bojar, Marta R. Costa-Jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, et al. 2019. Findings of the 2019 conference on machine translation (wmt19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61. https://doi.org/10.18653/v1/W19-5301

Ondřej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, and Christof Monz. 2018. Findings of the 2018 conference on machine translation (wmt18). In *Proceedings of the Third Conference on Machine Translation*, volume 2, pages 272–307. https://doi.org/10.18653/v1/W18-6401

Ondřej Bojar, Chatterjee Rajen, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, and Marco Turchi. 2017. Findings of the 2017 conference on machine translation (wmt17). In *Second Conference on Machine Translation*, pages 169–214. The Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-4717

Houda Bouamor, Hanan Alshikhabobakr, Behrang Mohit, and Kemal Oflazer. 2014. A human judgement corpus and a metric for Arabic MT evaluation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 207–213, Doha, Qatar. Association for Computational Linguistics. https://doi.org/10.3115/v1/D14-1026

Shweta Chauhan, Shefali Saxena, and Philemon Daniel. 2021. Monolingual and parallel corpora for Kangri low resource language. *arXiv preprint arXiv:2103.11596*.

Christos Christodouloupoulos and Mark Steedman. 2015. A massively parallel corpus: The bible in 100 languages. *Language Resources and Evaluation*, 49(2):375–395. https://doi.org/10.1007/s10579-014-9287-y, PubMed: 26321896

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451. https://doi.org/10.18653/v1/2020.acl-main.747

Chenchen Ding, Masao Utiyama, and Eiichiro Sumita. 2016. Similar southeast Asian languages: Corpus-based case study on Thai-Laotian and Malay-Indonesian. In *Proceedings of the 3rd Workshop on Asian Translation (WAT2016)*, pages 149–156.

Bonaventure F. P. Dossou and Chris C. Emezue. 2020. Ffr v1. 1: FonFrench neural machine translation. *arXiv preprint arXiv:2006.09217*.

Bonaventure F. P. Dossou and Chris C. Emezue. 2021. Crowdsourced phrase-based tokenization for low-resourced neural machine translation: The case of Fon language. *arXiv preprint arXiv:2103.08052*.

Abteen Ebrahimi, Manuel Mager, Arturo Oncevay, Vishrav Chaudhary, Luis Chiruzzo, Angela Fan, John Ortega, Ricardo Ramos, Annette Rios, Ivan Vladimir, Gustavo A. Giménez-Lugo, Elisabeth Mager, Graham Neubig, Alexis Palmer, Rolando A. Coto Solano, Ngoc Thang Vu, and Katharina Kann. 2021. AmericasNLI: Evaluating zero-shot natural language understanding of pretrained multilingual models in truly low-resource languages. *arXiv preprint arXiv:2104.08726*.

Ignatius Ezeani, Paul Rayson, Ikechukwu Onyenwe, Chinedu Uchechukwu, and Mark Hepple. 2020. Igbo-English machine translation: An evaluation benchmark. *arXiv preprint arXiv: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, Edouard Grave, Michael Auli, and Armand Joulin. 2020. Beyond English-centric multilingual machine translation. *arXiv preprint arXiv:2010.11125*.

Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Tajudeen Kolawole, Taiwo Fagbohungbe, Solomon Oluwole Akinola, Shamsuddee Hassan Muhammad, Salomon Kabongo, Salomey Osei, Sackey Freshia, Rubungo Andre Niyongabo, Ricky Macharm, Perez Ogayo, Orevaoghene Ahia, Musie Meressa, Mofe Adeyemi, Masabata Mokgesi-Selinga, Lawrence Okegbemi, Laura Jane Martinus, Kolawole Tajudeen, Kevin Degila, Kelechi Ogueji, Kathleen Siminyu, Julia Kreutzer, Jason Webster, Jamiil Toure Ali, Jade Abbott, Iroro Orife, Ignatius Ezeani, Idris Abdulkabir Dangana, Herman Kamper, Hady Elsahar, Goodness Duru, Ghollah Kioko, Espoir Murhabazi, Elan van Biljon, Daniel Whitenack, Christopher Onyefuluchi, Chris Emezue, Bonaventure Dossou, Blessing Sibanda, Blessing Itoro 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. *arXiv preprint arXiv:2010.02353*.

Markus Freitag and Orhan Firat. 2020. Complete multilingual neural machine translation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 550–560, Online. Association for Computational Linguistics.

Andargachew Mekonnen Gezmu, Andreas Nürnberger, and Tesfaye Bayu Bati. 2021. Extended parallel corpus for Amharic-English machine translation. *arXiv preprint arXiv: 2104.03543*.

Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc'Aurelio Ranzato. 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. https://doi.org/10.18653/v1/D19-1632

M. Johnson, M. Schuster, Q. V. Le, M. Krikun, Y. Wu, Z. Chen, N. Thorat, F. Viagas, M. Wattenberg, G. Corrado, M. Hughes, and J. Dean. 2016. Google's multilingual neural machine translation system: Enabling zero-shot translation. In *Transactions of the Association for Computational Linguistics*. https://doi.org/10.1162/tacl_a_00065

Philipp Koehn. 2005. In *Europarl: A parallel corpus for statistical machine translation*. Citeseer.

Sai Koneru, Danni Liu, and Jan Niehues. 2021. Unsupervised machine translation on Dravidian languages. *arXiv preprint arXiv:2103.15877*.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv preprint arXiv:1808.06226*. https://doi.org/10.18653/v1/D18-2012

Garry Kuwanto, Afra Feyza Akyürek, Isidora Chara Tourni, Siyang Li, and Derry Wijaya. 2021. Low-resource machine translation for low-resource languages: Leveraging comparable data, code-switching and compute resources. *arXiv preprint arXiv:2103.13272*.

Yachao Li, Jing Jiang, Jia Yangji, and Ning Ma. 2021. Finding better subwords for Tibetan neural machine translation. *Transactions on Asian and Low-Resource Language Information Processing*, 20(2):1–11. https://doi.org/10.1145/3448216

Chaitanya Malaviya, Graham Neubig, and Patrick Littell. 2017. Learning language representations for typology prediction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2529–2535. https://doi.org/10.18653/v1/D17-1268

Benjamin Marie, Atsushi Fujita, and Raphael Rubino. 2021. Scientific credibility of machine translation research: A meta-evaluation of 769 papers. *CoRR*, abs/2106.15195. https://doi.org/10.18653/v1/2021.acl-long.566

Nitika Mathur, Johnny Wei, Markus Freitag, Qingsong Ma, and Ondřej Bojar. 2020. Results of the wmt20 metrics shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 688–725, Online. Association for Computational Linguistics.

El Moatez Billah Nagoudi, Wei-Rui Chen, Muhammad Abdul-Mageed, and Hasan Cavusogl. 2021. Indt5: A text-to-text transformer for 10 indigenous languages. *arXiv preprint arXiv:2104.07483*.

Evander Nyoni and Bruce A Bassett. 2021. Low-resource neural machine translation for southern African languages. *arXiv preprint arXiv:2104.00366*.

Maja Popović. 2017. chrf++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-4770

Matt Post. 2018. A call for clarity in reporting BLEU scores. *arXiv preprint arXiv:1804.08771*. https://doi.org/10.18653/v1/W18-6319

Hammam Riza, Michael Purwoadi, Teduh Uliniansyah, Aw Ai Ti, Sharifah Mahani Aljunied, Luong Chi Mai, Vu Tat Thang, Nguyen Phuong Thai, Vichet Chea, Sethserey Sam, Sopheap Seng, K. Soe, K. Nwet, M. Utiyama, and Chenchen Ding. 2016. Introduction of the Asian language treebank. In *2016 Conference of The Oriental Chapter of International Committee for Coordination and Standardization of Speech Databases and Assessment Techniques (O-COCOSDA)*, pages 1–6. IEEE. https://doi.org/10.1109/ICSDA.2016.7918974

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2021. 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. https://doi.org/10.18653/v1/2021.eacl-main.115

Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, Armand Joulin, and Angela Fan. 2019. CCMatrix: Mining billions of high-quality parallel sentences on the web. *arXiv preprint arXiv:1911.04944*.

Stephanie Strassel and Jennifer Tracey. 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'16)*, pages 3273–3280, Portorož, Slovenia. European Language Resources Association (ELRA).

Ye Kyaw Thu, Win Pa Pa, Masao Utiyama, Andrew Finch, and Eiichiro Sumita. 2016. Introducing the Asian language treebank (ALT). In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1574–1578, Portorož, Slovenia. European Language Resources Association (ELRA).

Jörg Tiedemann. 2018. Emerging language spaces learned from massively multilingual corpora. In *Digital Humanities in the Nordic Countries DHN2018*, pages 188–197. CEUR Workshop Proceedings.

Jörg Tiedemann. 2020. The Tatoeba Translation Challenge – Realistic data sets for low resource and multilingual MT. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1174–1182, Online. Association for Computational Linguistics.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2019. Ccnet: Extracting high quality monolingual datasets from web crawl data. *arXiv preprint arXiv:1911.00359*.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. 2020. Improving massively multilingual neural machine translation and zero-shot translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1628–1639, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.148

Mike Zhang and Antonio Toral. 2019. The effect of translationese in machine translation test sets. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 73–81, Florence, Italy. Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-5208