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

Naman Goyal[*], **Cynthia Gao**[*], **Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato**[†], **Francisco Guzmán**[†], **Angela Fan**[†‡]
Facebook AI Research, [‡]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 multilingually 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*, i.e. languages without a lot of training data. In fact, many low-resource languages are not even supported by most popular translation engines. Yet, the majority 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 shifting focus towards low-resource languages.

Over the past decade, the research community has made a lot of recent progress on models for low-resource machine translation. Approaches like iterative backtranslation (Sennrich et al., 2015), multilingual machine translation (Johnson et al., 2016; Tang et al., 2020; Fan et al., 2020), and even unsupervised machine translation (Lample et al., 2018; Artetxe et al., 2018) have shown promising results. Beyond modeling, a major challenge for research in low-resource machine translation is evaluation. Low-resource evaluation is critical to the scientific progress of the field, because evaluation enables proper comparison of approaches and ultimately, a better understanding of what needs further investigation and improvement. Unfortunately, finding high-quality data suitable for the evaluation process is even more difficult in low-resource scenarios.

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., 2020b; ∀ 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 mul-

[*]Indicates equal contribution
[†]Indicates equal contribution

tilingual 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. As an additional challenge, there are no established practices for how to build such benchmark. Working with professional translators in low-resource languages is difficult because of their scarce availability, and because it is non-trivial to check the quality of their work (Guzmán et al., 2019).

We present the FLORES-101 benchmark, consisting of 3001 sentences sampled from 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 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 will support other kinds of evaluations and tasks, such as document level translation, multimodal translation and text classification. Fifth, we propose a BLEU metric based on sentence piece tokenization (Kudo and Richardson, 2018) (§5) that enables evaluation of all languages in the set in a unified and extensible framework. Finally, we publicly release both data and baselines used in our experiments (§6), to foster research in low-resource machine translation and related areas.

This paper is organized as follows: In Section 2, we describe related work to construct evaluation benchmarks in machine translation. In Section 3, we detail the construction process of FLORES-101, from sourcing sentences to translate to defining the translation workflow. Section 4 gives a detailed overview of the sentences, languages, and quality of FLORES-101. In Section 5, we describe our proposed SentencePiece BLEU metric which unifies and simplifies evaluation. Section 6 uses FLORES-101 to evaluate various public translation models, and breaks down model performance by amount of training data, domain, sentence length, and language family. We present our conclusions in Section 7.

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 (Ondrej et al., 2017), Kazakh in 2018 (rej Bojar et al., 2018), Gujarati and Lithuanian in 2019 (Barrault et al., 2019), and Inuktitut, Khmer, Pashto, and Tamil in 2020 (Barrault et al., 2020a). 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).

To date, the largest resource of parallel data which can also be used for evaluation purposes is OPUS (Tiedemann, 2012), which is itself a collection of publicly available parallel datasets. While OPUS has by far the largest coverage of languages, particularly to and from English, it consists of a mixture of manually translated and mined data, which results in a large variety of datasets and domains with varying level of quality. For instance, OPUS contains parallel data translated by humans coming from operating system handbooks like Ubuntu, or parallel data from religious documents (Liu et al., 2021) like Jehovah's Witness magazines (Agić and Vulić, 2019) and the Bible. These have recently been expanded to include more languages (Nicolai et al., 2021) as well. OPUS also

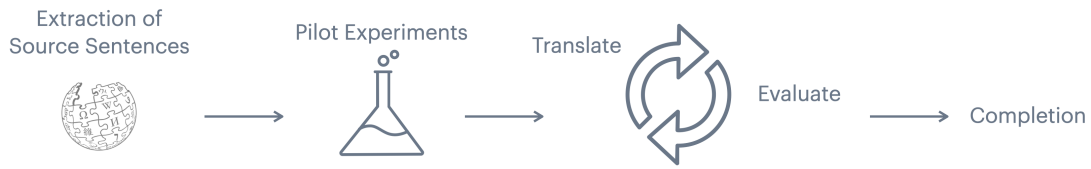

Figure 1: **Dataset Construction**. The workflow used to construct FLORES-101 has three phases: **(1)** sourcing sentences to translate from English Wikipedia, **(2)** designing pilot studies to define efficient and effective translation and evaluation processes, **(3)** launching the actual translation across all languages. The last stage is iterative, as translations may go through additional rounds of re-translation if the evaluation indicates that quality is insufficient; see Fig. 2 for further details.

contains a variety of other automatically-aligned datasets, such as various versions of TED talks, which are usually of lower quality (Ye et al., 2018; Zhang et al., 2020; Fan et al., 2020). Similarly, OPUS contains large parallel datasets generated via automatic filtering and alignment methods, such as WikiMatrix (Schwenk et al., 2021), ccMatrix (Schwenk et al., 2019), ccAligned (El-Kishky et al., 2020), and ParaCrawl (Esplà et al., 2019), which contain noisy translations. While these may be utilized for training, they are clearly unsuitable for evaluation purposes due to automatic alignment.

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.

Lastly, the current literature in low-resource translation provides very scarce guidance in terms of best practices and methodology to construct parallel datasets and perform quality assurance. The much lower number of translators is problematic because it makes the annotation process much more susceptible to variance in the proficiency level of such annotators. In Flores v1.0 (Guzmán et al., 2019), a mixture of human and automatic checks were used to filter and rework problematic translations. In TICO-19 (Anastasopoulos et al., 2020), a two-step translation and quality assurance process was followed. Despite its technical complexity, the annotation process for benchmark sets is in fact seldom documented in technical reports. This is still largely an uncharted territory. However, there are a lot of practical questions related to setting up and ensuring the quality of large-scale translation campaigns targeting low-resource languages which may have very few annotators. For example: What guidelines should be considered for translators and evaluators? What workflow is most efficient and effective? What automatic checks should be put in place to minimize human intervention? When can a dataset be declared to have reached a sufficient level of quality to be released? In this study, we document our choices and processes in a hope to build and consolidate best practices of dataset construction for the machine translation community.

3 Dataset Construction

The construction of FLORES-101 is intended to accomplish several goals: (*i*) to enable the evaluation of many-to-many multilingual models, meaning the evaluation of translations from any language to any other language including very long-tail languages; (*ii*) to enable other kinds of evaluation beyond machine translation, such as document-level translation, multi-modal translation, multilingual classification, and so on; (*iii*) most importantly, to build a high-quality evaluation benchmark.

To achieve the above goals, the overall construction process consisted of three phases, as outlined in Fig 1. First, we extracted sentences to translate from English Wikipedia. Second, we designed and ran pilot experiments to determine the translation process, and finally we launched the actual translation workflow for over 100 languages. In this section, we describe the process in detail. The reader who is more curious about general statistics can safely skip this section and go to Section 4.

3.1 Sourcing Sentences

We describe how the domains and sentences in FLORES-101 were selected. A high-level summary of the dataset can be found in Table 1.

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 is currently supported in 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 professional translators that could read both the source and target side.

Instead, we opted to source all sentences from English Wikipedia, 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 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 only 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. 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.

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

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

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

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

Sentence Selection. The sentence selection process consisted of selecting an article at random from each source, and then manually selecting a few (typically between 3 and 5) contiguous sentences from each article, avoiding segments with very short or malformed sentences. To avoid bias coming from the document structure, we carefully selected one paragraph per document, from either the beginning, middle or end of the article. We balanced the location selection to be equally distributed across the whole corpus — roughly one third of paragraphs were sampled from the beginning of the article, one third from the middle, and so on. For each sentence, we also extracted the Wikipedia URL, topic, and noted boolean flags to indicate whether the sentence contained entities linked to other Wikipedia pages and images. The selection process was performed by 10 different annotators in our lab, 6 male and 4 female; with different roles in research: researchers (scientists and engineers) and program/project managers; originally coming from different regions of the world: East Asia, South Asia, Southern Europe, Latin America and North America. We manually labeled all sentences by a more detailed *sub-topic*, one of 10 possibilities: crime, disasters, entertainment, geography, health, nature, politics, science, sports, and travel. Table 1 reports basic statistics of the originating English sentences.

Since several contiguous sentences are extracted from the same article and since we also provide the corresponding URL, we support evaluation of machine translation at the document level. With the additional meta-data, we also enable evaluation of multimodal machine translation.

3.2 Pilot Experiments

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 render the quality of the dataset insufficient or inconsistent for that language. Here, we describe the process we followed to define an efficient and high quality translation workflow. To this end, we report two pilot experiments we used to determine how we should proceed with the creation of this large-scale evaluation dataset.

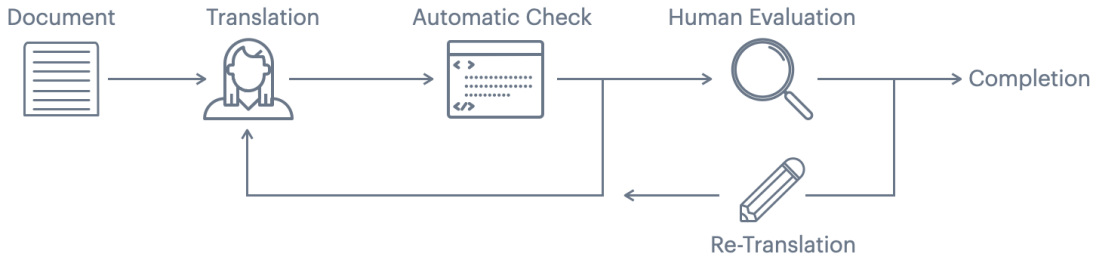

Figure 2: **Depiction of Overall Translation Workflow.** For each target language, sentences are translated by a translation Language Service Provider (LSP). The resulting translations are automatically checked. If these checks fail, translations are sent back for re-translation. If the automatic checks pass, translations are sent to another LSP for human evaluation. If the quality is not sufficient, translations are sent back to the original translation LSP for re-translation. Depending on the human score, the process can repeat for multiple rounds of human evaluation.

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 1: **Statistics of FLORES-101.** FLORES-101 contains 3001 sentences selected from 842 articles, divided into three splits: dev, devtest, and test. The articles are sourced from three domains, breaking down into 10 sub-topic classifications.

3.2.1 Providers and Workflows

To ensure the best possible level of quality for our translations, we designed two pilots aimed to determine the best workflow to follow for translating hundreds of languages. The first pilot experiment was meant to select the best translation providers for each language and the second, to determine the best translation-quality assurance workflow.

Language Service Providers. As a starting point, let us assume that each language can be translated by K different Language Service Providers (LSPs) and that they all charge the same price for translating a sentence. We randomly selected 100 sentences and 8 language pairs, and assigned each language to at least two LSPs. We then used another LSP to evaluate all translations.

Based on human assessment of translation quality, we selected the two LSPs that produced the highest quality translations. We chose two translation LSPs to make our translation process not rely entirely on a single-party, while reducing the communication overhead created by working with too many external parties.

Translation and Quality Assurance Workflow. Despite having reliable translation LSPs, we need to ensure that each translation conforms to the highest level of quality required by a benchmark. Therefore, we split our workflow into two parts: translation (which includes editing), performed by an initial LSP, and quality assurance (QA) performed by an independent LSP. After the QA process, a translation might need *re-translation* or minimal editing to improve its quality. Here, we explored the best workflow for when *re-translation* is needed. Assuming there are two translation LSPs, A and B, and a separate QA LSP C, we can have two possible workflows: (*i*) **A–C–B** we can have B re-translate translations flagged by C that were produced by A; (*ii*) **A–C–A** an alternative and simpler workflow is to have the same LSP take care of both translation and re-translation for a given language, and to have each translation LSP handle half of the languages.

The advantage of workflow (*i*) is the re-translation process is the least biased, particularly on low-resource languages where the re-translator and the translator could be the same person. On the other hand, the re-translator has less context

and the workflow has higher complexity because data comes in and out of LSPs at different times, making the whole process more error prone.

We tested both workflows and observed negligible differences between the two workflows, and therefore, we chose workflow (*ii*) , i.e. A–C–A, with the same LSP taking care of both translation and re-translation as it is operationally simpler.

3.2.2 Automatic Translation Quality

The second pilot experiment aimed to investigate how to assess translation quality automatically.

Implemented Checks. We implemented several checks to ensure the first round of translations were of acceptable quality: (*i*) language identification, (*ii*) checking whether the translation is a copy of the source sentence; (*iii*) checking whether the translation has significantly different length, (*iv*) checking translation fluency according to a language model, (*v*) and checking whether the translation is a copy of the translation produced by publicly available translation engines. Among all checks, we found that (*v*) was the most significant issue, despite formulating clear guidelines forbidding the use of translation engines. This is important, as we want to our translations to be as unbiased as possible. Relying on verbatim or post-edited translations from online engines would be misguiding, and give them unfair advantage when using a reference-based automatic metric for comparison.

To address the issue, we proposed a heuristic to detect and reject professional translations that are likely copies from translation engines based on their sentence-level similarity. Moreover, when the translations of two different engines are available, we check whether a sentence is very similar to the output of a specific translation engine while being different from the output produced by other translation engines.

The heuristic is as follows: let x be the translation produced by a translation LSP, y_A and y_B be translations produced by translation engine A and B; and let spBLEU(x, y) be the sentence-level SentencePiece BLEU (cfr. Sec. 5) between sentence x and the reference y. Then, we declare a that a translation was a copy if $\text{spBLEU}(x, y_A) - \text{spBLEU}(x, y_B) > 20$ and $\text{spBLEU}(x, y_A) > 50$, when the language in question is supported by both engines A and B; and if $\text{spBLEU}(x, y_A) > 50$ when the translation is only supported by translation engine A.

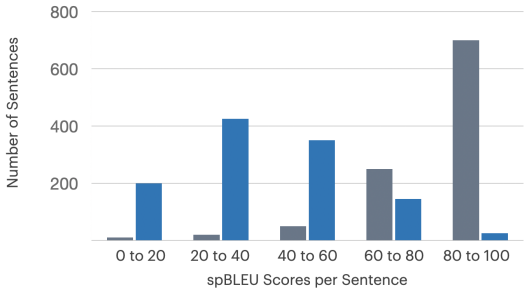

Figure 3: **Importance of Automatic Checks**. In *gray*, we show the sentence-level spBLEU of a language that displays indication of copy from a commercial translation engine. A large number of sentences have very high BLEU scores, mainly in the 80 to 100 BLEU bucket. In contrast, the bars in *blue* indicate a language that does not experience this issue. We discuss the spBLEU metric in greater detail in Section 5.

# 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 2: **Statistics of FLORES-101 Translation Workflow.** To ensure high quality, our translation workflow includes translation and re-translation steps. We break down the amount of re-translation required, and summarize that to complete one language, it takes on average two months.

The values of 50 and 20 were based on the analysis of the distribution of scores for translations of tens of languages, where we used clustering techniques to determine the right cutoff values.

Moreover, we established that any set of translations with more than 10% of the sentences violating the above criteria condition would need to be re-translated prior to perform any subsequent human evaluation. We show in Figure 3 an example of a language that passed this test and one which did not. Thanks to these automatic checks, we reduced the amount of copying from popular translation engines, streamlined the translation workflow before human evaluators assessed quality, and fully automated the process of translation, evaluation, and re-translation. This is described in the next section.

3.3 Translation and Evaluation

We describe the final workflow for collecting data for all languages in FLORES-101. We start with how we decide when a language is ready to be included in FLORES-101: the final translation quality score. Then, we detail the full translation process, including automatic and human quality checks.

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? Before we summarize the workflow to produce translations, we briefly discuss how we measure translation quality. We assess translation quality through a *Translation Quality Score*, calculated per language on a 0 to 100 scale. The translation quality score is 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, between minor, major, and critical. Based on tallying these different error types, the overall final score is determined. We encouraged evaluators to pay particularly high attention to unnatural translation errors. Based on our pilot experiments, we set the acceptable translation quality score to 90%.

Translation Workflow. The overall translation workflow is depicted in Figure 2. For each language, all source sentences are sent to a certain translation LSPs. 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. An additional verification step is added to this specific workflow with comparison of the translated data to translations from commercial engines as previously mentioned. 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 the accepted threshold, translations together with the assessment report are sent back to the translation LSP for re-translation. If the initial score is below another certain threshold (that is associated to good translation quality), the re-translated translations are evaluated by humans one more time. We summarize in Table 2 the overall statistics around the translation process. We include the guidelines used for quality evaluation in the Appendix.

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 3. 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.

4.2 Sentences in FLORES-101

Table 1 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.

4.3 Languages in FLORES-101

We summarize all 101 languages in FLORES-101 and their scripts and language families in Table 4. We note that language classification is a complex task with different classification hierarchies. We chose language families at a reasonable level of detail, i.e. fine enough such that languages can be grouped with a few other languages but not so fine that each language is in its own group. Overall, our selected languages cover a large percentage of people all over the world, with a large diversity of scripts and families. Most of these languages are spoken by millions of people, despite being considered low-resource in the research community.

	# Languages	Diverse Topics	Many to Many	Human Translations	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 3: **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 created by humans, and if the tasks of document-level translation or multimodal translation are supported.

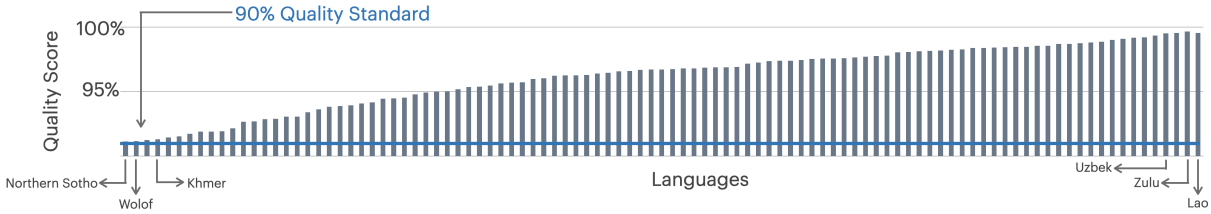

Figure 4: **Translation Quality Score across Languages.** We require the final translation quality score to be above 90% before the translation is of sufficient quality to include in FLORES-101. We depict the score distribution for all languages in FLORES-101.

In Table 4, we depict all FLORES-101 languages by their resource level. The amount of data available for a language is difficult to accurately denote, in part because quality is very important and thus, the amount of data does not necessary reflect its usefulness. Further, some data may be proprietary, and new datasets for new languages are actively being created by the research community. Thus, we report the amount of data to/from English and the amount of monolingual data available in OPUS, a public repository for multilingual data. OPUS is a heavily used resource, and itself is a collection of a large number of research datasets produced by the community over decades. The majority of languages have both bilingual data through English and monolingual data, though a number of languages have less than 100K sentences through English. Many of those also have no monolingual data available, making these truly low-resource. Examples include Shona and Nyanja.

4.4 Translation Quality

The translation quality score across all languages is depicted in Figure 4. All 101 languages in FLORES-101 meet our initial threshold of 90% quality based on human evaluation. Note that several languages were considered beyond our set of 101, but were unable to meet the bar after rounds of re-translation. 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.

We breakdown the main translation errors observed based on the quality assessments and re-translations. The largest error category across all languages was *mistranslation*, a broad error category that generally notes that the source text was not translated faithfully and the translation has rendered an incorrect meaning in the target language. Examples of mistranslation include (but not limited to) incorrect interpretation of the source text, literal translations and mistranslations of phrasal verbs and lack of disambiguation of ambiguous terms. For example, writing *...recommends hand washing **with** hand sanitizer rubs* instead of *...recommends hand washing **over** hand sanitizer rubs* would represent a mistranslation. Error categories with few errors include register, grammar, and punctuation.

We also examined if certain domains are more difficult to translate than others. Within a language, we did identify variation in the percentage of errors contributed by domain (often one domain could

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	—
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

ISO 639-3	Language	Family	Subgrouping	Script	Bitext w/ En	Mono Data
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 4: **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 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.

contribute up to 10% more errors than the others), but across languages, there was no clear trend. Overall, it appears that all domains are challenging to translate for human translators.

5 Metric: SentencePiece BLEU

How do we evaluate the performance of translation models at the scale of 101 languages? In this section, we propose the *SentencePiece BLEU* metric and analyze its performance across languages compared to various alternatives.

5.1 Motivation

Automatic evaluation of translation quality is an active field. Each year, the WMT Metrics shared task seeks to determine the automatic metric that better correlates with human evaluations (Mathur et al., 2020). While many metrics have been proposed through the years, the analysis has only included a handful of low-resource languages. Further, despite the progress in automatic metrics, the common practice is to use BLEU (Papineni et al., 2002) when reporting results. Unfortunately, using BLEU directly is suboptimal, as it relies on n-gram overlap which is heavily dependent on the particular tokenization used, i.e. tokenizing more aggressively can artificially raise the score and make it difficult to compare across reported results.

The challenge of making BLEU comparable by using equivalent tokenization schemes has been challenging for the translation community and has been partially addressed by `sacrebleu` (Post, 2018). Previous standards usually leverage the

mosestokenizer[5], which is the default tool in sacrebleu. However, for many languages, these existing tools and tokenizers are not sufficient.

For example, mosestokenizer supports a limited number of languages (often activated with the -tok flag). While its default tokenization rules might operate reasonably 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, which has incentivized the creation of BLEU variants (Bouamor et al., 2014). To further complicate matters, some languages like Hindi and Japanese already have custom tokenizers that are used when computing BLEU, although these appear scattered in various publication footnotes, while for others, no such special tokenizers have been developed yet. Further, 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 that can be applied 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 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.

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 this, we train our SPM model with temperature upsampling similar to (Conneau et al., 2020), so that low-resource languages are represented. In the future if a new language is added to FLORES-101 and this tokenizer does not support its script, we can easily add new tokens to encode it as desired.

Lang	Correlation spBLEU v. BLEU	Correlation char-BLEU v. BLEU
French	0.99	0.99
Italian	0.99	0.99
Spanish	0.99	0.99
Hindi	0.99	0.99
Tamil	0.41	0.31
Chinese	0.99	0.75

Table 5: **Spearman Correlation of spBLEU, BLEU, and char-BLEU**. We evaluated 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.

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. We dub this metric as *sentence-piece BLEU*, and we denote it as spBLEU. It is integrated into sacrebleu for ease of use[6] as the spm tokenizer.

5.3 Experiments and Analysis

We want to validate the spBLEU metric, to **(1)** see that it trends with the standard BLEU metric on languages where mosestokenizer is often used as the default, **(2)** for languages where custom tokenizers are currently used, see that spBLEU correlates with custom-tokenizer-BLEU more strongly than alternatives such as character-level BLEU, and finally **(3)** verify that spBLEU can be used for model selection purposes.

spBLEU correlates with BLEU. First, we examine if spBLEU has strong positive correlation with BLEU across various languages where the mosestokenizer is widely used by default. We

[5] https://github.com/moses-smt/mosesdecoder/blob/master/scripts/tokenizer/tokenizer.perl

[6] https://github.com/ngoyal2707/sacrebleu/tree/adding_spm_tokenized_bleu

	spBLEU v. Human Eval		spBLEU v. BLEU		Human Eval v. BLEU	
Language	Kendall τ	Same Best Model	Kendall τ	Same Best Model	Kendall τ	Same Best Model
Pashto	1.0	✓	1.0	✓	1.0	✓
Russian	0.71	✓	0.52	✓	0.80	✓
Chinese	0.90	✗	0.71	✓	0.80	✗

Table 6: **spBLEU Compared to Human Evaluation and BLEU ranking**. We analyze translation into Pashto, Russian, and Chinese. We indicate the Kendall τ between spBLEU and Human Eval, spBLEU and BLEU, and Human Eval and BLEU, as well as if the different metrics result in the selection of the same best model.

examine Spanish, Italian, and French. As shown in Table 5 (top), we find that spBLEU correlates very well with BLEU on these languages.

spBLEU is better than char-BLEU when custom tokenization is needed. Next, we examine the performance of spBLEU on languages where custom tokenizers are often used, or special rules are written for tokenization. We look at three languages: Chinese, Hindi, and Tamil. Chinese is supported by `mosetokenizer` with special rules. Hindi and Tamil have a popularly used tokenizer in the community from IndicNLP[7]. While these language-specific tokenizers are excellent, the challenge of scale and comparability exists: there are often different competing tokenizers and tokenizers need to be developed for each language we want to evaluate. A possible alternative is instead to eliminate tokenization, and evaluate characters directly and compute character-level BLEU (char-BLEU).

We examine if spBLEU is a better alternative to char-BLEU for languages that currently use special tokenizers. As shown in Table 5 (bottom), spBLEU correlates more strongly with custom tokenizer BLEU compared to the correlation between char-BLEU and custom tokenizer BLEU. While the development of custom tokenizers for specific languages produces much more accurate tokenization, spBLEU is a good alternative for comparability and scalability across a large number of languages.

spBLEU has similar performance as BLEU for model selection. Next, we turn to verifying that spBLEU can be used compare the quality of models for model selection purposes. This is important, as oftentimes automatic metrics are used in an outer loop of the training process, to select various hyper-parameters, such as model size, dropout rate, learning rate and so on. Thus, we replicate the selection of various models using spBLEU instead of BLEU, experimenting on three language directions: English to Pashto, English to Russian, and English to Chinese. We choose these languages because they were part of WMT2020 human evaluations, and thus we know the *ground-truth* ranking. We evaluate five models for Pashto, eight for Russian, and seven for Chinese. We focus on evaluation of models in directions out of English, as `mosestokenizer` works well on English.

For each of the three language directions, we compute the spBLEU and BLEU with language-specific tokenizers between different systems and the reference translation. Overall results are shown in Table 6. We first compare if the ranking of systems produced by spBLEU matches that of systems ranked by human evaluation and BLEU. We calculate exact match ranking accuracy using the Kendall τ coefficient, which ranges between -1 and 1. The ranking of systems by spBLEU matches human evaluation and BLEU perfectly for Pashto, and has strong correlation with both human evaluation and BLEU for Chinese and Russian.

However, the exact ranking may not be the most important part of a metric. Oftentimes, we want to use automatic metrics to understand which model improvement is the most effective — for example, which model to submit to WMT? Thus, it is important to check whether the best scoring model according to spBLEU matches the best scoring model according to BLEU. We find that spBLEU and BLEU indeed select the *same* best model on all three languages. Note that BLEU has no guarantee of selecting the same model as human evaluators[8].

Takeaway. Overall, we conclude that spBLEU functions fairly similarly to BLEU, especially on languages that usually default to the `mosestokenizer`. On languages that use custom tokenization, spBLEU correlates more strongly

[7] https://anoopkunchukuttan.github.io/indic_nlp_library/

[8] There are a number of ways that BLEU score can be improved that may not have an effect on human evaluation. Punctuation normalization is an example.

	Very Low	Low	Medium	High	Avg
	< 100K	(100K, 1M)	(1M, 100M)	> 100M	
Num Languages	15	40	38	6	
Very Low	1.60	2.29	6.98	9.14	5.00
Low	2.02	2.74	8.48	10.30	5.89
Medium	3.79	5.38	19.13	23.43	12.93
High	4.29	5.83	21.70	27.32	14.79
Avg	2.93	4.06	14.07	17.55	

Table 7: **Many-to-Many Performance by available Bitext data through English**. We show spBLEU on devtest of FLORES-101 for M2M-124 615M parameter model. We group languages into 4 bins. spBLEU is worse for low-resource languages compared to high resource languages, and translating into low-resource languages is harder than translating out of a low-resource language.

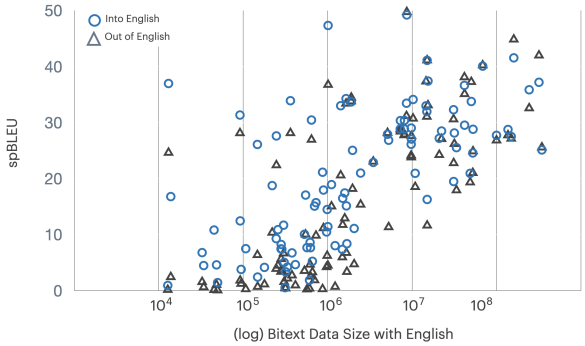

Figure 5: **spBLEU for directions in and out of English.** We compare performance against amount of available bitext data. For the same amount of data, translation into English is often stronger.

with BLEU than other alternatives, such as char-BLEU. Further, spBLEU often produces a very similar ranking of models and selects the same best model as BLEU and human evaluation.

For the vast majority of languages without custom tokenizers, spBLEU provides the ability to quantify performance in the same way, with one model. We believe that having a single model to perform tokenization will help the research community to make progress on low-resource research, while opening the door for improved versions of spBLEU that treat low-resource languages more fairly. In the subsequent rest of the work, we use spBLEU to evaluate model performance.

6 Evaluating Baselines on FLORES-101

In this section, we present evaluation of various models on FLORES-101. We describe the dev, devtest, and test splits and how we intend them to be used. We then analyze the performance of a many-to-many model based on Fan et al. (2020) and break down performance by resource level, sentence length, and language family. Finally, we compare various model variants.

6.1 Data Splits

FLORES-101 is divided into three splits: dev, devtest, and test. Unless otherwise stated, we report results on the devtest portion of FLORES-101. The dev set is meant to be used for hyper-parameter tuning. The devtest is meant to be used for testing purpose during the development phase. The test set will not be released, but will be available via a publicly available evaluation server, while the dev and devtest are publicly downloadable[9]. Through the evaluation server, the test set can be used by various evaluation campaigns, such as the WMT 2021 Large-Scale Multilingual Task[10]. 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 on different models to provide baselines for researchers who may be interested in the performance of certain directions, and to understand which languages and directions need substantial research improvements.

- **M2M-124**: Fan et al. (2020) trained the M2M-100 multilingual model by extending large-scale data mining to create training data for

[9] https://dl.fbaipublicfiles.com/flores101/dataset/flores101_dataset.tar.gz

[10] http://statmt.org/wmt21/large-scale-multilingual-translation-task.html

	Short <= 15 words	Medium (15, 25) words	Long > 25 words	Avg
Num Sentences	200	550	250	
→ English	19.11	20.42	20.47	20.00
English →	16.31	16.60	16.06	16.32
→ Chinese	10.44	10.52	10.01	10.32
Chinese →	9.81	10.21	9.34	9.79
→ Spanish	13.17	14.08	14.19	13.81
Spanish →	10.69	11.06	10.97	10.91
→ Hindi	13.58	13.82	14.73	14.04
Hindi →	10.42	10.75	10.54	10.57
→ Arabic	6.83	7.97	8.84	7.88
Arabic →	9.31	10.02	9.93	9.75
Many-to-Many	7.71	8.16	8.02	

Table 8: **Many-to-Many Performance by Sentence Length**. We show spBLEU on devtest of FLORES-101 for M2M-124 615M parameter model. We analyze if sentence length has an effect on performance across directions in English, Chinese, Spanish, Hindi, Arabic, and when averaging across all language directions. We find that length does not have a strong effect on performance.

language pairs not going through English. The original M2M-100 model does not have full coverage of the languages in FLORES-101. We extended their mined data with data from OPUS for the FLORES-101 languages not present in mined data, extending to 124 total languages. Note that for the additional languages added, OPUS does not contain a large quantity of data, and the OPUS data is rather noisy; see Table 4 for further details. On this parallel data, we trained two different sizes of models, namely a model with 615M and one with 175M parameters. Unless otherwise stated, we will report results using the 615M parameter model; this is our default throughout the rest of this paper.

- **OPUS-100**: Zhang et al. (2020) trained multilingual machine translation models on an English-centric OPUS training dataset with language-aware layers and random online backtranslation (RoBT). We evaluate the 24-layer model with backtranslation (dubbed `Ours + 24 layer + RoBT` in their work[11]) with 254M parameters.
- **Models open-sourced by Masakhane**: The Maskhane Participatory Research effort, focusing on Natural Language Processing for African languages, has developed and open-sourced for the community various machine translation models (∀ et al., 2020; Abbott and Martinus, 2019a,b). We evaluate models from English to six languages[12]: Yoruba, Zulu, Swahili, Shona, Nyanja, and Luo.

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

In this section, we report the results of the evaluation of the baseline approaches described above on the FLORES-101 devtest using spBLEU.

6.4.1 Findings From Evaluation on All Directions

All Directions. We evaluated our M2M-124 model with 615M parameters on all language pairs and report spBLEU scores in Figure 8. In Figure 8, the languages are organized alphabetically by language code, while in Figure 9, the rows and columns have been organized via spectral clustering. The spBLEU metric scores are used as affinity scores between each pair of languages. This produces clusters ordered by easiness to translate.

English-Centric Translation. Across the board, performance of translation *into* English is strong,

[11] https://github.com/bzhangGo/zero/tree/master/docs/multilingual_laln_lalt

[12] https://github.com/masakhane-io/masakhane-mt

	WikiNews	WikiJunior	WikiVoyage	Avg
Num Sentences	993	1006	1002	
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 9: **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 in that direction. We compute the performance into and out of English, Chinese, Spanish, Hindi, and Arabic, as well as average across all many-to-many directions. Overall, the News domain has slightly improved performance, but the domain effect is not strong.

with only a few languages with spBLEU below 10. Performance *out of* English is much worse. We display this graphically in Figure 5, where we show that performance into English (circle markers) is has higher spBLEU than performance out of English (triangle markers). Further, performance is overall heavily correlated with amount of training data, which we discuss in greater detail later.

Many-to-Many Translation. Across non-English-Centric directions, performance requires improvement — translation in and out of most African languages, for example, struggles to reach 5 spBLEU. In contrast, translation into many European languages, even low-resource languages such as Occitan, have 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 Resource Level. A challenge of analyzing performance of various language families is that performance is often closely tied to the amount (and quality) of available training data. Certain language families have much less data. For example, almost every single African language is considered a low-resource translation direction. Thus, we next evaluate performance based on resource level. We classify languages into four bins based on resource level of bitext data with English: *high*-resource languages, with more than 100M sentences of training data, *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. Translation between mid and high resource languages produces spBLEU scores around 20, whereas translating between very low and low-resource languages yields a mere spBLEU score of less than 5. Even translation between high resource and low-resource languages is still quite low, indicating that lack of training data strongly limits performance of current MT systems.

Performance by Sentence Length. In the previous paragraphs we have found that translation quality is affected by the amount of training data and the properties of the language. Next, we examine if translation quality is also affected a property of the sentences themselves. In particular, we calculate if the sentence length affects model performance, based on the hypothesis that longer sentences may be more complex and difficult to translate (Sutskever et al., 2014). The results in Table 8 show spBLEU on different subsets of the devtest, grouped by sentence length. The sentence

	Afro-Asiatic	Austronesian	Balto-Slavic	Bantu	Dravidian	Germanic	Indo-Aryan	Nilotic+Other AC	Romance	Sino-Tibetan+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-Tibetan+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 10: **Many-to-Many Performance on Family Groups**. We display the spBLEU on the devtest of FLORES-101 for the M2M-124 615M parameter model. We group languages into 11 families. Each cells represent 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.

length is determined by the number of tokens in the original English sentence. The bucket with short sentences collect all sentences with up to 15 tokens (in English), the medium bucket has sentences with a number of tokens in the range 16 to 25, and the last bucket has sentences with more than 25 tokens. The table shows that spBLEU is rather constant with respect to the sentence length, and in fact it slightly increases with the length of the sentence, contrary to our initial conjecture.

Performance by Domain. We analyze if model performance is affected by domain, to check whether certain domains are more difficult to translate than others. FLORES-101 contains three domains: WikiNews, WikiJunior, and WikiVoyage. 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 9 demonstrate that the factor that affects the most translation quality is the language we translate in and out of (with Arabic being the most challenging and English having the highest scores) rather than the domain. WikiNews is the easiest domain with slightly higher spBLEU, and WikiVoyage is the hardest domain, with an average spBLEU score lower by one point compared to WikiNews. We hypothesize that news-related content is often written in a certain fairly consistent, journalistic style, which could ease the challenge of translation, while WikiVoyage might be a little harder because it has more named entities of local regions of the world which might be harder to translate correctly. However, overall, there are no large differences in performance between domains.

Performance by Language Family. We also group languages into eleven families based on general language families[13] and report in Table 4 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. In fact, even translation within the language family (see values in the diagonal) works very poorly. For these languages, translating to/from Germanic and Romance languages works better. In general, Germanic, Romance, and Balto-Slavic are the language families that yield the largest spBLEU scores (above 10 spBLEU points in average). For these latter languages translation within the language family works the best. In this case, M2M-124 obtains

[13]Note: We define language subgroups for analysis purposes only. These are based on general language families, but are not completely aligned with an agreed upon linguistic taxonomy.

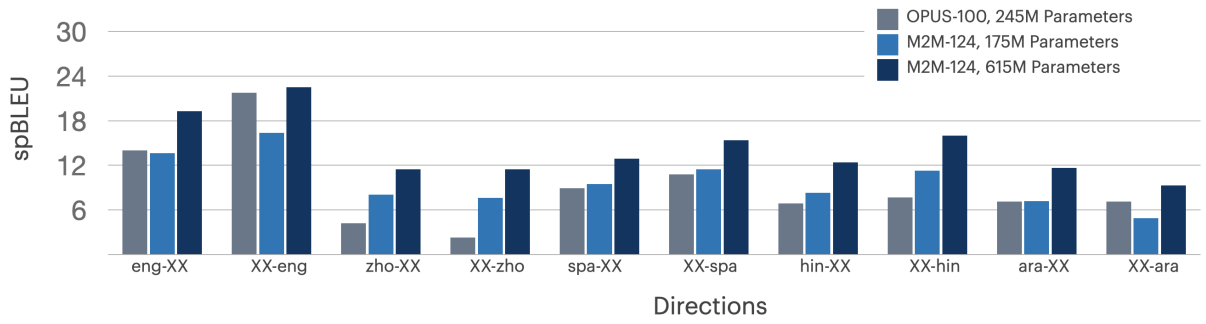

Figure 6: **Comparison between OPUS-100 and M2M-124** on several one-to-many and many-to-one translation tasks using five languages: English, Chinese, Spanish, Hindi, and Arabic. In each case, spBLEU is averaged over all languages in the set. Since the open-source OPUS-100 model covers only 80 languages of FLORES-101, we restrict the evaluation to only these languages in order to make a fair comparison.

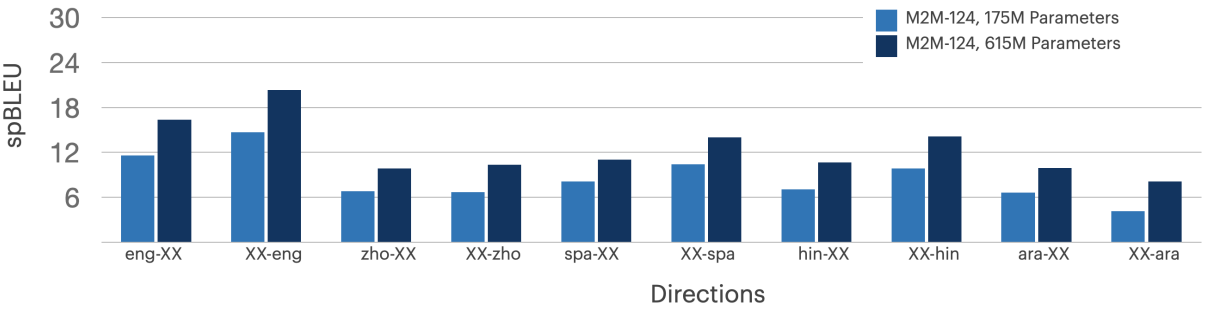

Figure 7: **Full results of M2M-124 Models** on several one-to-many and many-to-one translation tasks using five languages: English, Chinese, Spanish, Hindi, and Arabic. In each case, spBLEU is averaged over all languages in the set (all the remaining 100 languages of FLORES-101).

an spBLEU score above 20. Overall, translation between all languages in a many-to-many fashion requires improvement, as evidenced by the overall quite low average scores.

6.4.2 Comparison of Various Systems.

We end by comparing various baseline systems, to understand the performance of some existing models on FLORES-101.

Comparison to 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 6. 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. Overall, 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. For all systems we evaluated, translation to and from English works the best, while translation to and from Chinese and Arabic struggles the most. In general, spBLEU scores are relatively low, suggesting ample room for improvement and need for further research in this area.

We next display results of M2M-124 175M parameters and 615M parameters on the full set of FLORES-101 languages. This is shown in Figure 7. Comparing results with Figure 6, 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 Selected Masakhane Models. 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. Many of these models are created by people who speak these languages. Further, focusing on specific directions of interest rather than multilingual models could produce specialized models with potentially higher quality.

	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 11: **spBLEU of various models open sourced by Masakhane-MT and M2M-124**. We evaluate models on translating from English to six different African languages. We compare against the M2M-124 615M parameter model.

Masakhane is a participatory research effort that focuses on African NLP. We end by comparing our M2M-124 model with several publicly available models from the `Masakhane-MT` repository. We evaluate models from English to the following languages: Yoruba[14], Zulu[15], Swahili[16], Shona[17], Nyanja[18] and Luo[19]. The Masakhane models are trained on the JW300 dataset.

Results are shown in Table 11. 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 either similar or worse performance than M2M-124. Overall, all languages besides Swahili require significant improvement. Note that in many African countries, a large number of local and regional languages are spoken. We hope that FLORES-101 can be used to develop non-English-centric models that directly translate between African languages.

7 Conclusion

The potential to develop translation systems for languages all over the world is hindered by lack of reliable, high quality evaluation. Without the fundamental ability to measure translation quality, it is impossible to develop, iterate, and test various models and techniques. Particularly for low-resource languages where there is little available training data, new methods and algorithms must be developed to improve translation of these languages. In this work, we create and open-source FLORES-101, an evaluation benchmark covering 101 languages.

FLORES-101 supports many-to-many evaluation, meaning any of 10,100 language directions can be evaluated. With rich metadata, it also supports multimodal translation via images, and document-level translation. Unlike many other multilingual datasets, FLORES-101 is fully translated by humans using a detailed process with numerous quality control checks, including human evaluation during dataset creation.

Beyond translation, FLORES-101 can be used to evaluate tasks such as sentence and document classification, language identification, and multilingual domain adaptation. We hope that the release of this dataset and our baseline M2M models will be useful for the community.

We hope to continue to expand the number of languages covered in FLORES-101 and make the test set available to various community efforts to improve translation systems in shared tasks such as those from the Workshop on Machine Translation.

8 Acknowledgments

We'd like to thank Michael Auli and Sergey Edunov for enlightening discussions and advice. We'd like to thank Mona Diab and Denise Diaz for consulting on specific languages and providing invaluable guidance on translation quality. We'd like to thank Xian Li and Yuqing Tang for helping select the original sentences for translation as part of FLORES. Finally, we'd like to thank Brian Bui for helping with the organization of the data collection effort. We thank all of the translators and human evaluators, as well as the translation and quality assurance agencies we worked with, for helping create FLORES-101.

[14] https://github.com/masakhane-io/masakhane-mt/tree/master/benchmarks/en-yo/jw300-baseline-improve

[15] https://github.com/masakhane-io/masakhane-mt/tree/master/benchmarks/en-zu/jw300-baseline

[16] https://github.com/masakhane-io/masakhane-mt/tree/master/benchmarks/en-sw/fine-tuned-jw300-baseline

[17] https://github.com/masakhane-io/masakhane-mt/tree/master/benchmarks/en-sn/jw300-shona-baseline

[18] https://github.com/masakhane-io/masakhane-mt/tree/master/benchmarks/en-nya/jw-300-baseline

[19] https://github.com/masakhane-io/masakhane-mt/tree/master/benchmarks/en-luo/fine-tuned-jw300-baseline

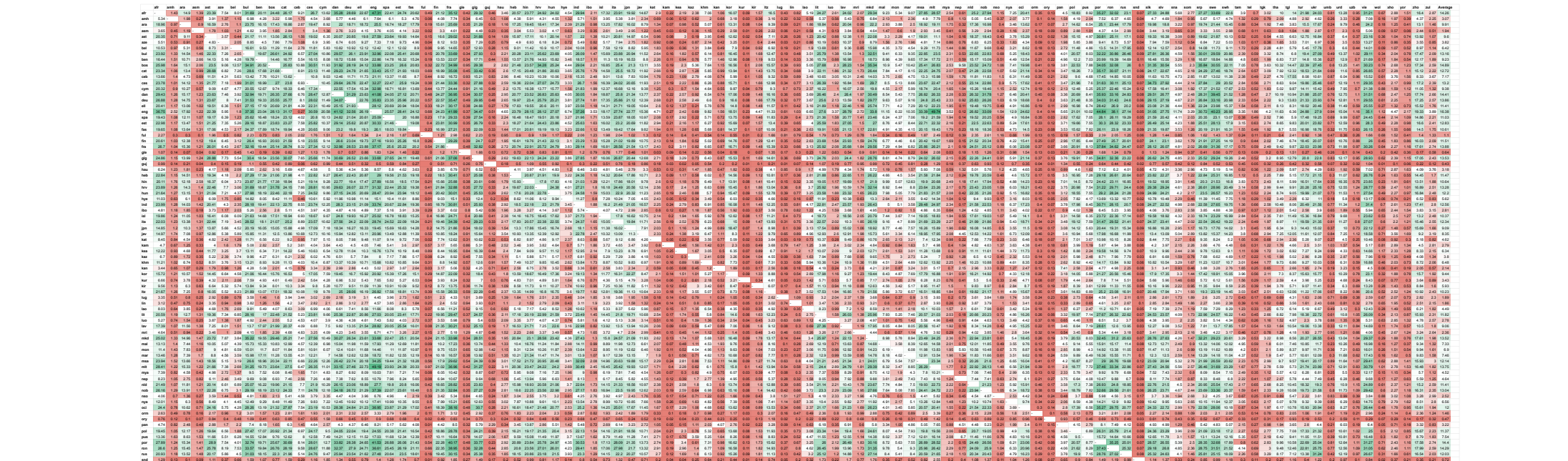

Figure 8: spBLEU of the M2M MMT model on all the language pairs of FLORES-101 dev-test set. Cell (i,j) reports spBLEU for translating from language i to language j. Therefore, each column shows spBLEU for translating in the same target language but using various source languages. Vice versa, each row shows spBLEU for translating into various target languages when starting from the same source language.

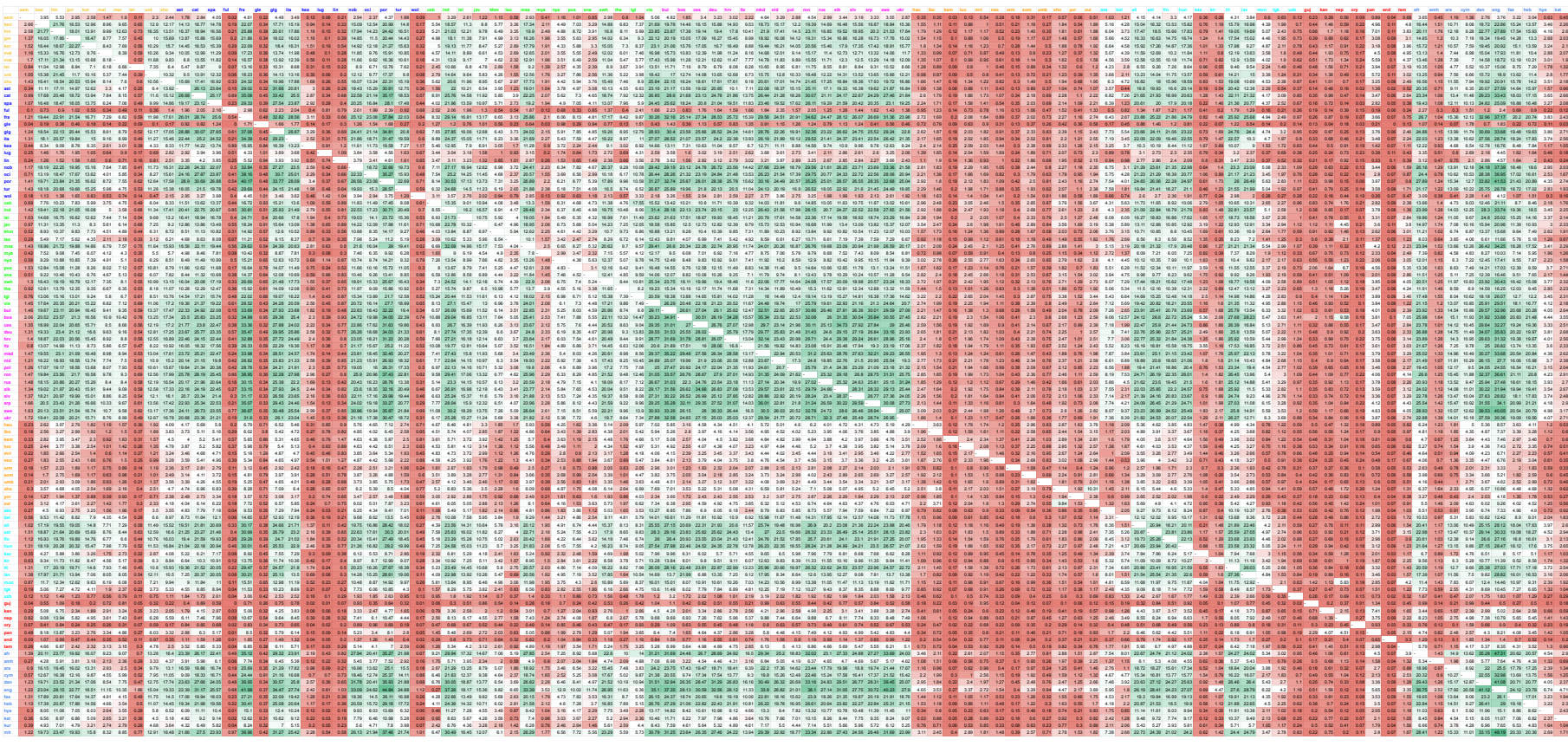

Figure 9: This is the same data as in Fig. 8, except that rows and columns have been organized according to spectral clustering with 8 clusters (identified by the color, see first row and first column).

References

Jade Abbott and Laura Martinus. 2019a. Benchmarking neural machine translation for southern african languages. In *Proceedings of the 2019 Workshop on Widening NLP*, pages 98–101.

Jade Abbott and Laura Martinus. 2019b. 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.

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 DMA Ali, Andrew Caines, and Jaimito LA 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*.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018. Unsupervised statistical machine translation. In *Empirical Methods in Natural Language Processing (EMNLP)*.

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*, pages 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, et al. 2020a. Findings of the 2020 conference on machine translation (wmt20). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1–55.

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.

LoÃ¯c Barrault, Magdalena Biesialska, OndÅ™ej Bojar, Marta R. Costa-jussÃ, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo, Nikola LjubeÅ¡iÄ‡, 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, Online. Association for Computational Linguistics.

Ond rej 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.

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.

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.

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.

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 FP Dossou and Chris C Emezue. 2020. Ffr v1. 1: Fon-french neural machine translation. *arXiv preprint arXiv:2006.09217*.

Bonaventure FP 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, et al. 2021. Americasnli: Evaluating zero-shot natural language understanding of pretrained multilingual models in truly low-resource languages. *arXiv preprint arXiv:2104.08726*.

Ahmed El-Kishky, Vishrav Chaudhary, Francisco Guzman, and Philipp Koehn. 2020. CCAligned: A massive collection of cross-lingual web-document pairs. In *Proc. of EMNLP*.

Miquel Esplà, Mikel Forcada, Gema Ramírez-Sánchez, and Hieu Hoang. 2019. ParaCrawl: Web-scale parallel corpora for the languages of the EU. In *MT Summit*, pages 118–119.

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, et al. 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, et al. 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.

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*.

Philipp Koehn. 2005. 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*.

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*.

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Phrase-based & neural unsupervised machine translation. In *Empirical Methods in Natural Language Processing (EMNLP)*.

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.

Ling Liu, Zach Ryan, and Mans Hulden. 2021. The usefulness of bibles in low-resource machine translation. In *Proceedings of the Workshop on Computational Methods for Endangered Languages*, volume 1, pages 44–50.

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.

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*.

Garrett Nicolai, Edith Coates, Ming Zhang, and Miika Silfverberg. 2021. Expanding the jhu bible corpus for machine translation of the indigenous languages of north america. In *Proceedings of the Workshop on Computational Methods for Endangered Languages*, volume 1, pages 1–5.

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

Bojar Ondrej, Rajen Chatterjee, Federmann Christian, Graham Yvette, Haddow Barry, Huck Matthias, Koehn Philipp, Liu Qun, Logacheva Varvara, Monz Christof, et al. 2017. Findings of the 2017 conference on machine translation (wmt17). In *Second Conference onMachine Translation*, pages 169–214. The Association for Computational Linguistics.

K. Papineni, S. Roukos, T. Ward, and W.J. Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*.

Matt Post. 2018. A call for clarity in reporting bleu scores. *arXiv preprint arXiv:1804.08771*.

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, et al. 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.

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.

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*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 86–96.

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).

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*, pages 3104–3112.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning. *arXiv preprint arXiv:2008.00401*.

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. 2012. Parallel data, tools and interfaces in opus. In *Proceedings of the 8th International Conference on Language Resources and Evaluation*.

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.

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*.

Qi Ye, Sachan Devendra, Felix Matthieu, Padmanabhan Sarguna, and Neubig Graham. 2018. When and why are pre-trained word embeddings useful for neural machine translation. In *HLT-NAACL*.

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.

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.

A Translation Quality Guidelines

Severities:

- **Critical Errors** are issues that render the content unfit for use. An error is only critical if it seriously distorts the meaning of the source, in such a way that it becomes completely incomprehensible or that the essence of the message is lost.
- **Major Errors** may confuse or mislead the user or hinder proper use of the product/service due to significant change in meaning or appear in a visible or important part of the content.
- **Minor Errors** don't lead to loss of meaning and wouldn't confuse or mislead the user but would be noticed, would decrease stylistic quality, fluency or clarity, or would make the content less appealing.

Error Categories:

1. **Grammar** — Noncompliance with target language's grammar rules. Grammar errors may be at the word or sentence level. Types of grammar errors may include:
    - Incorrect person, number or case: the person, number or in the translation does not match the person, number or case of the source text.
    - Incorrect tense: the tense used in the translation does not correspond to the tense used in the source.
    - Incorrect subject/verb agreement: The subject and verb of a sentence must agree with one another in number whether they are singular or plural. If the subject of the sentence is singular, its verb must also be singular; and if the subject is plural, the verb must also be plural.
    - Incorrect use of singular or plural: if a noun in the source text is plural, the corresponding noun and its qualifiers must be plural in the translation.
    - Incorrect word order: word order of the translation is non-standard in the target language, or the translator has made a preferential change to the word order of the source.
2. **Punctuation** — Punctuation is missing, non-standard in the target language or inconsistent with the source punctuation.
3. **Spelling** — Incorrect spelling in the target language. Types of spelling errors may include:
    - Use of the wrong homophone for the context e.g. 'bare with me'
    - Typos
    - Incorrect use of accents
4. **Capitalization** — Noncompliance with target language rules e.g. not capitalising the start of a sentence or a proper noun.
5. **Addition/Omission** — An essential element from the source text is missing in the translation or unnecessary/superfluous elements are present in the translation but were not originally present in the source text.
6. **Mistranslation** — Source text has not been translated faithfully. Types of mistranslation errors may include:
    - Incorrect interpretation of the source text
    - Literal translations and mistranslations of phrasal verbs, rendering incorrect meaning in the target
    - Lack of disambiguation of ambiguous terms
    - Using a subpar word
7. **Unnatural Translation** — Translation does not sound natural to a native speaker of the target language. Source text is translated word for word, rendering the translation unnatural, or is grammatically correct but unnatural to a native speaker.
8. **Untranslated Text** — Words are left untranslated from the source text. This is when there are words in the source language present in the translation which should have been translated into the target language.
9. **Register** — The style or register of the translation is inconsistent with the source and context.

B Additional Results

Details of Spectral Clusters The list of clusters formed by spectral clustering on spBLEU scores is shown in Table 12.

Comparison of Many-to-Many with English-Centric Pivoting. We compare the need of evaluation in a truly many-to-many sense. Instead of creating multilingual models that can translate directly

Cluster	Languages
1	Assamese, Bengali, Hindi, Japanese, Korean, Malayalam, Marathi, Telugu, Urdu, Chinese
2	Asturian, Catalan, Spanish, Fula, French, Irish, Galician, Italian, Kabuverdianu, Lingala, Norwegian, Occitan, Portuguese, Turkish, Wolof
3	Cebuano, Indonesian, Icelandic, Javanese, Khmer, Lao, Malay, Nyanja, Pashto, Shona, Swahili, Thai, Tagalog, Vietnamese
4	Bulgarian, Bosnian, Czech, German, Luxembourgish, Macedonian, Romanian, Russian, Slovak, Slovenian, Swedish, Ukrainian
5	Hausa, Igbo, Kamba, Luo, Maori, Northern Sotho, Oromo, Somali, Umbundu, Xhosa, Yoruba, Zulu
6	Azerbaijani, Belarusian, Greek, Estonian, Finnish, Hungarian, Kazakh, Kyrgyz, Lithuanian, Latvian, Mongolian, Tajik, Uzbek
7	Gujarati, Kannada, Nepali, Oriya, Punjabi, Sindhi, Tamil
8	Afrikaans, Amharic, Arabic, Welsh, Danish, English, Farsi, Armenian, Hebrew, Georgian, Kurdish, Maltese

Table 12: **Language clusters after applying spectral clustering on the full spBLEU matrix**: Interestingly, the spectral clustering identifies several clusters that are reminiscent of world regions, where these languages are often spoken together.

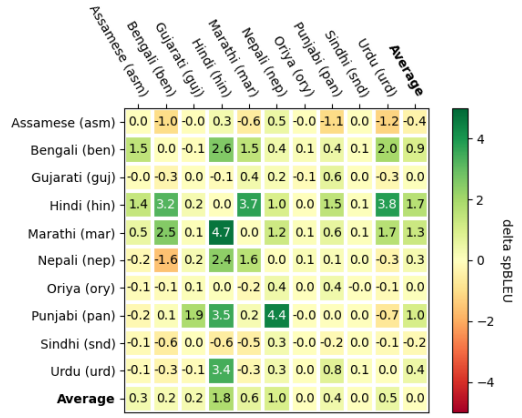

Figure 10: **Performance between Many-to-Many direction translation and English-Centric Pivoting.** We compare the difference in spBLEU (positive indicates direct translation has stronger performance) for 10 Indic languages. The results are computed using the M2M-124 615M model.

between any pair of languages, *pivoting* through English is also possible. Pivoting works by first translating from language X into English, then from English to language Y, instead of translating from X to Y. FLORES-101 supports the evaluation and comparison of these strategies. Unlike previous work such as Fan et al. (2020), which was unable to evaluate all directions of their many-to-many model, FLORES-101 enables evaluation of all 101 x 101 pairs.

In Figure 10, we compare direct translation with English-Centric Pivoting for 10 Indic languages: Assamese, Bengali, Gujarati, Hindi, Marathi, Nepali, Oriya, Punjabi, Sinhala, and Urdu. The spBLEU difference between direct translation and English pivoting is displayed in the heatmap. Overall, we see gains through 80% of the directions by translating directly in a many-to-many fashion. Some directions have gains of more than 3 spBLEU, while a majority of the quality decrease from pivoting is less than 1 spBLEU.

tSNE of Model Embeddings. We examine the similarity of various languages by visualizing the tSNE of language embedding of the trained M2M-124 615M model. Unlike spectral clustering, this examination is a reflection of the model embeddings, rather than the spBLEU score. Figure 11 shows that the languages belonging to the same language family are often grouped together, clustered next to each other.

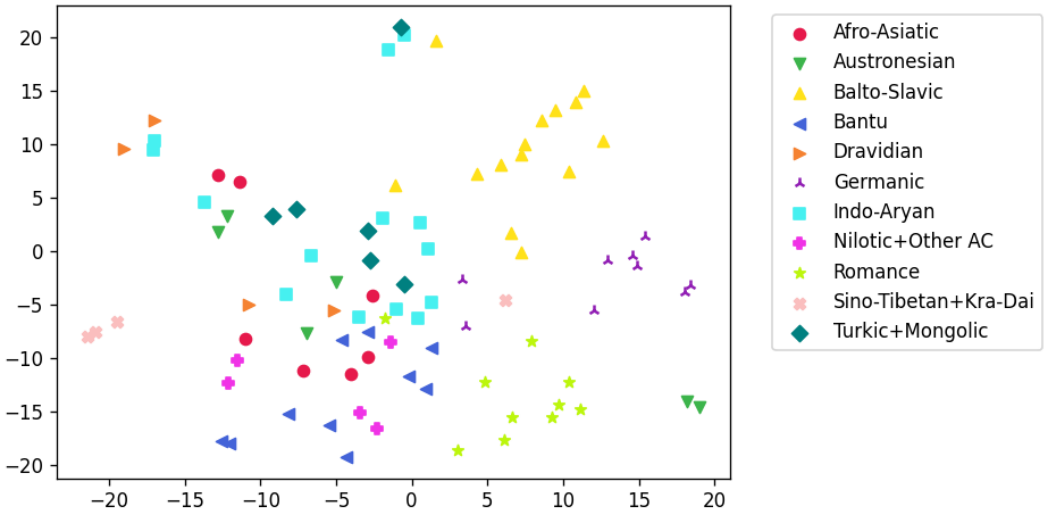

Figure 11: **tSNE plot of Language Embeddings.** We embed the data of various languages with our model and examine by language subgrouping. Oftentimes, languages in the same subgrouping cluster together.