A Shocking Amount of the Web is Machine Translated: Insights from Multi-Way Parallelism

Brian Thompson,[*][1] **Mehak Preet Dhaliwal,**[†][2] **Peter Frisch,**[1] **Tobias Domhan,**[3] **Marcello Federico**[1]

[1]AWS AI Labs   [2]UC Santa Barbara   [3]Amazon

briant@amazon.com

Abstract

We show that content on the web is often translated into many languages, and the low quality of these multi-way translations indicates they were likely created using Machine Translation (MT). Multi-way parallel, machine generated content not only dominates the translations in lower resource languages; it also constitutes a large fraction of the total web content in those languages. We also find evidence of a selection bias in the type of content which is translated into many languages, consistent with low quality English content being translated en masse into many lower resource languages, via MT. Our work raises serious concerns about training models such as multilingual large language models on both monolingual and bilingual data scraped from the web.

1 Introduction

Modern AI is enabled by huge amounts of training data, typically several hundred billion tokens to a few trillion tokens (Sun et al., 2021; Chowdhery et al., 2023; Touvron et al., 2023; Almazrouei et al., 2023). Training at this scale is only possible with web-scraped data.

We explore the effects that the long-term availability of low cost Machine Translation (MT) has had on the web.[1] We show that content on the web is often translated into many languages, and the quality of these multi-way translations indicates they were primarily created using MT: see Figure 1. Machine generated, multi-way parallel translations not only dominate the total amount of translated content on the web in lower resource languages where MT is available, it also constitutes
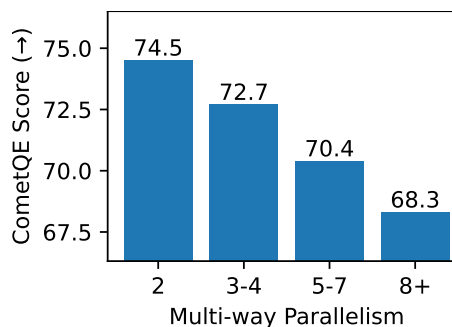

Figure 1: The more languages a sentence has been translated into ("Multi-way Parallelism"), the lower quality the translations are, suggesting a higher prevalence of machine translation. See § 4.4 for more details.

a *large fraction of the total web content* in those languages. We also find evidence of a selection bias in the *type* of content which is translated into many languages, and therefore over represented in lower resource languages: This content is shorter, more predictable, and has a different topic distribution compared to content translated into a single language. A limited investigation suggests this selection bias is the result of low quality content generated in English (likely produced to generate ad revenue) and translated en masse into many lower resource languages via MT (again, likely to generate ad revenue).

Our findings raise numerous concerns for multilingual model builders: Fluency (especially across sentences) and accuracy are lower for MT data,[2] which could produce less fluent models with more hallucinations, and the selection bias indicates the data may be of lower quality, even before considering MT errors. Data quality is crucial in Large Language Model (LLM) training, where high quality corpora like books and Wikipedia articles are

---

[*]Corresponding author

[†]Work conducted during an internship at Amazon.

[1]Free MT has been available online since late 1997 (Gaspari and Hutchins, 2007), around the same time that MT researchers began scraping the web for training data (Resnik, 1998), and commercial systems have been available since the 1970s (Hutchins, 1995).

[2]MT technology has improved dramatically over the last decade, but still falls short of human quality (Freitag et al., 2023). MT content has been added to the web over many years using MT systems available at the time, so much of the MT on the web is likely very low quality by modern standards.

typically upsampled several times (Brown et al., 2020; Gao et al., 2020; Rae et al., 2021; Le Scao et al., 2022).

Our findings also help to explain why low-resource MT (Khan et al., 2017; Duh, 2018; NLLB Team et al., 2022) is challenging, and why filtering noise (Khayrallah and Koehn, 2018) from web-scraped bitext (Junczys-Dowmunt, 2018; Chaudhary et al., 2019) is beneficial for MT training (Koehn et al., 2018, 2019, 2020; Sloto et al., 2023).

To enable analysis, we create the largest multi-way corpus to date, consisting of 6.4B unique sentences in 90 languages. We release code to reproduce our corpus and analysis.[3]

2 Related Work

Our work is inspired by several recent efforts which seek to understand the characteristics of large scale corpora (Mehmood et al., 2017; Dodge et al., 2021; Kreutzer et al., 2022; Brannon et al., 2023). Many works have detected machine translation (Kurokawa et al., 2009; Arase and Zhou, 2013; Aharoni et al., 2014), but we are not aware of prior work using multi-way parallelism to do so. Freitag and Firat (2020) explored multi-way parallelism with the goal of improving multilingual MT.

Exploring multi-way parallelism on the web requires a curated representation of translated content from the web. We build upon ccMatrix (Schwenk et al., 2021), which is in turn based on Common Crawl.[4] Common Crawl is a long running web-scraping project which maintains a free, open source repository of web-scraped data. ccMatrix is created by embedding Common Crawl sentences into a multilingual space using LASER (Artetxe and Schwenk, 2019) and then finding bilingual translation pairs using fast approximate nearest neighbor search (Johnson et al., 2019). We choose ccMatrix over a corpus from a traditional bitext mining process of document alignment (Resnik and Smith, 2003; Buck and Koehn, 2016; Thompson and Koehn, 2020) followed by sentence alignment (Gale and Church, 1993; Sennrich and Volk, 2010; Thompson and Koehn, 2019) because it is the largest corpus available at the time of writing (in both number of sentences and language coverage).

3 Corpus Creation: MWccMatrix

We create a multi-way parallel representation of the web, consisting of translation *tuples* containing *two or more* sentences in different languages which are translations of each other.[5] As a trivial example, ("hello", "hola") in English-Spanish and ("hello", "olá") in English-Portuguese combine to make (En:"hello", Es:"hola", Pt:"olá"). We denote this corpus Multi-Way ccMatrix (MWccMatrix).

We iterate through all bitext in ccMatrix, from highest to lowest LASER margin score, adding sentence pairs as new tuples in MWccMatrix when neither sentence is already in the new corpus, and expanding tuples already in the new corpus when one sentence or the other (but not both) is already present. This deduplicates the corpus (i.e. adds each unique sentence only once), but allows for more than one sentence in the same language to be added to a given tuple, which tend to differ primarily in punctuation/capitalization (i.e. near duplicates). Therefore, we remove all but the first sentence added to each tuple in a given language. Deduplication across language pairs brings the total number of sentences down from 21.7B total sentences (10.9B sentence pairs) to 7.9B unique sentences in 2.2B tuples, and near duplicate removal brings it down to 6.4B. Pseudocode and a description of the optimizations required to make corpus creation tractable are provided in Appendix A.

4 Analysis

4.1 Much of the Web is Translated

We compared the total number of unique sentences (before removing near-duplicates) in MWccMatrix to the total number of unique sentences from the Common Crawl snapshots that the data is based on, as reported by Schwenk et al. (2021). They only report the number of unique sentences for the 54 (of 90) largest resource languages, so we cannot compute the fraction of sentences with one or more translations in the 36 lowest-resource languages. The percentage of unique monolingual sentences which have at least one translation is quite high, even for some high resource languages (e.g. 9.4% of English, 17.5% of French): see Figure 2.

[5]Unless otherwise noted, we use the term "translation" to mean a sentence which appears in a translation tuple – i.e. we do not attempt to distinguish whether that sentence was translated into or out of a given language.
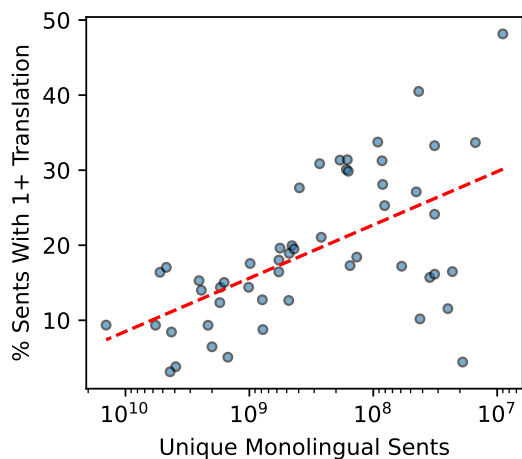
Figure 2: Fraction of the total monolingual data used to create ccMatrix with one or more translation, in the 54 languages for which we can compute it. See Appendix B for a larger plot with language codes.

Parallelism	# tuples	% tuples	# sents	% sents
2	1,368	62.5%	2,736	42.9%
3-4	573	26.2%	1,895	29.7%
5-7	177	8.1%	1,004	15.7%
8+	70	3.2%	745	11.7%
Total	2,188	100.0%	6,379	100.0%

Table 1: MWccMatrix statistics. Numbers in millions. 37.5% of tuples are multi-way parallel, but 57.1% of all sentences come from multi-way parallel tuples.

4.2 Translations on the Web are Highly Multi-way Parallel

Of the 6.38B sentences in our 2.19B translation tuples, 3.63B (57.1%) are in multi-way parallel[6] (3+ languages) tuples: see Table 1. Lower resource languages tend to be more multi-way parallel, with the 10 highest-resourced languages in ccMatrix having an average parallelism of 4.0, and the 10 lowest-resource languages in ccMatrix having an average parallelism of 8.6 (see Appendix C for all languages), and this increase is driven by an increase in highly multi-way parallel (8+) sentences: see Figure 3.

4.3 Multi-way Parallel Data is Shorter and Simpler

We perform monolingual analysis to explore how data varies with multi-way parallelism. We find that more multi-way parallel sentences are shorter

---

[6]We use "multi-way parallelism" (or simply "parallelism") to refer to the size of the translation tuple that that sentence is in. For example, a sentence with parallelism of 5 comes from a tuple of size 5, which contains the given sentence plus translations in 4 other languages.

Parallelism	De	En	Fr	Ja	Zh
2	**95.2**	**103.7**	**96.8**	**25.2**	**27.8**
3-4	90.4	86.0	88.7	23.6	24.8
5-7	82.6	71.2	80.5	23.2	22.8
8+	71.6	59.9	70.0	22.3	19.1

Table 2: Sentence length (in characters) as a function of multi-way parallelism, in several languages. Multi-way parallelism is associated with shorter content.

in length (see Table 2) and have lower perplexity (i.e. are easier to predict) as measured by GPT-2 (Radford et al., 2019): see Figure 4.

4.4 Multi-way Parallel Translations are Lower Quality

We evaluate the quality of translations on the web using Quality Estimation (QE), with the CometQE model (Rei et al., 2022), across different levels of multi-way parallelism. Modern quality estimation methods are nearly on par with reference-based metrics (Freitag et al., 2023) and have been shown to perform well on noisy web data (Peter et al., 2023). As QE does not require human annotation or human references, it allows us to evaluate a very large data sample (1M samples per language pair) and many language pairs.[7]

We find that highly multi-way parallel translations are significantly lower quality (6.2 CometQE points worse) than 2-way parallel translations. This trend is consistent across all 8 language pair directions we considered: see Table 3. Since length could interact with cometQE scores, we verified that these results also hold across sentence length: see Figure 5.

We also investigate how LASER margin score varies with multi-way parallelism. Multi-way parallel data tends to have higher margin scores: see Figure 6. Further investigation reveals that LASER has a strong bias for MT output over human translations (see Table 4), thus LASER margin scores for more multi-way parallel content are consistent with multi-way parallel data being MT. LASER's preference for MT is likely because LASER is based on a small MT model. Similar phenomenon has been observed (Freitag et al., 2021) in the Prism metric (Thompson and Post, 2020a,b), which is also based on an MT model.

---

[7]We select from WMT language pairs as CometQE is trained on WMT annotations, thus we expect CometQE to be most accurate in those language pairs.
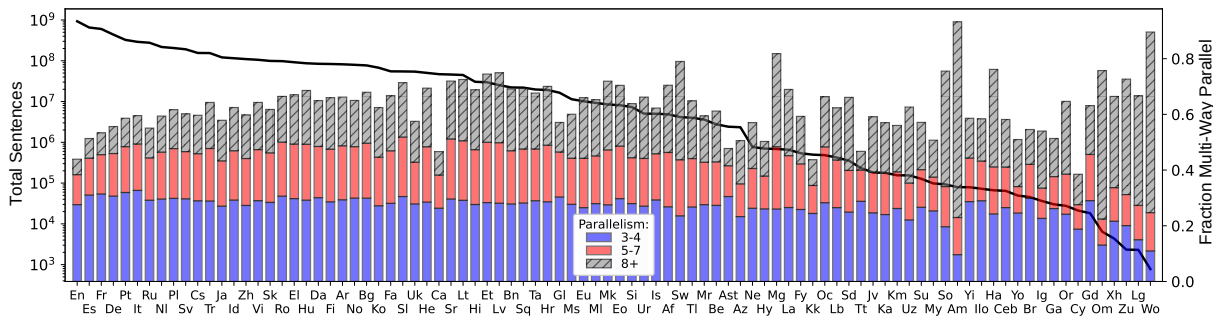
Figure 3: Fraction of parallel data in each language which is multi-way parallel (bar chart, right y-axis) and number of unique sentences (solid black line, left y-axis, log scale) by language (x-axis). Low-resource languages exhibit a dramatic increase in the amount of highly multi-way parallel data (hatched gray bars).

Parallelism	En→De	De→En	Fr→De	De→Fr	En→Ja	Ja→En	En→Zh	Zh→En	AVG
2	**76.5**	**76.1**	**73.3**	**74.6**	**73.6**	**71.9**	**74.8**	**75.4**	**74.5**
3-4	74.3	74.2	72.3	73.7	72.0	70.6	72.1	72.5	72.7 (-1.8)
5-7	71.9	71.8	70.0	71.3	70.5	69.2	69.1	69.5	70.4 (-4.1)
8+	69.7	69.8	67.5	68.6	69.7	68.6	66.1	66.6	68.3 (-6.2)

Table 3: Bitext quality (as measured by CometQE) as a function of multi-way parallelism, for random 1M subsets in various language pairs. Multi-way parallel translations are lower quality. Average scores are visualized in Figure 1.
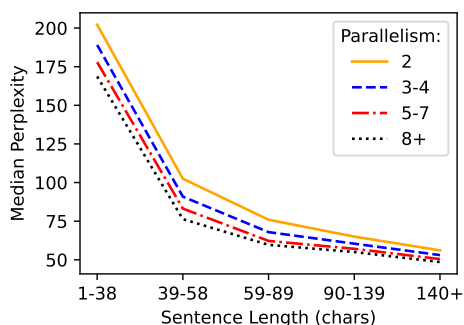

Figure 4: Median perplexity (measured by GPT-2) vs multi-way parallelism, in English. We stratify by sentence length, as shorter content tends to have higher perplexity, likely due to GPT-2 having no or little context for predicting the first few words. More multi-way parallel data has lower perplexity (i.e. easier to predict).
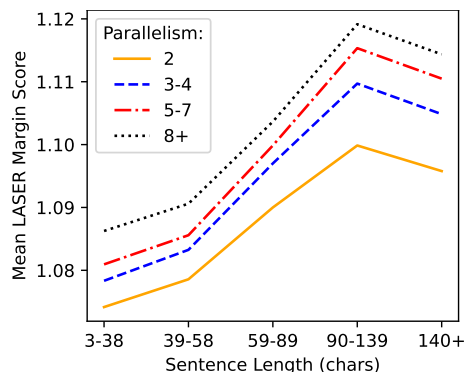

Figure 6: LASER margin scores as a function of multi-way parallelism and sentence length, in English. Trends in other languages that we investigated (French, German, Chinese, Japanese) were very similar (not shown).
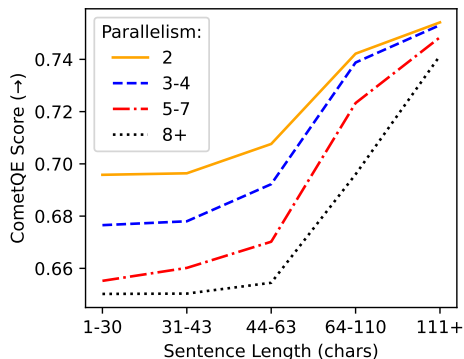

Figure 5: CometQE score vs sentence length (average of source and target, in characters), for Fr-De. Other language pairs (not shown) are very similar. Quality differences between levels of multi-way parallelism holds across sentence length.

Langs	Ref	MT	Δ
En→De	86.7%	89.5% +/- 0.6%	2.8%
De→En	87.4%	90.1% +/- 0.4%	2.8%
Zh→En	82.3%	85.3% +/- 0.3%	3.0%
En→Zh	82.6%	85.8% +/- 0.4%	3.1%
En→Ja	78.8%	82.7% +/- 1.1%	3.9%
De→Fr	87.4%	90.1% +/- 0.4%	2.8%
Fr→De	90.5%	91.4% +/- 0.6%	0.9%

Table 4: LASER cosine similarity between source and human reference ("Ref") vs mean and standard deviation for Online-Y, Online-G, Online-A, Online-W, and Online-B ("MT") from WMT 2022. In cases where there is more than one human reference, we average the cosine similarities. We find that LASER has a bias for MT output, of about 2.8% on average. Note that Ja→En was not included among the WMT language pairs.

Topic	Parallelism			
	2	3-4	5-7	8+
Autos/Vehicles/Transit	2.2	2.0	1.6	1.6
Beauty/Fitness/Health.	6.5	7.2	6.4	6.4
Books/Arts/Entertainment	8.0	7.3	6.3	5.4
Business/Industry/Finance	10.6	7.6	6.0	6.6
Computers/Electronics	3.7	3.4	4.5	3.5
Conversation/Opinion	22.5	29.9	33.3	40.1
Food/Drink	1.7	1.8	1.7	0.9
Hobbies/Leisure	0.9	0.9	1.4	1.0
Home/Garden	0.9	0.7	0.9	0.5
Internet/Telecom	3.2	3.0	2.1	2.6
Jobs/Education	6.6	4.7	5.0	5.0
Law/Government	7.2	6.0	5.2	4.8
News/Information	12.2	11.8	10.5	9.5
Pets/Animals	1.2	1.4	1.2	1.0
Real Estate	1.3	0.8	0.5	0.5
Religion	3.4	4.7	5.6	5.4
Shopping	1.1	1.2	1.2	0.7
Sports/Games	3.2	2.9	3.7	2.2
Travel	2.9	2.6	2.5	1.9
Other/Cannot tell	0.2	0.1	0.5	0.7

Table 5: Percentage of corpus which human annotators classified as each topic, for various levels of parallelism.

4.5 Multi-way Parallel Data has Topical Bias

We had professional linguists classify[8] 10,000 randomly selected English sentences as one of the 20 topics given in Table 5, based on the high-level Topics API categories.[9] We observe a dramatic shift in the distribution of topics when comparing 2-way to 8+ way parallel data, with CONVERSATION & OPINION increasing from 22.5% to 40.1%.

We manually inspected a random sample of 100 highly multi-way parallel sentences from the CONVERSATION & OPINION topic and found them hard to characterize due to the isolated sentences being very short (typically 5-10 words). However, searching the web for the sentences was enlightening: the vast majority came from articles that we characterized as low quality, requiring little or no expertise or advance effort to create, on topics like being taken more seriously at work, being careful about your choices, six tips for new boat owners, deciding to be happy, etc. Furthermore, we were unable to find any translationese or other errors that would suggest the articles were being translated into English (either by human translators or MT), suggesting it is instead being generated in English and translated to other languages.

---

[8]Full annotator guidelines are provided in Appendix D
[9]https://cloud.google.com/natural-language/docs/categories

5 Discussion & Conclusion

Experiments with QE (see § 4.4) strongly suggest that highly multi-way parallel translations are generated by MT. In lower resource languages, *most* translations are multi-way parallel (see § 4.2), suggesting that MT content dominates translation content. Furthermore, a large fraction of the *total* sentences in lower resource languages have at least one translation (see § 4.1), implying that a large fraction of the *total web* in those languages is MT generated.

Several observations point to a selection bias in the *type* of data which is translated into many languages, compared to data translated into a single language: it is shorter and more predictable (see § 4.3), and substantially more likely to be from the CONVERSATION & OPINION topic (see § 4.5). Since translations of this data constitute a substantial portion of the total data in low-resource languages, this bias will also appear in low resource languages.

An investigation into the increase in CONVERSATION & OPINION data suggests that this selection bias is the result of low quality content (likely produced to generate ad revenue) being translated via MT en masse into many lower resource languages (again likely for the purpose of generating ad revenue). It also suggests that such data originates in English and is translated into other languages. Additional investigation would be required to understand if this finding generalizes to other topics, languages, and levels of multi-way parallelism.

Our findings also suggest some ways to address the problem of MT output in web-scraped training data: MT detection, which has typically been proposed to filter bitext, could also help to filter monolingual text in lower resource languages. It also suggests that multi-way parallelism is a promising way to detect low quality, machine translated data, especially in lower resource languages, to filter both bilingual and monolingual data.

Limitations

Our study covers 90 of the most common languages on the web, where MT tends to be available. We would obviously not expect the trends we observe regarding low resource languages to extend into the long tail of low-resource languages that are not currently supported by MT.

All our analysis is performed at the sentence level, because ccMatrix is at the sentence level;

this makes some analysis (e.g. topic analysis) difficult and/or ambiguous. We would have preferred to use a corpus which is aligned at the document level to enable document level analysis and evaluation (Läubli et al., 2018; Toral et al., 2018; Vernikos et al., 2022), but no such corpus is publicly available.

Similarly, our analysis does not take advantage of any cues that might be present in a web page to indicate its content is MT generated. However, in personal communications with the authors of Paracrawl (Bañón et al., 2020), they note that in low-resource languages supported by popular MT systems, simple rules[10] to remove data from common translation plug-ins filter out most of their scraped bitext. This observation is consistent with our findings.

We use CometQE to evaluate translation quality. CometQE is trained on human annotations of translation quality from many years of WMT (Kocmi et al., 2023) evaluations. The web data in our experiments likely has a very different domain distribution than WMT data, and trained metrics like CometQE have been shown to exhibit a performance drop when moving from WMT to other domains (Zouhar et al., 2024).

Our analysis of the web is based on bitext mined from the web. As such, shortcomings or biases in web scraping and bitext mining could affect our results. Common Crawl provides only a sample of the web, and biases have been demonstrated in web scraping (Mehmood et al., 2017; Dodge et al., 2021). Common Crawl follows links within web pages to find new pages, and web pages sometimes have links to translations of the same page in another language, so Common Crawl may be more likely to find web pages which are translations of pages it has already found than other, new pages. This should be mitigated at least in part by combining many Common Crawl snapshots, as is done in ccMatrix.

The 32 snapshots of Common Crawl used in ccMatrix were collected between December 2017 to February 2020 (Schwenk et al., 2021). We are not aware of a more recent, publicly available corpus that would enable this kind of analysis.

The ccMatrix corpus creation process relies on LASER margin scores (as does our process to create MWccMatrix). LASER is known to have lower recall in lower-resource languages (Feng et al., 2022) and as we show in this work (Table 4), has a preference for MT over human translations. ccMatrix also uses approximate nearest neighbor search (Johnson et al., 2019), which trades off some recall performance in order to make searches computationally feasible.

Our analysis by language / language pair relies on automatic language identification (LID). Shortcomings have also been noted in automatic LID, especially in low-resource languages (Caswell et al., 2020; Kreutzer et al., 2022).

Acknowledgements

References

Roee Aharoni, Moshe Koppel, and Yoav Goldberg. 2014. Automatic detection of machine translated text and translation quality estimation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 289–295, Baltimore, Maryland. Association for Computational Linguistics.

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, et al. 2023. The falcon series of open language models. *arXiv preprint arXiv:2311.16867*.

Yuki Arase and Ming Zhou. 2013. Machine translation detection from monolingual web-text. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1597–1607, Sofia, Bulgaria. Association for Computational Linguistics.

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Marta Bañón, Pinzhen Chen, Barry Haddow, Kenneth Heafield, Hieu Hoang, Miquel Esplà-Gomis, Mikel L. Forcada, Amir Kamran, Faheem Kirefu, Philipp Koehn, Sergio Ortiz Rojas, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Elsa Sarrías, Marek Strelec,

Brian Thompson, William Waites, Dion Wiggins, and Jaume Zaragoza. 2020. ParaCrawl: Web-scale acquisition of parallel corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4555–4567, Online. Association for Computational Linguistics.

William Brannon, Yogesh Virkar, and Brian Thompson. 2023. Dubbing in practice: A large scale study of human localization with insights for automatic dubbing. *Transactions of the Association for Computational Linguistics*, 11:419–435.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Christian Buck and Philipp Koehn. 2016. Quick and reliable document alignment via TF/IDF-weighted cosine distance. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 672–678, Berlin, Germany. Association for Computational Linguistics.

Isaac Caswell, Theresa Breiner, Daan van Esch, and Ankur Bapna. 2020. Language ID in the wild: Unexpected challenges on the path to a thousand-language web text corpus. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6588–6608, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Vishrav Chaudhary, Yuqing Tang, Francisco Guzmán, Holger Schwenk, and Philipp Koehn. 2019. Low-resource corpus filtering using multilingual sentence embeddings. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 261–266, Florence, Italy. Association for Computational Linguistics.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. 2021. Documenting large webtext corpora: A case study on the colossal clean crawled corpus. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1286–1305, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Kevin Duh. 2018. The multitarget ted talks task. http://www.cs.jhu.edu/~kevinduh/a/multitarget-tedtalks/.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

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.

Markus Freitag, Nitika Mathur, Chi-kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Tom Kocmi, Frederic Blain, Daniel Deutsch, Craig Stewart, Chrysoula Zerva, Sheila Castilho, Alon Lavie, and George Foster. 2023. Results of WMT23 metrics shared task: Metrics might be guilty but references are not innocent. In *Proceedings of the Eighth Conference on Machine Translation*, pages 578–628, Singapore. Association for Computational Linguistics.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, George Foster, Alon Lavie, and Ondřej Bojar. 2021. Results of the WMT21 metrics shared task: Evaluating metrics with expert-based human evaluations on TED and news domain. In *Proceedings of the Sixth Conference on Machine Translation*, pages 733–774, Online. Association for Computational Linguistics.

William A. Gale and Kenneth W. Church. 1993. A program for aligning sentences in bilingual corpora. *Computational Linguistics*, 19(1):75–102.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. 2020. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.

Federico Gaspari and John Hutchins. 2007. Online and free! ten years of online machine translation: origins, developments, current use and future prospects. In *Proceedings of Machine Translation Summit XI: Papers*, Copenhagen, Denmark.

W John Hutchins. 1995. Machine translation: A brief history. In *Concise history of the language sciences*, pages 431–445. Elsevier.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3):535–547.

Marcin Junczys-Dowmunt. 2018. Dual conditional cross-entropy filtering of noisy parallel corpora. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 888–895, Belgium, Brussels. Association for Computational Linguistics.

Nadeem Jadoon Khan, Waqas Anwar, and Nadir Durrani. 2017. Machine translation approaches and survey for indian languages. *arXiv preprint arXiv:1701.04290*.

Huda Khayrallah and Philipp Koehn. 2018. On the impact of various types of noise on neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 74–83, Melbourne, Australia. Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Philipp Koehn, Benjamin Marie, Christof Monz, Makoto Morishita, Kenton Murray, Makoto Nagata, Toshiaki Nakazawa, Martin Popel, Maja Popović, and Mariya Shmatova. 2023. Findings of the 2023 conference on machine translation (WMT23): LLMs are here but not quite there yet. In *Proceedings of the Eighth Conference on Machine Translation*, pages 1–42, Singapore. Association for Computational Linguistics.

Philipp Koehn, Vishrav Chaudhary, Ahmed El-Kishky, Naman Goyal, Peng-Jen Chen, and Francisco Guzmán. 2020. Findings of the WMT 2020 shared task on parallel corpus filtering and alignment. In *Proceedings of the Fifth Conference on Machine Translation*, pages 726–742, Online. Association for Computational Linguistics.

Philipp Koehn, Francisco Guzmán, Vishrav Chaudhary, and Juan Pino. 2019. Findings of the WMT 2019 shared task on parallel corpus filtering for low-resource conditions. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 54–72, Florence, Italy. Association for Computational Linguistics.

Philipp Koehn, Huda Khayrallah, Kenneth Heafield, and Mikel L. Forcada. 2018. Findings of the WMT 2018 shared task on parallel corpus filtering. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 726–739, Belgium, Brussels. Association for Computational Linguistics.

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Ortiz Suarez, Iroro Orife, Kelechi Ogueji, Andre Niyongabo Rubungo, Toan Q. Nguyen, Mathias Müller, André Müller, Shamsuddeen Hassan Muhammad, Nanda Muhammad, Ayanda Mnyakeni, Jamshidbek Mirzakhalov, Tapiwanashe Matangira, Colin Leong, Nze Lawson, Sneha Kudugunta, Yacine Jernite, Mathias Jenny, Orhan Firat, Bonaventure F. P. Dossou, Sakhile Dlamini, Nisansa de Silva, Sakine Çabuk Ballı, Stella Biderman, Alessia Battisti, Ahmed Baruwa, Ankur Bapna, Pallavi Baljekar, Israel Abebe Azime, Ayodele Awokoya, Duygu Ataman, Orevaoghene Ahia, Oghenefego Ahia, Sweta Agrawal, and Mofetoluwa Adeyemi. 2022. Quality at a glance: An audit of web-crawled multilingual datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72.

David Kurokawa, Cyril Goutte, and Pierre Isabelle. 2009. Automatic detection of translated text and its impact on machine translation. In *Proceedings of Machine Translation Summit XII: Papers*, Ottawa, Canada.

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

Teven Le Scao, Thomas Wang, Daniel Hesslow, Stas Bekman, M Saiful Bari, Stella Biderman, Hady Elsahar, Niklas Muennighoff, Jason Phang, Ofir Press, Colin Raffel, Victor Sanh, Sheng Shen, Lintang Sutawika, Jaesung Tae, Zheng Xin Yong, Julien Launay, and Iz Beltagy. 2022. What language model to train if you have one million GPU hours? In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 765–782, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Muhammad Amir Mehmood, Hafiz Muhammad Shafiq, and Abdul Waheed. 2017. Understanding regional context of world wide web using common crawl corpus. In *2017 IEEE 13th Malaysia International Conference on Communications (MICC)*, pages 164–169.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Jan-Thorsten Peter, David Vilar, Daniel Deutsch, Mara Finkelstein, Juraj Juraska, and Markus Freitag. 2023.

There's no data like better data: Using QE metrics for MT data filtering. In *Proceedings of the Eighth Conference on Machine Translation*, pages 561–577, Singapore. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Jack W Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, et al. 2021. Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Philip Resnik. 1998. Parallel strands: a preliminary investigation into mining the web for bilingual text. In *Proceedings of the Third Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 72–82, Langhorne, PA, USA. Springer.

Philip Resnik and Noah A. Smith. 2003. The web as a parallel corpus. *Computational Linguistics*, 29(3):349–380.

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

Rico Sennrich and Martin Volk. 2010. MT-based sentence alignment for OCR-generated parallel texts. In *Proceedings of the 9th Conference of the Association for Machine Translation in the Americas: Research Papers*, Denver, Colorado, USA. Association for Machine Translation in the Americas.

Steve Sloto, Brian Thompson, Huda Khayrallah, Tobias Domhan, Thamme Gowda, and Philipp Koehn. 2023. Findings of the WMT 2023 shared task on parallel data curation. In *Proceedings of the Eighth Conference on Machine Translation*, pages 95–102, Singapore. Association for Computational Linguistics.

Yu Sun, Shuohuan Wang, Shikun Feng, Siyu Ding, Chao Pang, Junyuan Shang, Jiaxiang Liu, Xuyi Chen, Yanbin Zhao, Yuxiang Lu, et al. 2021. Ernie 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation. *arXiv preprint arXiv:2107.02137*.

Brian Thompson and Philipp Koehn. 2019. Vecalign: Improved sentence alignment in linear time and space. 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 1342–1348, Hong Kong, China. Association for Computational Linguistics.

Brian Thompson and Philipp Koehn. 2020. Exploiting sentence order in document alignment. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5997–6007, Online. Association for Computational Linguistics.

Brian Thompson and Matt Post. 2020a. Automatic machine translation evaluation in many languages via zero-shot paraphrasing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 90–121, Online. Association for Computational Linguistics.

Brian Thompson and Matt Post. 2020b. Paraphrase generation as zero-shot multilingual translation: Disentangling semantic similarity from lexical and syntactic diversity. In *Proceedings of the Fifth Conference on Machine Translation*, pages 561–570, Online. Association for Computational Linguistics.

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

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Giorgos Vernikos, Brian Thompson, Prashant Mathur, and Marcello Federico. 2022. Embarrassingly easy document-level MT metrics: How to convert any pretrained metric into a document-level metric. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 118–128, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Vilém Zouhar, Shuoyang Ding, Anna Currey, Tatyana Badeka, Jenyuan Wang, and Brian Thompson. 2024. Fine-tuned machine translation metrics struggle in unseen domains. *arXiv preprint arXiv:2402.18747*.

A MWccMatrix Creation: Additional Details

A simplified version of the algorithm used to create MWccMatrix is provided in algorithm 1.

In practice, several optimizations were required to make the process tractable. Instead of attempting to sort 10.9B sentence pairs by margin score, we approximate the search by binning margin scores and sorting the data into buckets corresponding to the (binned) margin scores, similar to a radix sort. The sentences are too large to fit in memory, so we represent the sentences as 64 bit hashes. Additionally, our scripts are written in python but we use the cykhash[11] package, which provides a native C int64 to int64 hashmap. Conversion from hashes back to sentences is done in small shards, and the hash→sent mappings required to reconstruct the data are sharded such that only the mappings required for one shard are loaded in memory at one time. Finally, we make extensive use of parallelization (e.g. computing margin score bins, sharding data by margin score bin, hashing sentence pairs, etc).

B Larger Version of Figure 2

A larger version of Figure 2, which includes language codes for each language, is provided in Figure 7. As previously noted, we only have total data sizes for the 54 highest-resource languages, as that is what was reported by Schwenk et al. (2021), so we cannot compute this percentage for the 36 lowest-resource languages used in this study.

C Multi-way Parallelism by Language

Average parallelism for each language is shown in Figure 8.

---

[11] https://github.com/realead/cykhash

Algorithm 1: Algorithm (simplified for comprehension) used to create multi-way parallel corpus.

```
 1  sent2row ← defaultdict(dict)
 2  sort(bitext, key = marginScore, descending = True)
 3  numRows ← 0
 4  for srcTxt, srcLang, tgtText, tgtLang, marginScore in bitext do
 5      if srcTxt not in sent2row[srcLang] and tgtText not in sent2row[tgtLang] then
 7          /* add new sentence pair                                                    */
 8          sent2row[srcLang][srcTxt] ← (numRows, marginScore)
 9          sent2row[tgtLang][tgtTxt] ← (numRows, marginScore)
10          numRows ← numRows + 1
11      else if srcTxt in sent2row[srcLang] then
13          /* srcText in table, join on it                                             */
14          srcRow ← sent2row[srcLang][srcTxt]
15          sent2row[tgtLang][tgtTxt] ← (srcRow, marginScore)
16      else if tgtTxt in sent2row[tgtLang] then
18          /* tgtText in table, join on it                                             */
19          tgtRow ← sent2row[tgtLang][tgtTxt]
20          sent2row[srcLang][srcTxt] ← (tgtRow, marginScore)
22      /* else both sentences already in table (with higher marginScore), do nothing   */
24  /* Invert sent2row                                                                  */
25  row2sent ← defaultdict(dict)
26  for lang in langs do
27      for sent, (row, marginScore) in sent2row[lang].items() do
28          if row in row2sent[lang] then
29              _, oldMarginScore ← row2sent[lang][row]
30          else
31              oldMarginScore ← −1
33          /* When we find duplicates/paraphrases, keep the sentence with the highest score */
34          if marginScore > oldMarginScore then
35              row2sent[lang][row] ← (sent, score)
37  /* Coalesce output translation tuples                                               */
38  output ← []
39  for row in range(numRows) do
40      translations ← dict()
41      for lang in langs do
42          if row in row2sent[lang] then
43              translations[lang] ← row2sent[lang][row]
44      output.append(translations)
```
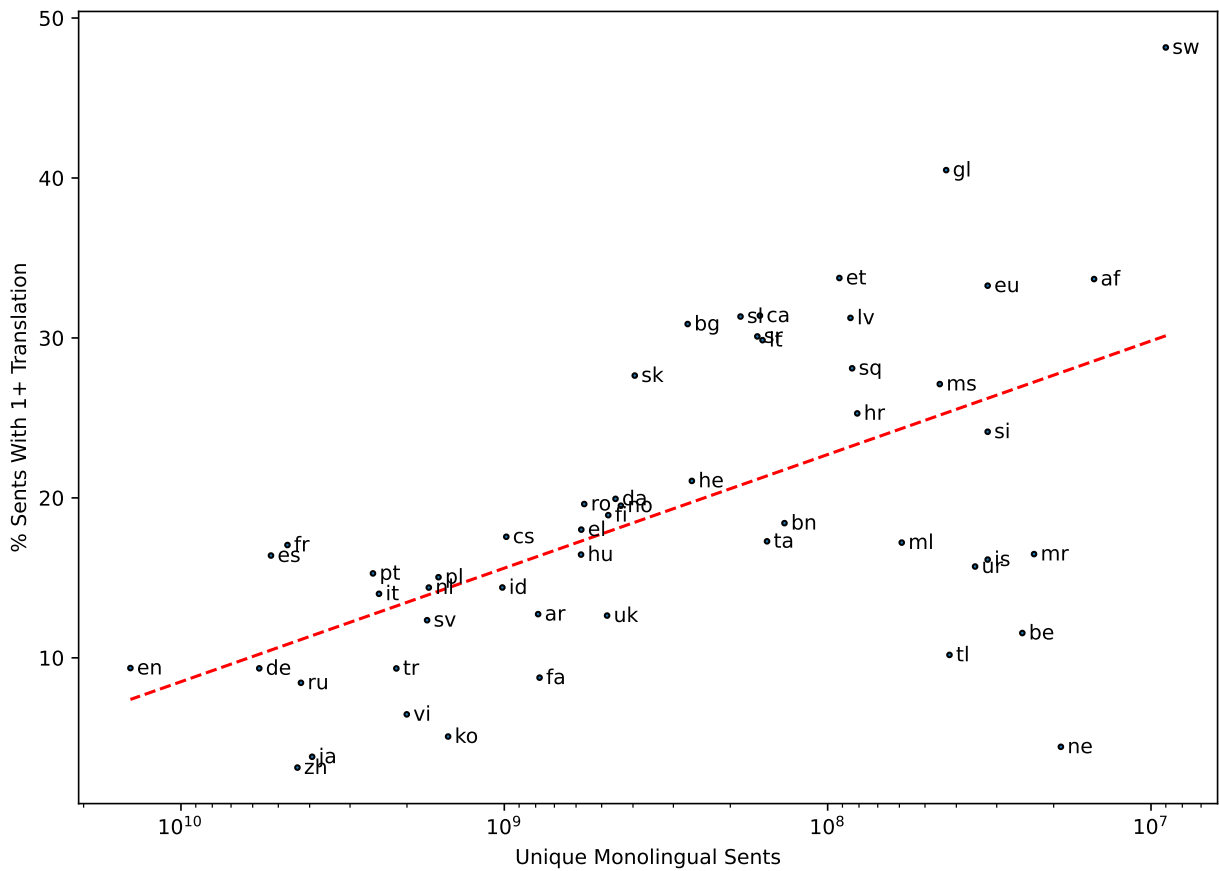
Figure 7: Percentage of unique monolingual sentences which have at least one translation, in each language for which we have the data to compute it.
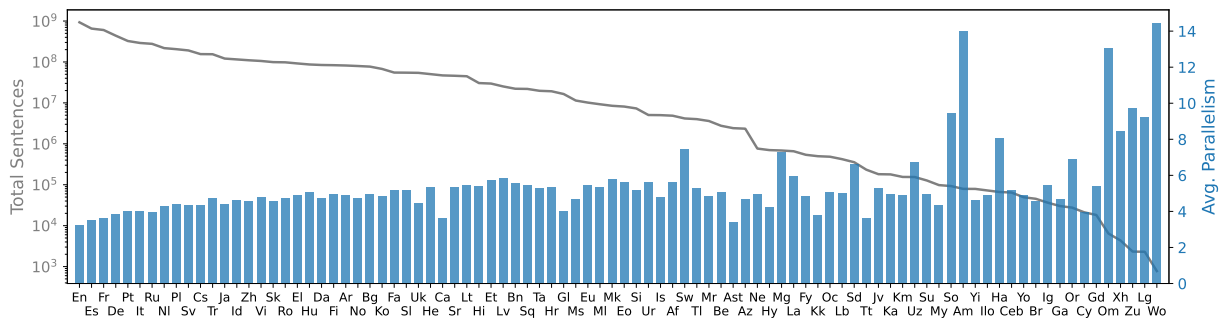

Figure 8: Average multi-way parallelism (blue bars, right y-axis) and number of unique sentences (gray line, left y-axis, log scale) by language (x-axis). Lower-resource languages tend to be more multi-way parallel.

D Topic Analysis Annotation Guidelines

Task: Please identify the most relevant topic for each sentence using the pre-defined list. Assign the correct label to each sentence. Make sure to familiarize yourself with the list before working on the task.

Index	Topic
1	Autos/Vehicles/Transit
2	Beauty/Fitness/Health.
3	Books/Arts/Entertainment
4	Business/Industry/Finance
5	Computers/Electronics
6	Conversation/Opinion
7	Food/Drink
8	Hobbies/Leisure
9	Home/Garden
10	Internet/Telecom
11	Jobs/Education
12	Law/Government
13	News/Information
14	Pets/Animals
15	Real Estate
16	Religion
17	Shopping
18	Sports/Games
19	Travel
20	Other/Cannot tell

Here is the list of sub-topics that may help you develop a better idea of what belongs to each of these topics: https://github.com/patcg-individual-drafts/topics/blob/main/taxonomy_v2.md

Note:

1. Do differentiate between a domain and a topic. A topic of the sentence is the main idea of the sentence. Where this sentence belongs is the domain. In this task we are classifying topics.

   (a) "Aiden was once a warrior who placed complete faith in his own abilities." - this belongs to literature/creative writing domain, but the topic of the sentence is CONVERSATION & OPINION.

   (b) "i keep telling you to leave me alone, this forum is not the right place for hate" - the domain is media, but the topic is CONVERSATION & OPINION.

2. If needed, please do a quick search of the sentence to identify the topic. Do limit the search to a quick scan of search results no longer than 30 sec.

3. If a sentence fits more than one topic equally and you cannot decide between the two, then select a primary and a secondary topic. Add a comment if needed to explain. Try not to abuse this option and always try to choose one.