

From Unaligned to Aligned: Scaling Multilingual LLMs with Multi-Way Parallel Corpora

Yingli Shen^{1*}, Wen Lai^{2,3*}, Shuo Wang¹, Ge Gao⁴
Kangyang Luo¹, Alexander Fraser^{2,3}, Maosong Sun^{1,5†}

¹ Department of Computer Science and Technology, Tsinghua University

² Technical University of Munich ³ Munich Center for Machine Learning

⁴ Minzu University of China ⁵ Institute for AI, Tsinghua University

syl@mail.tsinghua.edu.cn, wen.lai@tum.de

Abstract

Continued pretraining and instruction tuning on large-scale multilingual data have proven to be effective in scaling large language models (LLMs) to low-resource languages. However, the unaligned nature of such data limits its ability to effectively capture cross-lingual semantics. In contrast, *multi-way parallel data*, where identical content is aligned across multiple languages, provides stronger cross-lingual consistency and offers greater potential for improving multilingual performance. In this paper, we introduce a large-scale, high-quality multi-way parallel corpus, TED2025, based on TED Talks. The corpus spans 113 languages, with up to 50 languages aligned in parallel, ensuring extensive multilingual coverage. Using this dataset, we investigate best practices for leveraging multi-way parallel data to enhance LLMs, including strategies for continued pretraining, instruction tuning, and the analysis of key influencing factors. Experiments on six multilingual benchmarks show that models trained on multi-way parallel data consistently outperform those trained on unaligned multilingual data.¹

1 Introduction

Large language models (LLMs) have demonstrated remarkable performance on various tasks in high-resource languages (Huang et al., 2024; Qin et al., 2024). However, their performance still lags behind in low-resource languages (Huang et al., 2023; Lai et al., 2023a). To bridge this gap, recent efforts have focused on continued pretraining (Ji et al., 2024; Groeneveld et al., 2024) and instruction tuning (Lai et al., 2024; Üstün et al., 2024), utilizing large-scale unaligned multilingual text. However, these methods do not take full advantage of the explicit many-to-many alignments present in *multi-way parallel corpora*, which have been shown to

improve cross-lingual representations in multilingual NLP (Qi et al., 2018; Freitag and Firat, 2020; Xu et al., 2022; Wu et al., 2024). More recently, Mu et al. (2024) demonstrated that multi-way parallel inputs can also enhance in-context learning (Dong et al., 2024). However, scaling multilingual LLMs using multi-way parallel data remains underexplored.

Existing multi-way parallel datasets typically cover only a limited number of languages, domains and levels of parallelism (see Table 1). In contrast, TED Translators², a global community translating TED talk transcripts into over 100 languages, provides consistently high-quality, human-verified translations and serves as an ideal source for a large-scale multi-way parallel corpus. However, the largest TED-based datasets (Qi et al., 2018; Reimers and Gurevych, 2020) have not been updated since 2020, limiting their utility for LLM training and potentially exacerbating hallucinations (Ji et al., 2023).

To address these limitations, we introduce TED2025, a large-scale, high-quality multi-way parallel corpus derived from the latest TED talks. TED2025 covers 113 languages, 352 domain labels, and supports up to 50-way parallelism. Compared to existing resources, it offers more frequent updates and significantly broader coverage across both languages and domains, thereby strengthening the data foundation for multilingual LLM training.

Utilizing TED2025, we investigate three key research questions:

RQ1: How does fine-tuning on multi-way parallel data compare to training on unaligned multilingual text in terms of zero-shot cross-lingual transfer and representation alignment?

RQ2: Which strategies for selecting parallelism in multi-way parallel data (e.g., degree of parallelism and language subsets) lead to the greatest

*Equal contribution.

†Corresponding author.

¹<https://github.com/yl-shen/multi-way-llm>

²<https://www.ted.com/participate/translate>

Dataset	Source	#Langs	#Domains	Max Parallelism	Type	Collection Method	Date Range	Open-source?
Bible (Christodouloupoulos and Steedman, 2015)	Bible	100	1	55	Training	Human Translator	2015	✓
UN Corpus (Ziems et al., 2016)	UN	6	1	6	Training	Human Translator	2016	✓
MMCR4NLP (Dabre and Kurohashi, 2017)	Mix	59	5	13	Training	Mix Collection	2017	✗
GCP Corpus (Imamura and Sumita, 2018)	Speech	10	4	10	Training	Machine Translation	2018	✗
FLORES-101 (Goyal et al., 2022)	Wiki	101	10	101	Evaluation	Human Translator	2022	✓
FLORES-200 (Costa-Jussà et al., 2022)	Wiki	204	10	204	Evaluation	Human Translator	2022	✓
BPCC (Gala et al., 2023)	Many	22	13	22	Training	Mix Collection	2023	✓
XDailyDialog (Liu et al., 2023b)	DailyDialog	4	/	4	Training	Human Translator	2023	✓
MWccMatrix (Thompson et al., 2024)	Common Crawl	90	/	/	Training	Crawl	2024	✓
TED2018 (Qi et al., 2018)	TED	58	/	58	Training	Human Translator	2018	✓
TED2020 (Reimers and Gurevych, 2020)	TED	108	/	/	Training	Human Translator	2020	✓
Ours (TED2025)	TED	113	352	50	Training	Human Translator	2025	✓

Table 1: Comparison of existing multi-way parallel corpora and our constructed TED2025, highlighting key attributes such as the data source, the number of languages (#Langs), the number of domains (#Domains), the maximum parallelism supported, the type of data (training or evaluation), the collection method, and whether the corpus is open-source.

improvements in multilingual LLM performance?

RQ3: Which instruction-tuning objectives can most effectively leverage the advantages of multi-way parallel data?

We perform a comprehensive evaluation across six multilingual benchmarks to assess the benefits of using multi-way parallel data for scaling multilingual LLMs. Our results reveal that, at an equal data scale, fine-tuning on multi-way parallel data consistently outperforms training on unaligned multilingual text for both low-resource and high-resource languages (Section 4). Additionally, we identify the most effective configurations of parallelism (Section 5). Furthermore, we investigate how different instruction-tuning objectives impact LLM performance and cross-domain robustness (Section 6).

In summary, our contributions are as follows: **(i)** We construct TED2025, a 50-way parallel corpus derived from recent TED talk transcripts, covering 113 languages and 352 domains. **(ii)** To the best of our knowledge, this is the first work to leverage multi-way parallel data for scaling multilingual LLMs. We present a systematic comparison of multilingual LLM fine-tuning using multi-way versus unaligned data, analyzing their effects on zero-shot transfer and cross-lingual representation alignment. **(iii)** We explore instruction-tuning objectives specifically designed for multi-way parallel data and provide practical recommendations for optimizing multilingual LLM performance.

2 Related Work

Multi-Way Parallel Corpora. Datasets containing the same content across multiple languages (typically more than two) are known as *multi-way parallel corpora*. These corpora have demonstrated substantial benefits for machine translation (Freitag and Firat, 2020; Xu et al., 2022; Wu

et al., 2024) and cross-lingual representation alignment (Tran et al., 2020). Existing methods for constructing such corpora include mining comparable texts (Resnik et al., 1999), aligning independently collected monolingual corpora via translation pivots (Thompson et al., 2024), extracting multi-way subsets from large bilingual collections (Ramesh et al., 2022), and harvesting multilingual web crawls (Resnik and Smith, 2003; Qi et al., 2018). However, many of these resources are limited in terms of language and domain coverage. In contrast, we construct a multi-way parallel corpus derived from recent TED Talk transcripts, offering a broader and more diverse set of languages and domains.

Scaling Multilingual LLMs. The multilingual capabilities of LLMs have been significantly enhanced through two complementary strategies: continued pretraining on diverse multilingual corpora and multilingual instruction tuning. Continued pretraining on unaligned multilingual data has improved both in-language fluency and cross-lingual transfer (Ji et al., 2024; Groeneveld et al., 2024), while instruction tuning with human-curated multilingual prompts has boosted task performance across a wide range of languages (Lai et al., 2024; Üstün et al., 2024). More recently, Mu et al. (2024) demonstrated that incorporating multi-way parallel examples into in-context prompts leads to further gains in zero-shot transfer. Building on these insights, we systematically fine-tune multilingual LLMs on large-scale multi-way parallel data and quantify their impact compared to conventional unaligned approaches.

3 Experimental Setup

TED2025. We introduce TED2025, a new multi-way parallel corpus derived from the latest TED

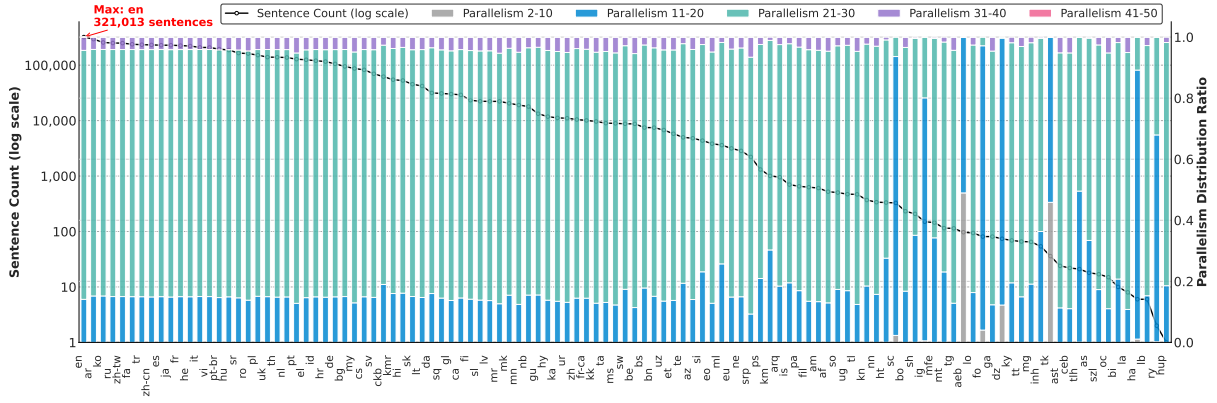


Figure 1: Distribution of sentence counts (line chart, left y-axis, log scale) and parallelism spans (bar chart, right y-axis, ratio) across languages (x-axis) in the TED2025 corpus. The parallelism spans, with a notable concentration between 21 and 30 languages, and high range even for low-resource languages.

Talk transcripts. It encompasses 113 languages with up to 50-way parallelism, making it one of the largest and most diverse resources for multilingual fine-tuning. Figure 1 illustrates the total number of sentences and the distribution of parallelism spans across languages in TED2025. Figure 2 compares translation quality, as measured by COMET-QE (Rei et al., 2020), across TED2025 (which includes 4,765 language pairs³) and other existing multi-way datasets: TED2018 (Qi et al., 2018), TED2020 (Reimers and Gurevych, 2020), and MWccMatrix (Thompson et al., 2024). Additional dataset statistics and construction details are provided in Appendix A.

We observe that: (1) the most common parallelism span in TED2025 ranges from 21 to 30 languages. Notably, many low-resource languages also achieve high degrees of parallelism, providing a solid foundation for multilingual research. (2) TED2025 contains significantly more high-quality translations (with a COMET-QE score greater than 60) compared to previous multi-way corpora. Unlike TED2020, which segments English based on punctuation, or MWccMatrix, which relies on LASER score (Artetxe and Schwenk, 2019), we use human-provided timestamps to generate cleaner and more reliable sentence alignments.

Training. To isolate the effects of multi-way parallel data, we conduct both continued pre-training (Parmar et al., 2024) and instruction tuning (Zhang et al., 2023) on TED2025. We experiment with two model families and sizes: LLaMA-3.1-8B / LLaMA-3.1-8B-Instruct and Qwen-2.5-14B / Qwen-2.5-14B-Instruct (available on Hug-

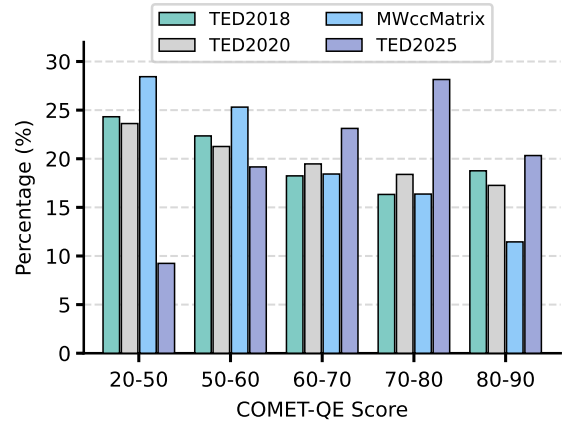


Figure 2: Comparison of translation quality between TED2025 and existing multi-way datasets, including TED2018 (Qi et al., 2018), TED2020 (Reimers and Gurevych, 2020), MWccMatrix (Thompson et al., 2024), using COMET-QE score.

ging Face⁴). To make fine-tuning feasible, we employ Low-Rank Adaptation (LoRA) (Hu et al., 2022) instead of performing full parameter updates. Full hyperparameter settings and training configurations are provided in Appendix B.

Evaluation and Metrics. We evaluate our models on five widely adopted multilingual benchmarks in a zero-shot setting, covering a range of tasks: understanding (MMMLU), reasoning (XCOPA; Ponti et al., 2020), generation (FLORES-101; Goyal et al., 2022 and FLORES-200; Costa-Jussà et al., 2022), instruction following (xIFEval; Huang et al., 2025), and text classification (SIB-200; Adelani et al., 2024). Table 2 summarizes these benchmarks along with their associated evaluation metrics. Ad-

³For language pairs with more than 10,000 sentence pairs, we randomly sample 10,000.

⁴<https://huggingface.co>

Benchmark	Task	#Langs	Metric
MMMLU	Understanding	14	Acc
XCOPA	Reasoning	11	Acc
FLORES-101	Generation	101	BLEU/COMET
FLORES-200	Generation	204	BLEU/COMET
xIFEval	Instruction Following	17	Acc
SIB-200	Text Classification	204	Acc

Table 2: Overview of evaluation benchmarks, including task types, the number of languages (#Langs) involved, and the metrics used for assessment.

ditionally, we categorize the languages in each benchmark as low-resource or high-resource based on the classification in Costa-Jussà et al. (2022).

4 Effectiveness of Multi-Way Corpora

We investigate the impact of training on multi-way parallel data on the performance of a multilingual LLM across three key dimensions: downstream performance on multilingual benchmarks (Section 4.1), zero-shot cross-lingual transfer to unseen languages (Section 4.2) and cross-lingual alignment of representations within the model’s internal embeddings (Section 4.3).

For fairness, we fix the total continued pre-training data at 5 million tokens (5M) and evaluate two LLM backbones (LLaMA-3-8B and Qwen-2.5-14B) under three conditions: (1) **Multi-Way**: Pretraining on our multi-way parallel corpus (TED2025); (2) **Unaligned**: Pretraining on an unaligned multilingual corpus⁵ (DCAD-2000; Shen et al., 2025); (3) **Baseline**: The original pretrained checkpoint without additional data.

4.1 Downstream Performance

We evaluate all variants in a zero-shot setting across four benchmarks—MMMLU, XCOPA, FLORES-101, and xIFEval—that cover understanding, reasoning, generation, and instruction following. The full results are reported in Table 3. Our findings show that, across all tasks, *Multi-Way* consistently outperforms both *Baseline* and *Unaligned* across both low- and high-resource languages. For example, on MMMLU, *Multi-Way* achieves accuracies of 22.48/41.38 in low/high-resource languages, compared to 18.27/33.72 for the *Baseline* and 19.64/36.26 for *Unaligned*. Similar improvements are observed on FLORES-101 and xIFEval. These results demonstrate that multi-way alignment provides stronger cross-lingual supervision,

⁵Unaligned data is sourced from DCAD-2000 (Shen et al., 2025) rather than TED2025, due to considerations related to scalability and domain coverage.

thereby enhancing both discriminative and generative capabilities. Although our main experiments focus on English–X translation directions, we also evaluate the trained models on non-English-centric translation pairs. The results, which show that models trained on aligned data significantly outperform those trained on unaligned data even for non-English-centric tasks, are provided in Appendix C.

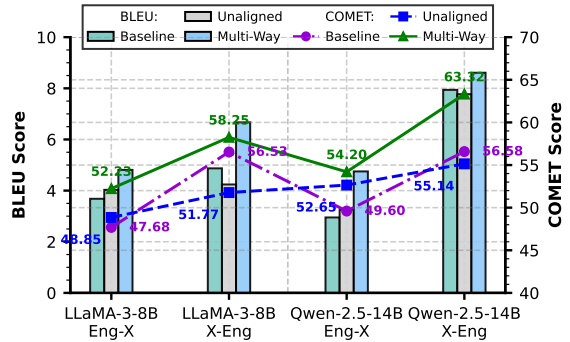


Figure 3: Cross-lingual transfer performance comparison between Baseline, Unaligned and Multi-Way pre-training on the FLORES-200 benchmark with BLEU (bar chart, left y-axis) and COMET (line chart, right y-axis) for LLaMA-3-8B and Qwen-2.5-14B models.

4.2 Zero-Shot Cross-Lingual Transfer

To assess generalization to unseen languages (Lai et al., 2022a; Zhao et al., 2025), we evaluate on FLORES-200. We exclude all languages in the evaluation subset from training and assess the English↔X translation quality. As shown in Figure 3, the *Multi-Way* model significantly outperforms both *Baseline* and *Unaligned* in both translation directions. This highlights that explicit multi-way supervision promotes language-agnostic representations, enabling robust zero-shot transfer. We further explore this hypothesis in Section 4.3, where we analyze differences in cross-lingual representation alignment across models.

4.3 Cross-Lingual Representation Alignment

We further analyze the alignment of internal representations across models. To be specific, for 32 randomly selected aligned languages (with 100 sentences each) from TED2025, we compute the following four metrics: average *cosine similarity* between parallel sentence embeddings, *Centered Kernel Alignment* (CKA) between representation matrices (Kornblith et al., 2019), Cross-lingual sentence retrieval accuracy at P@1, P@5, and P@10 (Conneau et al., 2017) and SVCCA score (Raghu et al., 2017).

	MMMLU		XCOPA		FLORES-101 (Eng-X)				FLORES-101 (X-Eng)				xIFEval	
					BLEU		COMET		BLEU		COMET			
	low	high	low	high	low	high	low	high	low	high	low	high	low	high
Baseline	18.27	33.72	23.46	34.29	6.03	11.67	57.15	61.03	13.37	22.49	75.24	82.32	17.14	24.43
Unaligned	19.64	36.26	24.62	34.76	6.12	11.78	57.51	62.11	13.84	22.74	75.82	82.58	17.28	24.44
Multi-Way	22.48	41.38	27.58	57.22	6.32	12.08	58.06	67.44	14.45	25.03	76.25	86.43	18.79	27.41

(a) LLaMA-3-8B

	MMMLU		XCOPA		FLORES-101 (Eng-X)				FLORES-101 (X-Eng)				xIFEval	
					BLEU		COMET		BLEU		COMET			
	low	high	low	high	low	high	low	high	low	high	low	high	low	high
Baseline	35.24	49.55	62.25	72.00	7.45	11.05	57.22	67.16	16.54	20.24	67.23	74.29	27.63	32.40
Unaligned	35.61	51.32	62.59	74.06	7.62	11.60	57.85	70.85	16.86	21.02	67.61	75.97	27.92	35.54
Multi-Way	36.64	55.81	63.24	79.52	8.07	13.11	58.94	80.56	17.36	23.26	68.59	81.33	28.64	40.95

(b) Qwen-2.5-14B

Table 3: Performance (%) comparison of different models across multilingual benchmarks. The **Multi-Way** approach with blue background demonstrates consistent superiority in both low-resource (left columns) and high-resource (right columns) scenarios.

		LLaMA-3-8B			Qwen-2.5-14B		
		Baseline	Unaligned	Multi-Way	Baseline	Unaligned	Multi-Way
Cosine ($\uparrow$)		0.27	0.27 _{-0.00}	0.30 _{+0.03}	0.29	0.27 _{-0.02}	0.32 _{+0.03}
CKA ($\uparrow$)		0.54	0.54 _{-0.00}	0.60 _{+0.06}	0.56	0.57 _{+0.01}	0.63 _{+0.07}
Retrieval ($\uparrow$)	P@1	0.09	0.07 _{-0.02}	0.13 _{+0.04}	0.12	0.14 _{+0.02}	0.19 _{+0.07}
	P@5	0.24	0.23 _{-0.01}	0.27 _{+0.03}	0.26	0.28 _{+0.02}	0.33 _{+0.07}
	P@10	0.35	0.33 _{-0.02}	0.39 _{+0.04}	0.36	0.38 _{+0.02}	0.42 _{+0.06}
SVCCA ($\uparrow$)		0.55	0.55 _{-0.00}	0.61 _{+0.06}	0.57	0.58 _{+0.01}	0.63 _{+0.06}

Table 4: Cross-lingual representation alignment results. Improvement margins (colored red/blue) are shown relative to Baseline. **Bolded** Multi-Way results with **red** annotations indicate consistent improvements, particularly in cross-lingual retrieval tasks (e.g., +0.07 P@1 improvement for Qwen-2.5-14B).

Table 4 shows that *Multi-Way* outperforms the other models, yielding higher CKA and better retrieval accuracy. Figure 4 further corroborates these results: *Multi-Way* demonstrates denser SVCCA alignments, particularly for linguistically distant language pairs. These metrics confirm that multi-way pretraining promotes a more coherent, language-agnostic embedding space, which drives the observed improvements in downstream performance and cross-lingual transfer.

5 Impact Factors

In Section 4, we demonstrate that using multi-way parallel data can significantly enhance the multilingual capabilities of LLMs. To better understand the factors driving this improvement, we analyze two key aspects, while keeping the total pretraining fixed at 5 million tokens: (1) **Degree of parallelism**: the number of languages aligned in each

training example (Section 5.1). (2) **English as a pivot**: the impact of including versus excluding English in multi-way groups (Section 5.2). (3) **Language combinations**: the role of language family composition in training data (Section 5.3). (4) **Training data size**: how the amount of pretraining data affects model performance (Section 5.4).

5.1 Degree of Parallelism

We construct datasets with parallelism levels ranging from 2 to 40 languages per example, sampled from the TED2025 corpus, while always keeping 5M tokens. Figure 5 shows model performance across a range of tasks for each setting. For bidirectional machine translation (FLORES-101, Eng $\rightarrow$ X and X $\rightarrow$ Eng), performance steadily improves with higher parallelism, suggesting that broader semantic alignment enhances cross-lingual generation and fluency. In contrast, for non-generative tasks (reasoning and understanding), accuracy tends to

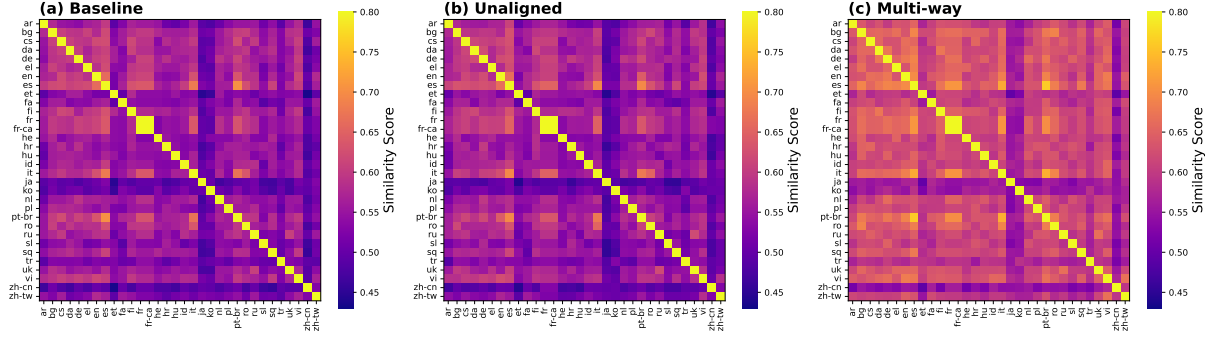


Figure 4: SVCCA alignment comparison between the Multi-Way, Unaligned and Baseline models across 32-way language pairs.

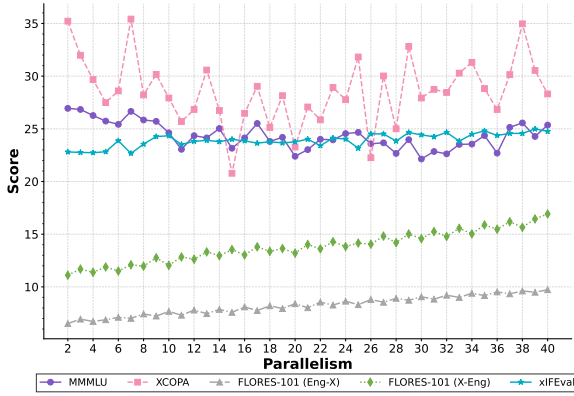


Figure 5: Performance (%) of continued pretraining models on downstream tasks with varying degrees of parallelism.

peak at small parallelism (around 6–10 languages) before deteriorating. We attribute this decline to two factors: (1) Excessive linguistic diversity can obscure shared semantic patterns. (2) With a fixed token budget, each language receives fewer tokens, limiting the ability to learn language-specific features.

5.2 English as Pivot

English is widely used as a pivot language in multilingual MT and NLP (Kim et al., 2019; Mallinson et al., 2018; Lai et al., 2023b). To explore its impact, we form five groups⁶, each with both English-included and English-excluded variants across six languages. Table 5 compares their performance across various tasks.

In tasks involving understanding and reasoning, groups that include English consistently outperform those in other language groups by an average of 2–4 percentage points. This suggests that English, as a high-resource “semantic anchor”, helps stabilize embedding alignment and facilitates trans-

⁶We provide the configurations of each group in Table 10 of Appendix D

		MMMLU	XCOPA	FLORES (Eng-X)	FLORES (X-Eng)	xIFEval
Group 1	with	23.17	30.93	5.80	13.75	23.39
	w/o	20.84	30.07	6.63	14.56	24.14
Group 2	with	22.19	35.33	6.74	13.88	22.19
	w/o	18.51	31.60	7.13	14.18	22.19
Group 3	with	26.47	36.73	6.40	13.48	22.69
	w/o	23.83	35.78	6.38	14.00	22.91
Group 4	with	23.42	33.84	6.20	12.11	23.66
	w/o	20.26	30.84	6.67	14.76	23.67
Group 5	with	23.89	39.64	6.96	13.07	23.19
	w/o	20.39	34.15	7.83	14.79	23.85

Table 5: Performance (%) comparison of models with and without (w/o) English across five different language groupings.

fer learning, particularly on complex tasks. Interestingly, for machine translation (FLORES-101) and some instruction-following tasks (xIFEval), English-inclusive groups show slightly lower performance. We attribute this to two primary factors: (1) English occupies tokens that could otherwise be used to align non-English language pairs directly. (2) The model may overly rely on English as an intermediary, which reduces its ability to directly transfer knowledge between non-English languages. These findings highlight that English’s role as a pivot language is task-dependent. While it can enhance semantic coherence in understanding and reasoning tasks, it may hinder direct multilingual transfer in generative tasks. Consequently, the inclusion of English in multilingual pretraining should be carefully considered based on the target application.

5.3 Language Combinations

In Section 5.2, we investigate the influence of including English in sampling combinations on model performance. In this section, we extend that analysis by investigating an alternative sampling strategy: whether the selected languages belong to

	MMMLU		XCOPA		FLORES-101 (Eng-X)				FLORES-101 (X-Eng)				xIFEval	
					BLEU		COMET		BLEU		COMET			
	low	high	low	high	low	high	low	high	low	high	low	high		
Baseline	41.96	46.68	64.59	66.79	5.51	10.50	56.95	60.02	14.98	18.41	68.29	73.51	35.99	38.56
MT	45.28	51.06	68.17	69.38	11.26	13.25	63.03	65.61	22.25	21.60	70.76	75.41	41.72	44.52
CLTS	39.99	45.44	62.87	66.47	3.63	9.90	56.48	59.28	13.25	16.60	66.89	73.15	34.91	38.38
MTC	40.68	44.49	64.19	65.16	5.02	9.08	56.04	57.84	14.38	18.16	67.86	73.12	34.03	38.01
CLP	42.49	47.01	65.41	68.42	6.38	11.46	57.23	61.28	15.76	19.27	69.87	73.91	36.17	40.07
MT + CLTS	42.23	47.94	65.00	68.77	6.19	10.51	58.53	60.33	16.83	18.98	68.70	74.91	36.47	40.30
MT + MTC	42.69	47.54	65.12	66.83	6.35	10.73	57.53	60.24	15.13	18.45	69.28	74.00	36.10	39.03
MT + CLP	43.20	49.07	67.29	68.47	7.31	10.51	58.40	62.64	17.75	20.77	69.18	74.47	38.49	39.67
CLTS + MTC	41.78	45.17	63.21	65.62	5.38	9.77	55.19	58.21	14.67	16.59	67.45	72.35	34.57	36.63
CLTS + CLP	42.82	46.74	65.30	67.12	5.58	10.88	57.84	60.41	15.41	19.33	68.84	73.58	36.99	38.83
MTC + CLP	42.62	46.70	65.16	67.56	6.48	10.84	57.74	60.98	15.29	19.22	68.76	73.77	36.87	38.96
MT + CLTS + MTC	41.03	46.59	64.15	66.05	5.14	10.35	56.14	59.71	14.64	18.06	67.64	73.39	35.24	38.14
MT + CLTS + CLP	42.72	47.67	64.83	67.57	6.17	10.94	57.78	60.11	15.96	19.16	68.67	74.03	36.54	38.77
MT + MTC + CLP	42.33	46.72	65.26	67.58	5.92	10.98	57.93	60.76	15.38	18.81	68.98	73.55	36.22	39.15
CLTS + MTC + CLP	41.56	46.65	64.16	66.45	4.67	10.38	56.20	59.63	14.95	17.62	68.00	73.03	35.27	38.21
MT + CLTS + MTC + CLP	43.07	47.67	66.64	67.38	7.59	12.42	57.75	60.79	17.62	19.62	71.13	75.24	36.95	40.52

(a) LLaMA-3-8B-Instruct

	MMMLU		XCOPA		FLORES-101 (Eng-X)				FLORES-101 (X-Eng)				xIFEval	
					BLEU		COMET		BLEU		COMET			
	low	high	low	high	low	high	low	high	low	high	low	high		
Baseline	63.61	68.14	78.76	82.22	7.92	12.79	61.49	66.64	16.76	20.48	68.16	70.32	47.68	52.40
MT	68.52	72.83	82.95	86.95	11.90	16.60	64.96	70.90	20.43	26.22	72.25	73.51	53.24	55.83
CLTS	61.28	67.61	76.77	79.74	7.20	10.70	60.90	64.50	15.75	18.21	67.01	69.83	45.73	51.83
MTC	62.59	65.63	76.87	81.46	6.07	12.29	61.48	65.61	13.84	19.71	67.12	68.65	47.53	50.53
CLP	64.28	70.06	79.69	83.02	9.54	13.10	62.18	68.35	17.54	21.68	69.24	70.68	48.61	52.66
MT + CLTS	63.93	69.53	78.94	82.75	8.73	13.91	61.87	67.09	18.20	21.50	68.45	71.90	48.16	54.14
MT + MTC	65.33	69.12	80.27	83.34	8.64	13.94	61.71	67.42	17.87	22.27	68.60	72.01	49.46	53.53
MT + CLP	64.95	70.51	80.68	83.52	10.30	13.65	62.37	69.32	18.43	22.19	70.01	70.80	48.85	53.15
CLTS + MTC	62.67	67.19	78.44	81.33	7.61	12.60	61.04	65.90	16.25	19.93	67.47	70.00	46.85	51.63
CLTS + CLP	64.38	68.65	78.89	82.40	8.65	13.57	61.75	67.57	17.73	20.81	68.65	71.30	47.82	53.04
MTC + CLP	64.04	68.92	79.16	82.56	8.81	13.05	62.26	67.34	16.97	21.31	68.98	71.20	48.06	53.32
MT + CLTS + MTC	64.13	68.55	79.51	82.63	8.26	13.30	62.24	67.01	17.69	21.29	68.38	70.74	48.54	52.80
MT + CLTS + CLP	64.50	68.82	80.43	83.22	9.05	13.93	62.38	67.56	18.66	22.01	68.76	71.62	48.96	53.74
MT + MTC + CLP	64.71	69.52	80.26	83.32	9.15	13.37	62.82	67.66	17.75	21.70	69.33	71.50	48.93	53.41
CLTS + MTC + CLP	63.81	69.00	79.70	82.52	8.69	13.48	62.25	67.04	17.55	20.75	68.23	70.96	47.68	52.55
MT + CLTS + MTC + CLP	64.94	68.19	80.25	82.42	9.87	13.56	62.86	67.04	16.94	21.69	68.17	71.34	49.03	53.72

(b) Qwen-2.5-14B-Instruct

Table 6: Downstream task performance (%) of LLaMA-3-8B-Instruct and Qwen-2.5-14B-Instruct models trained with different instruction objectives, including MT, CLTS, MTC and CLP.

the same language family. To this end, we construct four language groups. Groups 1 and 2 consist of languages from the same language family, whereas Groups 3 and 4 include languages from diverse families. The specific language configurations for each group are detailed in Appendix D (Table 10).

Figure 6 illustrates the impact of language family composition on model performance. The results indicate that sampling from cross-family language combinations more effectively leverages the benefits of multi-way parallel corpora, resulting in greater improvements in multilingual task performance. In contrast, sampling languages from the same family yields only marginal gains. This may

be attributed to the structural and lexical similarities among related languages, which can introduce redundancy during training and limit the model’s ability to generalize across typologically diverse languages.

5.4 Training Data Size

Figure 7 presents the impact of training data size on model performance. We conduct experiments by randomly sampling varying amounts of tokens—10K, 50K, 100K, 500K, 1M, 5M, 10M, 50M, 100M, 500M, and 1B—from the constructed TED2025 dataset. The results demonstrate that model performance is notably constrained when

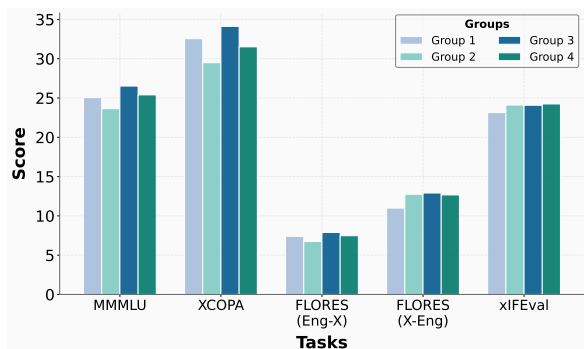


Figure 6: Impact of language family composition on model performance.

trained on smaller datasets (typically under 100K tokens). However, as the size of training data increases, performance consistently improves across all evaluated tasks.

For MMMLU and XCOPA, performance exhibits early gains with additional data but plateaus beyond a certain threshold. This trend likely reflects the nature of these tasks, which emphasize general language understanding and reasoning. Once the model acquires the necessary core linguistic and world knowledge, the marginal gains from further data diminish. Interestingly, performance on FLORES and xFEEval continues to improve with increasing data volume. These tasks, which involve cross-lingual understanding and translation—particularly for low-resource languages or semantically nuanced alignments—appear to benefit more substantially from large-scale data. This suggests that extensive training data is crucial for enhancing translation quality and evaluation accuracy in such settings.

6 Instruction Tuning

In this section, we further investigate whether instruction tuning can also enhance multilingual performance effectively. Specifically, we address the following key questions⁷: (1) Which of the different instruction fine-tuning objectives, built on multi-way parallel data, is most effective? (2) Do models trained with multi-way parallel data exhibit better generalization across domains?

6.1 Instruction Tuning Objectives

Using our constructed multi-way parallel data (TED2025), we define four instruction tasks: machine translation (MT), cross-lingual text similarity (CLTS), multilingual text classification (MTC), and

cross-language paraphrasing (CLP). Table 6 reports their impact on downstream benchmarks. Table 7 summarizes the prompt templates and output formats for each task.

We observe that the improvements in MT are the largest and most stable in both high-resource and low-resource languages. This can primarily be attributed to the fact that, as a token-level supervised generation task, translation strengthens cross-lingual syntactic and semantic consistency, making it a particularly robust task for broad multilingual generalization. In contrast, CLTS and MTC show smaller drops across different tasks, which we attribute to the coarser granularity of similarity judgments that may not provide fine-grained alignment signals. Moreover, discrete class labels lack the expressive power to capture subtle semantic distinctions. Although CLP shows similar advantages to MT in generation tasks, its advantages are narrower in scope and cannot be widely generalized.

Interestingly, the combination of tasks did not significantly improve performance, which may be due to several reasons: First, the objectives of MT and CLP differ; while MT emphasizes accuracy, CLP focuses more on the diversity of expression, which may make it difficult for the model to balance these two goals during training, thereby affecting performance. Second, interference between multiple tasks may arise, making it challenging for the model to focus on optimizing the optimal goal of each task, particularly when the task objectives are too similar, amplifying the interference effect. Furthermore, there is overlap in task design between MT and CLP, which may cause the model to encounter redundant training signals when processing both tasks, preventing it from fully utilizing the unique value of each task. Finally, joint training of multiple tasks increases the complexity of model training, potentially leading to unstable gradient updates and affecting the model’s convergence.

6.2 Cross-Domain Generalization

To evaluate cross-domain transfer (Lai et al., 2022b; Liu et al., 2023a), we extract domain-labeled subsets from TED2025 according to the taxonomy of the SIB-200 benchmark. Instruction tuning is then performed on each domain using the MTC objective, with the resulting models evaluated across all other domains within SIB-200. Figure 8 illustrates the transfer performance.

Overall, instruction tuning with multi-way parallel data significantly improves domain transfer.

⁷Additionally, we explore the cross-lingual alignment of representations in instruction tuned models (Appendix E.)

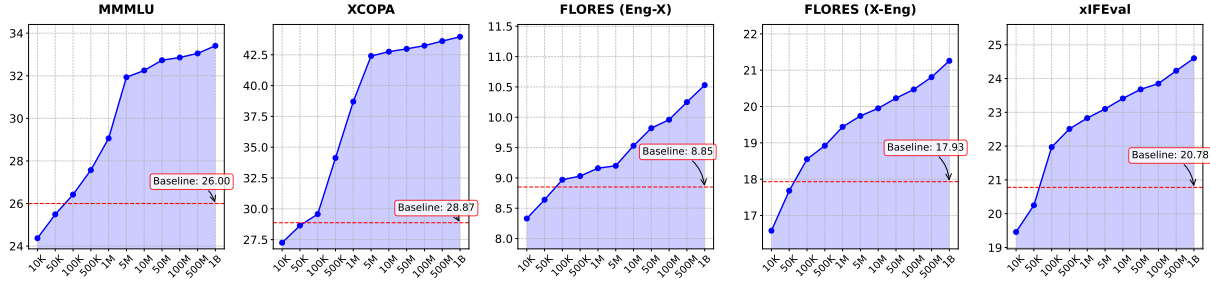


Figure 7: Impact of training data size on model performance across different token amounts (ranging from 10K to 1B) sampled from the TED2025 dataset.

Task	Prompt
Machine Translation (MT)	Translate the following {src_lang_1}, {src_lang_2}, ..., {src_lang_m} sentence to {tgt_lang_1}, {tgt_lang_2}, ..., {tgt_lang_n}. {src_lang_1} Sentence: {src_txt_1}. {src_lang_2} Sentence: {src_txt_2}. ... {src_lang_m} Sentence: {src_txt_m}. Translation: {tgt_lang_1} Sentence: {tgt_txt_1}. {tgt_lang_2} Sentence: {tgt_txt_2}. ... {tgt_lang_n} Sentence: {tgt_txt_n}.
Cross-Lingual Text Similarity (CLTS)	Given the sentences below in different languages, rate how similar their meanings are on a scale of 0 to 1, where 0 means completely dissimilar and 1 means identical meanings. {lang_1} Sentence: {txt_1}. {lang_2} Sentence: {txt_2}. ... {lang_m} Sentence: {txt_m}. Similarity: {sim_score}.
Multilingual Text Classification (MTC)	Classify the following sentence in {lang_1}, {lang_2}, ..., {lang_m} into one of the following categories: {domain_list}. {lang_1} Sentence: {txt_1}. {lang_2} Sentence: {txt_2}. ... {lang_m} Sentence: {txt_m}. Categories: {target_domain}.
Cross-Lingual Paraphrasing (CLP)	Paraphrase the following {src_lang} sentence in {tgt_lang}. {src_lang} Sentence: {src_txt}. Paraphrasing: {tgt_lang} Sentence: {tgt_txt}.

Table 7: Instruction prompts used for four multilingual tasks: machine translation (MT), cross-lingual text similarity (CLTS), multilingual text classification (MTC), and cross-lingual paraphrasing (CLP). Each prompt is designed to reflect the task’s specific objective and structure.

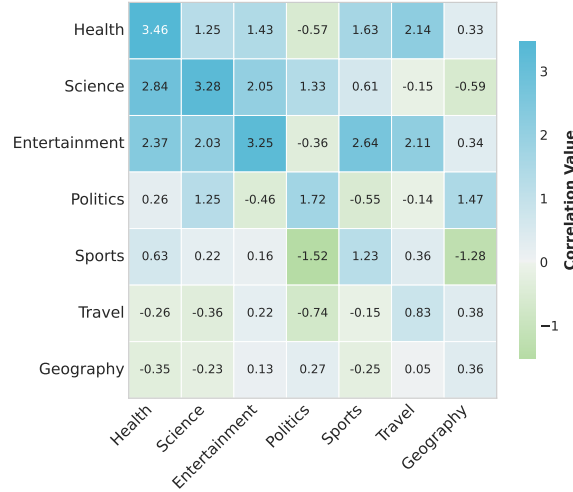


Figure 8: Cross-domain generalization performance of instruction-tuned models using multi-way parallel data. Models trained on one domain are evaluated on all other domains from the SIB-200 benchmark.

The rich cross-lingual and cross-domain signals allow the model to learn domain-invariant features, enhancing robustness when confronted with novel topics and linguistic contexts. However, transfer performance remains limited in domains such as politics, sports, travel, and geography. We hypothesize that the high topical diversity, coupled with the

relative sparsity of domain-specific examples in the training data, hinders the model’s ability to capture specialized patterns. Overcoming these limitations may require more balanced domain coverage or targeted data augmentation strategies.

7 Conclusion

In this paper, we construct a large-scale, high-quality multi-way parallel dataset covering 113 languages, with a maximum parallel degree of 50. This dataset provides a strong foundation for investigating the multilingual adaptation of LLMs. Using this dataset, we systematically explore best practices for adapting LLMs to multilingual tasks via multi-way parallel data. Our experiments reveal that multi-way data offers substantial advantages for both continued pretraining and instruction tuning, resulting in improved cross-lingual and cross-domain generalization. We further analyze key factors influencing model performance, including the degree of parallelism, the language combination strategies, and instruction training objectives.

Limitations

This work has the following limitations: (i) Due to limited computational resources, we employed parameter-efficient fine-tuning (PEFT) methods (LoRA) instead of full-parameter fine-tuning. While recent studies have demonstrated that LoRA achieves performance comparable to full fine-tuning across various tasks, our conclusions may still benefit from validation under full fine-tuning or alternative PEFT methods such as adapters or prefix tuning. (ii) Although our constructed dataset surpasses existing multi-way parallel corpora in both language coverage and maximum parallel degree, its overall size remains modest compared to large-scale unaligned multilingual datasets. To fully unlock the potential of multi-way parallel data for LLM adaptation, future work will focus on scaling up the dataset to further enhance multilingual performance.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grants No. 62236011 and No. T2341003), and by a grant from the Guoqiang Institute, Tsinghua University. Additionally, support was received from the European Research Council (ERC) under the European Union’s Horizon Europe Research and Innovation Programme (Grant Agreement No. 101113091), as well as from the German Research Foundation (DFG; Grant FR 2829/7-1).

References

- David Ifeoluwa Adelani, Hannah Liu, Xiaoyu Shen, Nikita Vassilyev, Jesujoba O. Alabi, Yanke Mao, Haonan Gao, and En-Shiun Annie Lee. 2024. [SIB-200: A simple, inclusive, and big evaluation dataset for topic classification in 200+ languages and dialects](#). In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 226–245, St. Julian’s, Malta. 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.
- Christos Christodouloupoulos and Mark Steedman. 2015. A massively parallel corpus: the bible in 100 languages. *Language resources and evaluation*, 49:375–395.
- Alexis Conneau, Guillaume Lample, Marc’Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*.
- Marta R Costa-Jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, and 1 others. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.
- Raj Dabre and Sadao Kurohashi. 2017. Mmc4nlp: multilingual multiway corpora repository for natural language processing. *arXiv preprint arXiv:1710.01025*.
- Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. 2024. [A survey on in-context learning](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1107–1128, Miami, Florida, USA. 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.
- Jay P Gala, Pranjal A Chitale, AK Raghavan, Varun Gumma, Sumanth Doddapaneni, Kumar M Aswanth, Janki Atul Nawale, Anupama Sujatha, Ratish Pudupully, Vivek Raghavan, and 1 others. 2023. Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages. *Transactions on Machine Learning Research*, 2023.
- Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc’Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. [The Flores-101 evaluation benchmark for low-resource and multilingual machine translation](#). *Transactions of the Association for Computational Linguistics*, 10:522–538.
- Dirk Groeneveld, Iz Beltagy, Evan Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, Shane Arora, David Atkinson, Russell Authur, Khyathi Chandu, Arman Cohan, Jennifer Dumas, Yanai Elazar, Yuling Gu, Jack Hessel, and 24 others. 2024. [OLMo: Accelerating the science of language models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15789–15809, Bangkok, Thailand. Association for Computational Linguistics.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. [LoRA: Low-rank adaptation of large language models](#). In *International Conference on Learning Representations*.
- Haoyang Huang, Tianyi Tang, Dongdong Zhang, Xin Zhao, Ting Song, Yan Xia, and Furu Wei. 2023. [Not all languages are created equal in LLMs: Improving multilingual capability by cross-lingual-thought prompting](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12365–12394, Singapore. Association for Computational Linguistics.
- Kaiyu Huang, Fengran Mo, Xinyu Zhang, Hongliang Li, You Li, Yuanchi Zhang, Weijian Yi, Yulong Mao, Jinchun Liu, Yuzhuang Xu, and 1 others. 2024. A survey on large language models with multilingualism: Recent advances and new frontiers. *arXiv preprint arXiv:2405.10936*.
- Xu Huang, Wenhao Zhu, Hanxu Hu, Conghui He, Lei Li, Shujian Huang, and Fei Yuan. 2025. Benchmark: A comprehensive multilingual evaluation suite for large language models. *arXiv preprint arXiv:2502.07346*.
- Kenji Imamura and Eiichiro Sumita. 2018. [Multilingual parallel corpus for global communication plan](#). In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).
- Shaoxiong Ji, Zihao Li, Indraneil Paul, Jaakko Paavola, Peiqin Lin, Pinzhen Chen, Dayyán O’Brien, Hengyu Luo, Hinrich Schütze, Jörg Tiedemann, and 1 others. 2024. Emma-500: Enhancing massively multilingual adaptation of large language models. *arXiv preprint arXiv:2409.17892*.
- Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM computing surveys*, 55(12):1–38.
- Yunsu Kim, Petre Petrov, Pavel Petrushkov, Shahram Khadivi, and Hermann Ney. 2019. [Pivot-based transfer learning for neural machine translation between non-English languages](#). 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 866–876, Hong Kong, China. Association for Computational Linguistics.
- Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. 2019. Similarity of neural network representations revisited. In *International conference on machine learning*, pages 3519–3529. PMLR.
- Viet Dac Lai, Nghia Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Huu Nguyen. 2023a. [ChatGPT beyond English: Towards a comprehensive evaluation of large language models in multilingual learning](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13171–13189, Singapore. Association for Computational Linguistics.
- Wen Lai, Alexandra Chronopoulou, and Alexander Fraser. 2022a. [m⁴ adapter: Multilingual multi-domain adaptation for machine translation with a meta-adapter](#). In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4282–4296, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Wen Lai, Alexandra Chronopoulou, and Alexander Fraser. 2023b. [Mitigating data imbalance and representation degeneration in multilingual machine translation](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14279–14294, Singapore. Association for Computational Linguistics.
- Wen Lai, Jindřich Libovický, and Alexander Fraser. 2022b. [Improving both domain robustness and domain adaptability in machine translation](#). In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 5191–5204, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.
- Wen Lai, Mohsen Mesgar, and Alexander Fraser. 2024. [LLMs beyond English: Scaling the multilingual capability of LLMs with cross-lingual feedback](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 8186–8213, Bangkok, Thailand. Association for Computational Linguistics.
- Fangyu Liu, Qianchu Liu, Shruthi Bannur, Fernando Pérez-García, Naoto Usuyama, Sheng Zhang, Tristan Naumann, Aditya Nori, Hoifung Poon, Javier Alvarez-Valle, Ozan Oktay, and Stephanie L. Hyland. 2023a. [Compositional zero-shot domain transfer with text-to-text models](#). *Transactions of the Association for Computational Linguistics*, 11:1097–1113.
- Zeming Liu, Ping Nie, Jie Cai, Haifeng Wang, Zhengyu Niu, Peng Zhang, Mrinmaya Sachan, and Kaiping Peng. 2023b. [XDailyDialog: A multilingual parallel dialogue corpus](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12240–12253, Toronto, Canada. Association for Computational Linguistics.
- Jonathan Mallinson, Rico Sennrich, and Mirella Lapata. 2018. [Sentence compression for arbitrary languages via multilingual pivoting](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2453–2464, Brussels, Belgium. Association for Computational Linguistics.
- Yongyu Mu, Peinan Feng, Zhiquan Cao, Yuzhang Wu, Bei Li, Chenglong Wang, Tong Xiao, Kai Song, Tongran Liu, Chunliang Zhang, and JingBo Zhu. 2024.

- Revealing the parallel multilingual learning within large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6976–6997, Miami, Florida, USA. Association for Computational Linguistics.
- Jupinder Parmar, Sanjev Satheesh, Mostofa Patwary, Mohammad Shoeybi, and Bryan Catanzaro. 2024. Reuse, don’t retrain: A recipe for continued pre-training of language models. *arXiv preprint arXiv:2407.07263*.
- Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. [XCOPA: A multilingual dataset for causal common-sense reasoning](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2362–2376, Online. Association for Computational Linguistics.
- Ye Qi, Devendra Sachan, Matthieu Felix, Sarguna Padmanabhan, and Graham Neubig. 2018. [When and why are pre-trained word embeddings useful for neural machine translation?](#) In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 529–535, New Orleans, Louisiana. Association for Computational Linguistics.
- Libo Qin, Qiguang Chen, Yuhang Zhou, Zhi Chen, Yinghui Li, Lizi Liao, Min Li, Wanxiang Che, and Philip S Yu. 2024. Multilingual large language model: A survey of resources, taxonomy and frontiers. *arXiv preprint arXiv:2404.04925*.
- Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. 2017. Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. *Advances in neural information processing systems*, 30.
- Gowtham Ramesh, Sumanth Doddapaneni, Aravindh Bheemaraj, Mayank Jobanputra, Raghavan AK, Ajitesh Sharma, Sujit Sahoo, Harshita Diddee, Mahalakshmi J, Divyanshu Kakwani, Navneet Kumar, Aswin Pradeep, Srihari Nagaraj, Kumar Deepak, Vivek Raghavan, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh Shantadevi Khapra. 2022. [Samanantar: The largest publicly available parallel corpora collection for 11 Indic languages](#). *Transactions of the Association for Computational Linguistics*, 10:145–162.
- Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. [COMET: A neural framework for MT evaluation](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.
- Nils Reimers and Iryna Gurevych. 2020. [Making monolingual sentence embeddings multilingual using knowledge distillation](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525, Online. Association for Computational Linguistics.
- Philip Resnik, Mari Broman Olsen, and Mona Diab. 1999. The bible as a parallel corpus: Annotating the ‘book of 2000 tongues’. *Computers and the Humanities*, 33:129–153.
- Philip Resnik and Noah A. Smith. 2003. [The web as a parallel corpus](#). *American Journal of Computational Linguistics*, 29(3):349–380.
- Yingli Shen, Wen Lai, Shuo Wang, Xueren Zhang, Kangyang Luo, Alexander Fraser, and Maosong Sun. 2025. Dcad-2000: A multilingual dataset across 2000+ languages with data cleaning as anomaly detection. *arXiv preprint arXiv:2502.11546*.
- Brian Thompson, Mehak Dhaliwal, Peter Frisch, Tobias Domhan, and Marcello Federico. 2024. [A shocking amount of the web is machine translated: Insights from multi-way parallelism](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1763–1775, Bangkok, Thailand. Association for Computational Linguistics.
- Chau Tran, Yuqing Tang, Xian Li, and Jiatao Gu. 2020. Cross-lingual retrieval for iterative self-supervised training. *Advances in Neural Information Processing Systems*, 33:2207–2219.
- Ahmet Üstün, Viraat Aryabumi, Zheng Yong, Wei-Yin Ko, Daniel D’souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. 2024. [Aya model: An instruction fine-tuned open-access multilingual language model](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15894–15939, Bangkok, Thailand. Association for Computational Linguistics.
- Di Wu, Shaomu Tan, Yan Meng, David Stap, and Christof Monz. 2024. [How far can 100 samples go? unlocking zero-shot translation with tiny multi-parallel data](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15092–15108, Bangkok, Thailand. Association for Computational Linguistics.
- Yulin Xu, Zhen Yang, Fandong Meng, and Jie Zhou. 2022. [EAG: Extract and generate multi-way aligned corpus for complete multi-lingual neural machine translation](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8141–8153, Dublin, Ireland. Association for Computational Linguistics.
- Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, and 1 others. 2023. Instruction tuning for large language models: A survey. *arXiv preprint arXiv:2308.10792*.

- Yiran Zhao, Wenxuan Zhang, Huiming Wang, Kenji Kawaguchi, and Lidong Bing. 2025. [AdaMergeX: Cross-lingual transfer with large language models via adaptive adapter merging](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 9785–9800, Albuquerque, New Mexico. Association for Computational Linguistics.
- Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, and Zheyang Luo. 2024. [LlamaFactory: Unified efficient fine-tuning of 100+ language models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 400–410, Bangkok, Thailand. Association for Computational Linguistics.
- Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. 2023. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*.
- Michał Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. 2016. [The United Nations parallel corpus v1.0](#). In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)*, pages 3530–3534, Portorož, Slovenia. European Language Resources Association (ELRA).

A Statistics of the Constructed TED2025

In Section 3, we introduce TED2025 dataset and present the distribution of sentence counts across languages, along with the overall degree of parallelism (Figure 1) and translation quality compared with other multi-way corpora (Figure 2). To provide a more comprehensive overview, we further analyze domain coverage, fine-grained variations in parallelism, and the distribution of bitexts with respect to both quantity and quality.

Domain Coverage. The TED2025 dataset is a multi-domain, multi-way parallel corpus encompassing 352 domains, with domain labels derived from TED Talks. Table 11 presents statistics on the number of talks per domain. We observe that the domains of global issues, education, technology, art, and business constitute the top five in terms of talk count. This statistical overview offers valuable insights into the dataset’s structure and facilitates a deeper understanding of its composition. Moreover, the domain labels enhance the dataset’s suitability for cross-lingual text classification tasks, positioning TED2025 as a robust benchmark for multilingual text classification.

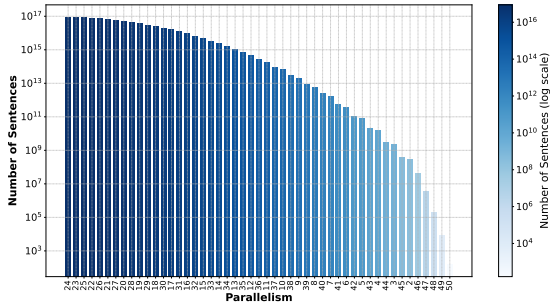


Figure 9: Fine-grained parallelism in TED2025 dataset by tuple size.

Fine-grained Parallelism. In Figure 1, we present an approximate count (ratio) of parallelism across different languages. For a more intuitive understanding of fine-grained parallelism in the TED2025 dataset, Figure 9 shows the tuple size corresponding to parallelism. Note that we count tuples rather than individual sentences. For instance, in a combination of six languages—English, French, Spanish, Russian, Arabic, and Chinese—a large tuple (en, fr, es, ru, ar, zh) encompasses all possible language combinations. The total number of such combinations is given by $C_6^1 + C_6^2 + C_6^3 + C_6^4 + C_6^5 + C_6^6 = 6 + 15 + 20 + 15 + 6 + 1 = 63$

In contrast, the corresponding sentence count for this tuple is $6 \times 1 + 15 \times 2 + 20 \times 3 + 15 \times 4 + 6 \times 5 + 1 \times 6 = 192$ sentences. This tuple-based counting method offers a more precise analysis of parallelism in the dataset.

Quantity and Quality of Bitext. In Figure 2, we compare the translation quality of TED2025 with existing multi-way parallel datasets, demonstrating the overall effectiveness of TED2025 translations. To offer a more comprehensive and intuitive view of translation quality, Table 12, 13, 14, 15, 16, and 17 report COMET-QE score for all 4,765 language pairs included in the dataset. This analysis highlights that, despite being a multi-way parallel corpus, TED2025 can be readily decomposed into bilingual sentence pairs, making it suitable for training machine translation models or fine-tuning LLMs on specific language pairs. By providing both the number of bilingual pairs and their associated translation quality, we aim to support researchers in selecting suitable data for their specific translation tasks and in optimizing performance on targeted language pairs.

B Details of Experimental Setup.

LoRA		Training	
rank	8	batch size	8
alpha	32	learning rate	1e-04
dropout	0.1	lr schedule	cosine
target	all	warmup ratio	0.1

Table 8: Hyper-parameters for continued pretraining and instruction tuning using LLaMA-Factory.

Training and Inference Setup. Due to computational resource constraints, we adopt LoRA (Hu et al., 2022) for continued pretraining and instruction fine-tuning of LLMs. All experiments are conducted using the LLaMA-Factory platform⁸ (Zheng et al., 2024), with training hyper-parameters detailed in Table 8. Each experiment is run on 8 NVIDIA A100 (80GB) GPUs. For inference, we utilize the vLLM toolkit⁹, and the prompt templates used for each benchmark are partially sourced from PromptSource¹⁰ as well as the respective original papers.

⁸<https://github.com/hiyouga/LLaMA-Factory>

⁹<https://github.com/vllm-project/vllm>

¹⁰<https://github.com/bigscience-workshop/promptsources>

Evaluation Benchmarks. We evaluate the trained model across a diverse set of tasks to comprehensively assess its capabilities in natural language understanding, commonsense reasoning, text generation, instruction following, and text classification. These tasks span multiple languages and domains, enabling us to measure both general and multilingual performance. Below, we outline the benchmarks used for each evaluation category:

- **Natural Language Understanding:** We use the MMMLU benchmark, a multilingual extension of the widely adopted MMLU dataset (Hendrycks et al., 2020), designed for evaluating the multitask language understanding abilities of large language models. MMMLU covers 14 languages and includes questions from a wide range of domains. We report accuracy across all tasks to measure performance.
- **Commonsense Reasoning:** We evaluate the model using the XCOPA dataset (Ponti et al., 2020). XCOPA tests a model’s ability to perform causal commonsense reasoning in multiple languages. The task involves selecting the most plausible cause or effect of a given premise from two alternatives, thus requiring both language understanding and reasoning skills.
- **Text Generation:** We assess multilingual text generation performance using two benchmarks. First, FLORES-101 (Goyal et al., 2022) is used to evaluate the model’s general generation quality across a broad set of high- and low-resource languages. Second, FLORES-200 (Costa-Jussà et al., 2022) is employed to test zero-shot cross-lingual transfer capabilities—specifically, the model’s ability to generate high-quality outputs in languages it was not directly trained on.
- **Instruction Following:** We use the multilingual variant of the IFEval benchmark (Zhou et al., 2023), implemented by the Benchmax framework (Huang et al., 2025), to evaluate the model’s ability to follow human instructions across diverse languages and task types. This benchmark focuses on the alignment between user instructions and model responses, which is critical for real-world applications of instruction-tuned models.

- **Text Classification:** For evaluating domain-robust classification performance, we adopt the SIB-200 benchmark (Adelani et al., 2024), which contains text classification tasks across 200 languages and multiple domains. This benchmark is particularly suited for testing the generalization and robustness of instruction-tuned models in a multilingual setting.

C Non-English-Centric Translation Performance

In addition to the primary experiments on English–X translation directions presented in Section 4.1, we also investigate the performance of our models on non-English-centric translation pairs. This analysis is motivated by the potential of multilingual models to perform well in settings where neither language involved is English. To this end, we randomly select five language pairs (sk-ta, kk-ko, eo-sw, ja-zh, and nl-tr) that do not involve English as either the source or target language. We then evaluate the translation performance of the pretrained models under three different data configurations: Baseline, Unaligned, and Multi-way, as described in Section 4.1. The results, including BLEU and COMET scores, are presented in Table 9.

We observe that: (1) Models trained on aligned data consistently outperform those trained on unaligned data across all translation pairs, indicating that aligned corpora are essential for achieving high translation quality, regardless of whether English is involved. (2) Models trained on unaligned data show a significant improvement over the baseline, suggesting that even in the absence of high-quality aligned corpora, leveraging unaligned multilingual data can still enhance translation performance. (3) The multi-way approach consistently yields superior results compared to both the baseline and unaligned models. This highlights the advantages of multi-way training for multilingual translation tasks, as it not only improves performance on English-centric tasks but also significantly boosts translation quality for non-English-centric language pairs.

D Configuration of Different Language Combinations

As discussed in Section 5.3, we define four distinct language groups for our analysis. The specific

	sk-ta		kk-ko		eo-sw		ja-zh		nl-tr	
	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET
Baseline	3.14	51.24	3.11	50.53	2.93	48.56	5.02	54.71	3.63	52.35
Unaligned	4.16	52.75	3.82	51.97	3.46	49.73	6.39	55.84	4.75	52.27
Multi-Way	6.31	53.25	5.38	52.66	5.14	51.35	8.06	56.27	6.82	53.50

Table 9: Translation performance of models trained on aligned, unaligned, and multi-way data setups for five non-English-centric language pairs.

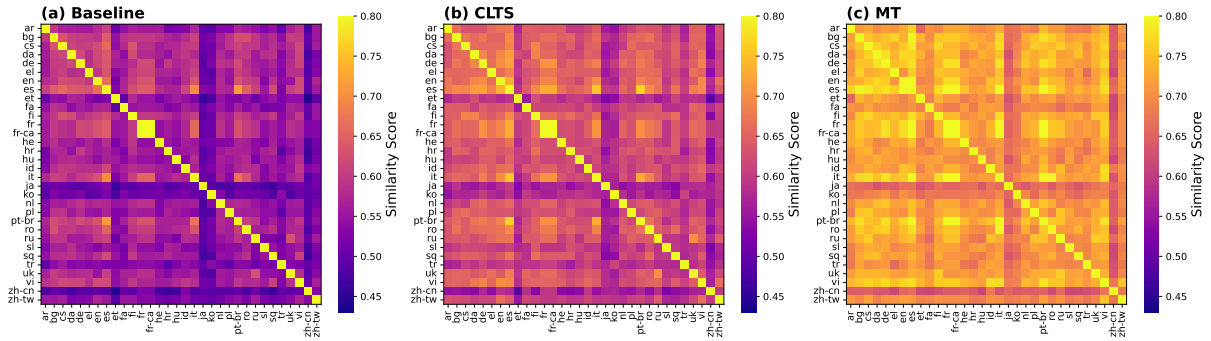


Figure 10: Representation alignment with instruction tuning across different training objectives.

English Pivoting Combinations		
Group	Language List	
group 1	en,vi,ar,bg,de,es,fr,he,it,ja,ko	
group 2	en,nl,pl,pt-br,zh-cn,ar,bg,de,es,fr,he	
group 3	en,el,vi,ar,bg,de,es,fr,he,it,ja	
group 4	en,el,ar,bg,de,es,fr,he,it,ja,ko	
group 5	en,fa,hu,ar,bg,de,es,fr,he,it,ja	
Language Family Combinations		
Group	Language List	Language Family
group 1	bg,pl,ru,sr,uk	Slavic
group 2	el,it,kmr,sq,tr	Indo-European
group 3	en,ko,my,sq,zh-tw	Indo-European (en, sq), Sino-Tibetan (zh-tw), Koreanic (ko)
group 4	fr,hu,hy,lv,vi	Indo-European (fr, hy), Uralic (hu, lv), Austroasiatic (vi)

Table 10: Language configurations for different sampling groups.

configurations of these groups are presented in Table 10.

E Representation Alignment with Instruction Tuning

In this section, we explore how each tuning objective reshapes the model’s internal multilingual embeddings, using the same alignment metrics as in Section 4.3. Figure 10 shows that MT-tuned models achieve the highest SVCCA score, indicat-

ing tighter alignment among semantically equivalent sentences across languages. In contrast, CLTS yields minimal alignment gains despite its similarity focus, likely because binary similarity labels lack the contextual richness of translation pairs.

Topic	#Talks	Topic	#Talks	Topic	#Talks	Topic	#Talks	Topic	#Talks
global issues	10334	race	154	coronavirus	70	inclusion	36	resources	18
education	9126	writing	154	sex	69	solar system	35	paleontology	17
technology	6323	physics	153	pandemic	69	sight	35	dinosaurs	17
art	5858	gender	144	product design	68	Audacious Project	35	television	17
business	5694	exploration	143	algorithm	68	industrial design	34	exercise	17
Life	5623	neuroscience	142	aging	67	sound	34	blockchain	17
health	5561	family	140	empathy	67	behavioral economics	34	fungi	16
science	4819	architecture	138	justice system	67	online privacy	34	public speaking	16
entertainment	4717	emotions	137	goals	66	travel	33	nuclear energy	16
design	2313	humor	136	urban planning	65	magic	33	PTSD	16
Humanities	1048	happiness	132	crime	65	trees	33	driverless cars	16
TED-Ed	941	universe	131	machine learning	65	weather	33	bees	15
animation	879	AI	131	compassion	64	public space	33	geology	15
culture	847	illness	131	investing	62	TED Prize	32	smell	15
social change	783	entrepreneur	130	fish	62	bioethics	31	Alzheimer's	15
TEDx	734	decision-making	130	demo	62	military	30	shopping	15
society	712	language	130	disability	62	illusion	30	Autism spectrum disorder	15
history	617	self	130	conservation	62	human rights	29	vulnerability	15
innovation	518	photography	127	feminism	61	theater	28	TED Connects	15
humanity	500	policy	126	Planets	60	solar energy	28	toys	14
biology	482	religion	124	renewable energy	57	biosphere	28	Antarctica	14
communication	453	media	121	code	57	maps	27	science fiction	14
future	434	democracy	120	women in business	57	spoken word	27	Islam	14
creativity	413	india	120	chemistry	57	heart	27	neurology	14
climate change	412	energy	118	Middle East	57	Slavery	27	Moon	13
community	407	motivation	115	immigration	57	sexual violence	27	3D printing	13
personal growth	395	philosophy	114	Best of the Web	56	surveillance	27	body language	13
environment	392	film	113	natural disaster	55	manufacturing	27	hearing	13
activism	371	violence	113	dance	55	trust	27	meditation	13
sustainability	341	literature	112	consciousness	54	archaeology	26	Brazil	12
performance	325	parenting	111	marine biology	54	AIDS	25	graphic design	12
psychology	325	journalism	111	virus	53	virtual reality	25	ebola	12
medicine	323	money	111	statistics	52	gardening	25	suicide	12
brain	323	potential	110	depression	52	nanotechnology	25	wind energy	12
music	316	ancient world	109	microbiology	52	Europe	25	coral reefs	12
work	316	social media	108	china	51	biomimicry	24	rivers	12
economics	308	love	107	electricity	51	drones	24	international relations	12
health care	307	poetry	107	plants	51	mindfulness	24	glaciers	12
nature	300	law	107	Vaccines	51	quantum	24	augmented reality	12
collaboration	288	pollution	106	Asia	49	aliens	24	worklife	12
politics	279	biodiversity	105	ethics	49	friendship	24	rocket science	11
animals	271	software	103	bacteria	48	encryption	24	Christianity	11
women	269	visualizations	103	farming	48	medical imaging	24	Sun	11
TED Fellows	267	teaching	100	prison	47	South America	23	bullying	11
human body	267	international development	99	refugees	47	telescopes	23	CRISPR	11
storytelling	259	finance	99	TED en Español	47	birds	23	String theory	10
invention	250	genetics	96	gaming	46	Big Bang	22	homelessness	10
identity	245	death	96	fear	46	Mission Blue	22	grammar	9
kids	243	books	94	natural resources	46	Surgery	22	typography	8
engineering	241	TED Residency	94	LGBTQIA+	46	protest	22	Buddhism	8
leadership	234	biotech	92	terrorism	44	painting	22	asteroid	8
government	233	beauty	92	philanthropy	44	interview	21	deextinction	8
equality	223	work-life balance	89	personality	44	library	21	cryptocurrency	8
public health	219	robots	88	marketing	43	plastic	21	metaverse	8
medical research	218	poverty	87	drugs	43	Mars	21	conducting	7
Internet	214	water	84	sociology	43	Egypt	21	bionics	7
computers	201	transportation	83	curiosity	43	pregnancy	21	NASA	7
Africa	197	cancer	83	TEDMED	43	synthetic biology	21	atheism	5
data	194	astronomy	82	indigenous peoples	42	Transgender	21	pain	5
cities	193	agriculture	82	sleep	41	prosthetics	20	forensics	4
disease	192	DNA	79	diversity	40	museums	20	microbes	4
space	187	success	79	fashion	40	addiction	20	NFTs	4
war	183	ecology	77	discovery	40	TED Membership	19	Judaism	3
United States	181	infrastructure	77	corruption	39	primates	18	street art	3
food	174	cognitive science	76	time	39	cyber security	18	Hinduism	1
math	173	sports	74	consumerism	38	astrobiology	18	reproductive health	1
ocean	168	youth	73	productivity	38	dark matter	18	crowdsourcing	1
Countdown	168	comedy	73	Anthropocene	38	botany	18	veganism	1
evolution	162	insects	71	capitalism	38	fossil fuels	18		
mental health	159	memory	70	flight	36	blindness	18		
relationships	154	anthropology	70	UX design	36	TED Books	18		

Table 11: Statistics on the number of talks per domain in the TED2025 dataset.

Lang Pair	#Sents	COMET-QE	Lang Pair	#Sents	COMET-QE	Lang Pair	#Sents	COMET-QE	Lang Pair	#Sents	COMET-QE	Lang Pair	#Sents	COMET-QE	Lang Pair	#Sents	COMET-QE
ar_en	1242114	74.58	ko_nl	603056	61.26	en_id	414970	78.6	bg_cs	290870	72.28	la_fa	144330	62.35	hi_ja	101055	64.02
en_es	1171895	78.54	en_ru	602559	67.92	ar_uk	414459	63.47	de_uk	290812	70.01	pt-br_sq	143939	65.7	ca_fa	100797	65.67
ar_es	1076362	67.49	fr	601008	66.3	el_nl	411252	71.5	ro	289965	66.43	pl_sq	143091	58.49	it_th	100559	64.29
ar_zh-cn	994593	56.73	hu_ru	599111	66.95	el_ro	410986	72.64	ro_th	289049	68.12	sk_th	142947	67.93	kmr_vi	99796	57.87
en_zh-cn	973667	68.84	ja_zh-cn	597792	58.76	hr_zh-tw	409256	65.9	pt_tr	288876	65	da_he	142920	68.94	my_ro	99349	59.51
en_fr	942555	78.07	en_pl	597234	71.96	hr_ko	406822	64.74	id	288405	70.38	es_it	142542	66.23	da_uk	99161	71.01
ar_fr	930909	65.64	es_sq	595397	67.29	el_fa	405959	65.36	de_fa	286615	70.47	da_ja	142244	60.65	my_pl	99872	64.29
ko_ko	922379	59.11	fr	595201	63.42	hr_ru	403875	70.08	de_id	286157	68.35	pt_th	142194	68.07	pt_br	97477	76.76
es_zh-cn	917132	61.02	it_ro	594433	69.53	bg_hu	402557	67.52	el	284286	69.11	he_it	141827	65.44	ca_pl	96318	65.14
ru_ru	916532	76.85	ja_pl	592311	60.64	id_zh-cn	402550	63.86	pt_vi	283273	68.58	sv_uk	141678	71.52	hi_hu	95946	69.46
ar_zh-tw	904984	58.14	pt-br_ro	591880	69.36	uk_zh-tw	402129	59.77	bg_uk	280180	66.54	id_zh-cn	140519	58.59	he_hi	95463	65.91
cs_fr	891473	70.34	pl-pt-br	587885	64.03	es_id	400666	69.39	sw_uk	274803	67.47	da-pt-br	139360	70.55	id_it	95440	64.5
en_ko	890124	70.57	he_hu	587631	67.31	en_pt	400349	77.17	el_id	271766	70.96	hu_sq	139156	59.85	da_th	95366	68.44
ar_pt-br	879212	67.73	pl_tr	586000	61.75	el_hu	399729	68.3	de	270867	65.05	da_tr	139059	63.87	kmr_pl	94413	65.86
en_it	878170	77.63	fr_hu	585533	67.48	ru_uk	398976	66.93	cs_el	269611	71.02	it_tr	138684	60.07	da_id	94277	69.27
en_ru	877084	62.75	ro_tr	583927	68.53	ko_uk	397906	60.44	fa_pt	266363	69.15	sv_th	138653	70.03	ar_fi	93954	63.87
ar_zh-tw	875551	70.77	ru_tr	583912	62.87	hu_sr	395571	68.03	ja	264092	61.77	fr_it	138611	66.97	id_my	93952	63.59
es-pt-br	852199	68.21	fr_nl	581625	68.29	en_uk	395281	74.99	el_th	263718	65.59	it_vi	137996	64.4	kmr_ro	93814	63.76
en_ru	850952	73.92	it_nl	579681	69.3	es_hr	394545	72.79	bg	262916	71.63	es_my	137828	62.85	ca_nl	93527	68.23
ar_it	847883	68.63	ar_de	579476	63.12	he_hr	393849	71.16	hr_uk	259520	69.92	nl_sq	137756	65.76	hi_vi	93155	65.35
es_ko	840427	63.57	fa_ja	577117	60.12	pl_sr	392935	65.49	bg_th	259002	63	da_vi	137054	69.4	ar_mrk	92840	63.44
ko_zh-tw	836982	57.84	ja_ru	577022	58.02	de_el	392903	69.08	cs	251810	67.23	ja_it	136887	59.28	ckb_zh-cn	92736	55.79
ar_vi	835188	60.61	en_nl	574465	74.32	hr_tr	391818	64.98	he_pt	251208	69.61	ko_my	136833	62.55	ca_de	92300	65.77
en_ru	834085	57.72	nl_zh-cn	573272	62.01	bg_fa	391024	63.69	es_hr	246496	68.76	de_sq	136631	62.52	id_zh-tw	92202	62.77
es_zh-tw	826342	62.19	fa_it	569719	69.54	hr_ji	389151	71.47	id_sr	244934	70.81	cs_sv	136553	69.84	mk_zh-tw	91924	60.04
ko_zh-cn	820514	55.92	he_nl	569437	67.38	ar_th	387921	60.43	th_uk	240998	65.41	my_zh-cn	136213	54.99	ko_mrk	91599	58.57
en_vi	819601	74.7	de_ko	566540	60.88	fr_hr	387881	72.49	hr	233343	69.03	bg_sq	136032	64.2	fi_ko	91568	62.31
ko_ru	817329	58.43	de_ru	566165	66.03	en_hr	387342	80.44	hu_pt	231917	66.73	en_sq	135269	75.78	fr_ru	91544	67.73
zh-cn_zh-tw	814804	59.22	zh-cn_zh-tw	565984	60.88	fr_zh-cn	386648	70.97	id	231647	69.17	my	135017	63.49	en	90728	63.49
fr_ko	812164	61.91	hu_ru	565854	61.79	hr_zh-cn	386582	63.07	ar	225106	60.12	id_sq	132944	68.69	mk_ru	90777	64.36
ru_zh-tw	808709	57.9	fa-pt-br	563585	68.58	nl_sr	383972	70.94	cs_uk	224284	70.41	fa_it	133962	64.81	sk_sv	90046	67.92
ar_zh-tw	807365	61.74	ru_pt-br	562777	69.25	ro_sr	383556	69.62	ar_sq	219186	66.47	it-pt-br	133915	63.4	kmr_sr	89752	63.3
en_tr	803643	69.44	hu_zh-cn	562549	61.05	fr_id	380333	69.33	ru_sk	216633	70.34	da_pl	133738	67.33	bg_ca	89729	67.71
fr_zh-cn	802960	59.75	fa_he	561511	63.16	id_zh-tw	380316	65.19	pt_ro	216447	67.21	fa_pl	133336	67.54	tr	89546	63.01
es_it	802859	71.06	el_vi	559884	63.9	ja_uk	380328	60.28	ko	216434	60.12	id_sq	132944	68.69	mk_ru	90777	64.36
es_ru	801560	67.16	hu-pt-br	559813	66.58	hr_ja	378739	62.86	sk_zh-tw	215458	63.18	da_pl	132314	68.62	fi_zh-cn	89334	62.16
fr_ru	792541	66.25	ja_nl	558716	60.73	ar_pt	378332	68.92	ru_sv	214533	67.74	it_pl	132181	64.39	ca_ro	88907	66.49
ar_ja	784497	56.38	nl_tr	557514	64.76	it_uk	377321	70.33	cs_th	213428	67.02	my_tr	131897	63.15	es_fi	88720	67.55
fr-pt-br	782007	69.2	de_fr	555403	67.25	tr_uk	376704	62.37	sv_zh-tw	211023	65.56	bg_da	131380	71.93	mk_tr	88567	60.92
ru_zh-cn	782001	56.46	de_es	552597	67.4	es_uk	376019	70.15	it_uk	210780	69.89	da_hu	130980	68.1	it_mrk	88419	67.46
es_vi	773676	66.34	de_zh-cn	550745	59.68	he-pt-br	375441	70	ko	210646	61.02	my_zh-tw	130701	58.12	es_mrk	88223	70.91
fr-pt-zh-cn	771080	61.87	de	548882	68.86	fr_uk	374373	67.84	he_uk	209261	68.5	hu	130019	63.01	fr_mrk	87857	69.59
ko-pt-br	769640	60.88	hu_ja	548815	60.43	id_ru	373973	66.87	id_uk	208344	70.49	el_sq	130019	67.52	mk_zh-cn	87776	58.56
ar_he	769088	63.17	hu	546990	68.89	he_uk	373946	66.87	es_uk	207954	70.01	da_de	129984	67.71	fi_it	87350	68.87
it_zh-cn	768684	62.08	fa_vi	540771	62.78	ru_th	371985	62.57	fr_sk	207151	70.04	fa_my	129218	51.56	kmr_nl	87266	65.77
en_ja	767381	67.19	de-pt-br	539297	68	he_el	370591	71.05	id	206594	70.04	ro	129134	65.92	it	87260	71.04
fr-pt-zh-tw	765142	63.79	ro_vi	538958	66.99	id	370340	64.68	en	206079	78.4	ar_kmr	129089	58.57	fi_he	87220	64.39
it_ko	763728	63.57	pl_ro	538936	64.87	th_zh-tw	370162	62.54	es_sq	204734	70.93	ro_sq	129001	62.8	ckb_tr	87019	55.63
ko_tr	762690	62.69	de	538274	65.08	ko_th	369872	59.23	sk_zh-cn	204644	61.05	en_it	128870	77.27	fi_fr	86735	68.32
fr_it	759270	70.8	he	537399	65.4	hr_vi	367480	69.78	fr_th	203474	70.21	ar	128754	71.48	ca_sr	86525	69.29
it_zh-tw	758320	64.39	ar_bg	535907	64.18	uk_zh-cn	365583	68.56	pl_pt	203742	65.16	ar_hi	128727	61.94	vi_ru	85431	66.65
vi_zh-cn	758118	58.82	de_hu	534905	60.27	id	365422	69.2	id	203628	67.15	de	128628	61.02	fi	85384	64.03
es_tr	758036	63.25	bg_ko	529377	60.53	es_pt	364577	69.57	ja_uk	202798	61.4	da_en	127603	77.74	fi_ja	85205	61.57
it_ru	755860	69.14	ru_ko	527966	64.19	uk_vi	363859	63.49	sv_tr	202544	65.03	en_hi	127575	74.11	iv_zh-tw	85107	61.77
tr_zh-tw	754441	60	nl_pl	526107	67.7	id-pt-br	361694	69.89	sr	202354	63.07	fr_my	127214	61.5	he_mrk	84836	69.53
pt-br_ru	748606	67.94	nl_vi	524453	67.49	en_th	361068	74.44	it_sv	202096	71.71	vi	126920	58.12	el_kmr	84364	69.19
ru_ru	743416	59.84	de_uk	522026	62.2	ga_uk	360247	62.62	sv_zh-cn	202046	63.58	kmr	126910	64.74	fi	84275	66.2
ja_ko	741597	63.41	de_en	521558	72.11	fa_ru	359517	71.19	sk_vi	200977	66.32	da_ru	124242	69.18	ko_lv	84272	58.64
ja_ru	738225	57.25	de_en	520960	58.58	id_it	358047	70.92	es_sv	200731	69.67	de_it	123263	67.47	hu	84242	67.94
ja_zh-tw	738006	56.72	he_it	518477	72.39	he_th	356554	63.74	nl_pt	200282	70.49	sq_sr	122703	65.64	de_kmr	84191	66.03
tr_zh-cn	735014	57.29	ar_el	516272	67.15	hr_pl	355076	66.9	pt-br_sk	199502	69.46	bg_it	122347	69.67	ja_mrk	83889	57.56
he_ko	731416	55.34	bg_es	515032	67.15	ar_cs	353364	68.36	pl_sk	199706	64.89	hr_my	122318	57.19	fa_fi	83877	65.84
fr_tr	730216	63.45	he_zh-tw	513676	67.78	pt-br_uk	352167	71.05	fr	198546	69.98	kmr_zh-tw	122148	57.91	de	83504	67.01
en_he	727009	77.91	he_hu	513379	71.16	it	352152	68.46	fa_sv	195714	68.61	cs_pt	121276	65.99	pt-br	83408	66.91
he_ru	725834	63.89	bg_tr	512323	70.2	he_id	351414	68.37	ja_sv	195050	61.82	kmr_ko	120297	61.03	fi_hu	83367	64.89
pt-br_zh-tw	724228	58.85	de_pl	511972	66.79	ja_th	351200	59.38	pt-br_sv	193655	70.9	ja_my	120050	62.06	my_th	83330	58.22
it_fr	721692	58.11	fa_hu	507281	66.84	hr_hu	349614	67.19	sv_vi	192166	69.76	kmr_ru	118804	64.93	sl_zh-tw	83277	62.43
es_ja	718188	59.91	nl_ro	506646	69.14	th_zh-cn	348388	60.81	id	191244	69.65	ar	118198	54.63	it	83022	66.27
it-pt-br	715066	68.82	hu_ru	503814	63.4	cs_ru	347890	67.99	pl_uk	190231	66.59	my_ru	117759	62.18	en_uk	82927	60.33
ko_vi	710975	59.13	ar_sr	502347	68.59	tr_vi	346956	65.55	en_sk	189541	77.09	fr_kmr	117606	65.11	ru_sl	82964	66.75
es_he	707993	68.38	el_ru	5													

Lang Pair	#Sents	COMET-QE	Lang Pair	#Sents	COMET-QE	Lang Pair	#Sents	COMET-QE	Lang Pair	#Sents	COMET-QE	Lang Pair	#Sents	COMET-QE	Lang Pair	#Sents	COMET-QE	Lang Pair	#Sents	COMET-QE
ro_sl	62309	65.04	nb_th	41483	68.87	id_mr	28855	62.92	bn_vi	23238	66.1	en_sw	20023	77.74	mn_sk	17389	57.28	sk_ta	13162	62.77
fa_gl	62195	66.65	hy_sq	41172	66.65	de_mm	28786	55.66	az_ru	23136	58.27	it_kk	19975	64.07	hi_lv	17362	70.18	gu_uk	13126	67.71
hi_th	61766	63.11	mr_zh-tw	41051	59.5	hr_mr	28741	61.72	bs_tr	23103	64.35	kk_vv	19960	57.71	fi_my	17350	62.66	fr-ca_hi	13071	67.4
fr-ca_hu	61740	67.53	ca_pt	40944	67.44	ar_ta	28701	59.23	fr_zh	23059	55.3	kk_ru	19953	62.81	lt_mn	17323	50.67	fa_uz	12972	58.59
nb_tr	61721	64.88	ko_mr	40861	57.46	ru_ru	28670	66.47	es_zh	23042	55.89	eo_vv	19940	62.58	bn_el	17306	66.77	da_mn	12894	57.22
gl_zh-cn	61551	61.4	mr_ru	40852	61.09	bg_mr	28482	59.18	az_sl	23039	62.1	be_fa	19909	61.3	hi_mk	17286	66.4	bs_sv	12879	71.79
ckb_ro	61414	64.44	ca_da	40750	64.95	mn_ru	28440	54.6	be_fa	23027	68.83	lv_mk	19841	65.22	kk_pl	17265	64.42	es_uz	12862	56.45
hy_id	61346	69.19	sl_sv	40580	68.09	ca_hy	28369	69.32	bs_pt-br	22950	70.16	mr_sv	19837	65.29	be_zh-cn	17197	55.22	ko_uz	12825	57.94
lv_uk	61183	66.36	hr_ka	40504	67.28	pt_sl	28315	67.3	az_zh-tw	22930	57.4	bg_bn	19824	65.58	be_fr	17161	64.58	es_uz	12820	55.61
he_nb	61043	69.77	ar_et	40194	60.93	hi_kmr	28312	61.12	bs_it	22902	71.89	ms_th	19782	66.91	ka_lv	17126	63.77	eo_sv	12787	67.71
fi_th	60689	65.39	el_ka	40126	63.99	ar_ru	28260	63.75	nl_ru	22882	71.69	gu_it	19781	71.55	bn_uk	17121	66.27	be_de	12733	63.47
cs_lv	60607	61.5	nb_uk	40126	71.13	ru_ta	28176	59.77	nb_pt	22869	71.02	mn_sv	19765	57.9	my_nb	17079	63.27	el_sw	12682	65.42
fi_uk	60369	67.11	fr_mr	39613	59.97	fr-ca_it	28019	66.5	ar_uz	22828	54.89	bs_de	19763	69.36	bn_cs	17031	71.02	gl_mk	12643	68.66
gl_ja	60292	61.48	et_zh-tw	39470	59.6	de_mr	27919	60.14	eo_it	22807	65.94	ta_uk	19755	60.35	kk_th	17025	58.53	en_uz	12605	68.42
cs_sl	60257	68.26	et_it	39396	66.73	gu_ko	27897	65.42	fa_gu	22775	68.07	bn_nl	19743	71.43	eo_sr	16994	64.07	et_hy	12568	68.97
gl_he	60181	64.29	ar_mn	39327	53.35	fa_ru	27841	64.79	bs_es	22773	71.13	es_eu	19707	63.62	pt-br_sw	16991	68.23	it_uz	12541	58.68
nb_pt-br	59946	70.65	et_ko	39184	60.15	ar_gu	27822	58.92	gu_ru	22694	67.19	he_kk	19685	60.69	bg_zh	16956	55.2	ja_uz	12530	59.06
gl_vi	59715	66.1	sk_sl	39176	64.98	ta_zh-tw	27729	65.35	eo_tr	22681	67.26	ar_sv	19635	61.62	mr_uk	16946	64.61	hi_ru	12494	64.75
hy_sr	59393	71.74	fr_mr	39081	66.39	da_fr-ca	27718	70.59	ja_zh	22668	52.78	id_ta	19626	64.77	az_br	16927	61.06	sq_ta	12451	59.64
ja_nb	59370	61.38	et_tr	39034	57.95	gl_sv	27669	69.53	bu_mk	22665	66.94	eo_ko	19608	60.61	sw_tr	16892	59.86	fr-ca_kmr	12444	64.61
cs_my	59125	60.86	fa_mr	38967	59.34	ta_tr	27550	61.7	de_ta	22650	63.8	bn_de	19583	67.93	ja_sv	16827	60.78	uz_ru	12421	54.52
da_sq	59094	65.53	et_zh-cn	38912	57.51	es_ta	27516	63.76	hu_ms	22587	62.27	de_eo	19563	63.3	gu_id	16791	69.13	bs_sk	12421	68.98
hy_th	59014	67.56	mi_sk	38860	61.54	ca_mk	27510	68.37	ja_ms	22565	59.52	ca_gl	19537	65.35	cs_zh	16676	55.99	hi_ta	12404	65.6
gl_it	58966	69.96	ka_ro	38839	65.8	hr_mn	27478	56.43	it_zh	22539	58.55	az_en	19496	67.92	mr_pt	16585	64.5	pl_sw	12365	63.34
he_nb	58519	67.93	he_mr	38829	59.79	fr_zh	27476	69.57	fr_zh	22476	64.07	eo_hu	19488	63.85	az_th	16562	57.18	gl_sl	12339	66.12
ckb_hu	58034	64.8	mr_tr	38771	59.25	et_tr	27474	63.71	eo_id	22475	59.18	fr_tr	19470	63.44	ka_pt	16527	68.13	fr_uz	12310	56.14
en_fr-ca	57838	77.03	et_fr	38519	64.46	ko_ru	27368	63.68	gu_pt-br	22474	70.83	sw_zh-tw	19459	62.71	mk_my	16510	59.47	az_sk	12257	61.21
en_nb	57805	79.09	de_mn	38466	55.85	ur_zh-tw	27255	63.45	nl_ta	22441	65.43	fr-ka_lv	16509	63.92	fr-ca_lv	16509	64.06	mr_sq	12256	58.03
mk_th	57666	63.23	ca_sr	38408	68.84	mr_uk	27138	61.46	be_fr	22433	71.93	eu_zh-cn	19430	54.12	da_et	16428	67.72	pt_tr	12243	71.63
fr-ca_ro	57611	67.98	es_kr	38366	60.75	ca_lv	27123	61.62	gu_ru	22393	67.16	id_ru	19426	71.55	gl_hi	16407	66.6	et_lv	12218	60.78
nb_pl	57576	68.01	lv_sv	38141	66.86	tr_ur	27086	63.83	eo_zh-tw	22319	60.18	eu_tr	19418	57.48	th_zh	16333	56.85	fr-ca_my	12206	61.04
gl_hu	57234	67.24	mm_ru	38121	54.51	he_ta	27010	59.85	ms_pt-br	22286	65.62	az_ro	19418	56.86	sw_zh-cn	16325	60.37	be_cs	12195	66.89
hy_uk	57028	68.8	ko_mn	38082	57.55	mr_th	26991	56.83	bn_ko	22276	64.43	eo_pt-br	19397	66.27	eo_uk	16320	64.88	gu_hr	12183	72.61
fi_id	56914	64.69	et_ja	38026	57.93	ta_zh-cn	26931	56.98	az_es	22258	58.49	gu_he	19393	66.39	kk_nl	16284	64.74	bn_sv	12169	70.14
sl_th	56724	66.49	lt_sl	37983	60.73	bn_ru	26903	60.73	bn_de	22257	65.49	bn_hr	19389	71.84	ca_ka	16175	67.12	ar_ml	11966	60.09
gl_nb	56600	67.12	bg_uk	37961	65.81	ka_sk	26776	67.12	az_de	22254	59.82	lv_sl	19365	62.39	kk_pt-br	16119	63.87	sk_tr	11937	68.87
pt-pt-br	56578	67.96	ka_et	37846	66.3	gu_ru	26764	64.27	ka_sq	22240	62.36	eu_ja	19355	58.77	uk_zh	16116	54.37	sv_zh	11928	59.52
lv_th	56527	65.73	et_pt-br	37667	65.16	fr_zh	26766	61.89	pt-br	22163	67.85	hi_th	19351	67.85	th_sl	16112	67.18	fr_uz	11915	66.91
pt_sq	56464	66.31	es_et	37614	64.9	el_mn	26676	54.79	uk_ru	22142	66.13	az_el	19316	58.55	hy_nb	16068	71.38	sv_ta	11922	66.31
fa_nb	56419	69.16	et_pl	37564	64.03	bg_mn	26670	51.37	kmr_sl	22102	61.45	hr_ta	19224	66.11	eo_id	16054	65.23	eo_lu	11836	59.67
fr_fr-ca	56408	71.27	hi_my	37548	63.11	da_hi	26669	69.97	az_tr	22094	63.69	gu_ja	19218	65.61	kmr_mr	16046	51.64	sw_uk	11826	66.39
cs_hi	56176	68.12	et_he	37462	62.98	en_ru	26655	76.77	eo_he	22093	65.22	be_zh-tw	19178	56.75	az_id	16033	59.39	bn_ru	11818	68.23
bg_nb	55914	71.8	mn_zh-tw	37406	53.22	ko_ja	26604	61.38	bn_fr	22070	69.88	fa_sw	19173	66.28	be_pt-br	15982	65.6	ms_pt	11789	66.75
cs_fr-ca	55450	68.81	hy_sv	37356	72.18	gu_ru	26581	67.08	bn_zh-cn	22057	60.76	ku_sw	19124	73.52	fr_zh	15948	68.8	fr_uz	11745	65.34
fr-ca_hr	55429	72.75	cs_ru	37300	64.56	fr_tr	26391	69.22	bs_en	21998	79.63	ms_sr	19102	66.98	eu_hr	15942	63.02	be_id	11737	64.08
id_mk	55413	70.51	mr_zh-cn	37210	56.63	fr_ta	26235	62.64	fa_zh	21985	55.54	eo_fa	19099	63.13	eo_th	15885	62.9	ml_zh-tw	11695	59.21
de_nb	55166	68.54	ja_mr	36933	57.31	nb_sq	26220	65.87	bs_vi	21947	69.69	kk_ko	19078	64.2	hr_zh	15833	58.34	pt-br_uz	11625	57.15
nl_sl	55141	66.25	ja_mn	36928	58.89	ar_ms	26161	64.18	gl_kk	21942	67.17	be_it	19049	66.37	hy_my	15825	62.51	ka_nb	11598	67.74
nb_nl	55142	69.03	en_mn	36906	66.06	fr_ru	26152	69.1	eo_ru	21931	64.08	ml_uk	19061	61.26	be_nl	15823	67.16	fr-ca_lv	11568	64.39
hi_pt	55241	71.39	pt_br	36983	63.49	et_zh	26123	65.49	az_ko	21923	68.76	hi_tr	19023	59.61	lt_mr	15776	67.1	ckb_uk	11524	70.32
ca_sk	54232	66.05	mn_vv	36734	54.09	es_ru	26057	70.59	ar_ja	21888	58.95	cs_sv	19009	66.7	hi_nb	15738	72.79	bs_sq	11559	62.87
gl_sq	54156	56.71	et_hu	36657	61.03	fr-ca_pt	25925	68.34	az_fr	21880	58.92	eo_hr	18989	65.3	ca_nb	15731	66.5	th_uz	11534	56.71
lt_pt	53925	63.74	it_mr	36650	62.65	et_th	25852	64.12	az_he	21849	58.59	be_ru	18988	64.69	ca_et	15707	61.66	ml_tr	11495	61.02
ca_sv	52925	66.43	cs_gl	36604	62.84	cs_mn	25800	54.98	sr_uz	21795	70.89	bn_pl	18970	65.02	bg_pl	15631	65.02	bn_ru	11463	63.24
fr-ca_uk	52898	66.73	sl_sq	36579	58.68	et_sq	25797	68.09	eo_sq	21737	64.21	cs_ru	18914	67.37	az_sr	15629	60.31	eo_sk	11450	65.18
gl_pl	51375	66.6	mr_ru	36547	57.31	gu_zh-tw	25733	65.37	bn_pl	21720	70.36	ru_sv	18891	63.09	kk_sr	15552	63.45	ka_sl	11429	65.76
ckb_pl	51099	66.74	fa_mn	36461	55.77	it_ru	25707	71.27	vi_vv	21692	56.94	hi_tr	18899	61.57	be_hu	15549	61.16	hi_mn	11406	53.06
el_nb	51088	72.82	fr-ca_sq	36388	61.53	ar_zh	25696	51.97	bn_fa	21691	64.71	fi_nb	18779	69.15	bg_kk	15478	57.82	hi_ka	11394	65.98
ka_ru	51082	65.91	et_nl	36312	66.24	hu_ru	25645	68.3	gl_ta	21685	63.15	kk_zh-tw	18769	58.06	eu_cs	15408	61.18	th_uk	11366	66.36
gl_nl	50958	69.47	hy_mr	36226	59.99	he_ru	25522	67.7	pl_ru	21638	67.81	eu_sl	18749	64.21	hu_sw	15407	64.91	ko_ml	11353	59.43
fr-ca_id	50569	68.86	it_pt	36199	58.48	hy_pt	25483	71.41	hy_lv	21634	66.56	hy_sl	18727	69.37	az_uk	15389	60.31	ms_kk	11302	65.45
ckb_pt	50551	68.2	de_et	36069	64.42	ca_kmr	25446	65.15	id_ms	21607	63.96	ku_he	18677	63.2	de_kk	15385	62.4	de_sw	11236	65.08
fr-ca_sq	50486																			

Lang Pair	#Sents	COMET-QE	Lang Pair	#Sents	COMET-QE	Lang Pair	#Sents	COMET-QE	Lang Pair	#Sents	COMET-QE	Lang Pair	#Sents	COMET-QE	Lang Pair	#Sents	COMET-QE	Lang Pair	#Sents	COMET-QE
gl_nb	9763	69.76	it_ne	7276	68.96	ne_uk	5902	67.79	ca_sw	4657	64.38	kn_zh-tw	3939	57.23	en_kn	2984	69.3	ka_uz	2385	57.03
sw_th	9717	66.26	fil_tr	7235	55.82	eo_lv	5847	60.91	ar_sh	4655	68.13	de_kn	3937	62.02	bn_ckb	2975	64.3	fi_ne	2380	66.46
lv_mr	9713	60.05	fil_ja	7210	57.82	id_ml	5817	63.08	az_my	4643	62.26	cs_srp	3933	68.93	cs_si	2972	63.69	de_tl	2370	58.68
ckb_sq	9677	56.9	eu_mk	7177	59.87	fil_id	5816	61.9	ckb_gu	4633	62.64	pt_srp	3924	71.02	cs_kn	2971	58.32	de_it	2361	61.46
kk_pt	9660	64.35	te_th	7167	60.45	ne_th	5800	62.89	is_zh-tw	4615	58	kn_tr	3923	61.92	nn_uz	2970	54.7	pt-br_tl	2361	62.52
fi_mr	9659	61.05	et_kmr	7154	59.83	eu_pt	5798	64.97	sr_srp	4608	71.38	is_gl	3913	61.64	bs_zh	2969	70.56	el_kn	2357	58.97
ar_te	9648	57.75	sq_uz	7149	49.5	az_kmr	5783	54.73	bg_us	4598	59.21	hgn_kn	3904	60.9	lm_zl	2964	57.47	eu_nn	2354	52.06
az_sq	9604	53.53	gu_kmr	7124	63.73	ru_srp	5779	69.01	gu_sk	4595	69.04	hr_kn	3902	63.07	et_kk	2935	59.62	en_nn	2353	78.32
be_pt	9569	65.05	es_fil	7116	61.7	ne_pl	5764	71.22	hy_uz	4586	60.16	et_ms	3885	66.56	gu_mk	2926	61.36	gu_mk	2333	68.46
da_eo	9527	64.91	fil_ru	7116	62.01	si_vi	5758	65.81	is_vi	4581	62.97	kn_ko	3872	59.7	fil_hi	2919	62.84	ar_qi	2325	49.95
mn_sl	9527	55.84	eu_fi	7114	59.51	gu_sw	5750	71.19	gl_ms	4574	64.76	gl_gu	3870	70.56	bs_ta	2914	66.79	nl_tl	2316	65.27
fi_mr	9525	55.82	ckb_fi	7103	64.57	be_lv	5737	61.45	is_tr	4574	59.29	3865	68.78	nn_zh-cn	2911	61.28	lv_ml	2316	57.06	
hu_ml	9498	61.66	fil-pt-br	7103	63.93	el_fil	5728	62.71	sq_sw	4573	59.73	cs_so	3864	55.14	arq_en	2911	66.15	ja_nn	2312	61.3
te_tr	9494	61.52	ckb_nn	7092	51.96	hi_te	5722	63.4	kk_mr	4569	57.44	ms_tr	3859	69.28	eo_mr	2902	57.31	et_fil	2310	54.28
be_sk	9480	66.02	az_pt	7046	60.25	my_uz	5706	59.45	is_it	4563	65.91	be_si	3848	62.48	kn_th	2883	63.63	ca_si	2309	61.73
et_nl	9456	63.91	fil_jv	7044	61.25	fr-ca_ms	5705	64.66	ckb_zh	4560	53.24	da_ne	3856	68.24	ml_sl	2883	62.81	az_kk	2307	60.95
hi_ur	9437	67.83	hi_sw	7040	63.17	bn_hy	5674	68.58	fil_tl	4539	56.82	be_ka	3849	60.36	it_nn	2875	68.03	hi_kn	2298	65.96
ca_ur	9410	66.46	nl_te	7032	65.38	pl_srp	5660	66.07	ne_ro	4538	69.61	eu_gl	3829	61.49	is_sk	2859	64.42	eu_ta	2291	65.15
de_ml	9288	63.38	eo_fr-ca	7027	64.87	bn_ka	5659	65.63	it_so	4533	50.67	sw_ur	3826	63.29	nn_tr	2858	63.63	is_uk	2289	64.13
my_ur	9268	60.22	fil_fr	7023	64.04	cs_ne	5651	67.02	he_so	4533	49	mk_zh	3817	54.45	arq-pt-br	2851	56.19	ar_qd	2288	55.13
ml_nl	9258	63.17	fil_hu	7022	61.23	pt-br_si	5640	68.48	ar_so	4529	46.84	nn_so	3779	62.39	hu_nn	2843	64.9	nn_srp	2272	64.27
hy_nn	9234	56.11	ml_uk	7022	61.36	hr_ne	5637	68.08	is_ko	4515	60.02	fa_kn	3771	63.31	bn_te	2832	63.3	mk_ml	2272	59.35
ca_eu	9229	60.19	ne_vi	7020	63.39	ka_mr	5619	59.93	spr_th	4513	69.8	el_is	3765	63.58	it_tl	2830	63.47	nn_th	2265	69.8
en_te	9221	71.9	be_sl	7008	62.64	eo_ka	5618	62.89	pl_sh	4502	66.51	sl_uz	3758	55.19	nn_ru	2825	66.42	arq_de	2259	47.72
id_sw	9203	67.71	he_ne	6991	65.35	ckb_lv	5614	63.74	ja_so	4497	63.74	kk_ms	3699	62.53	he_nn	2719	54.76	fr-ca_ne	2256	64.93
fa_te	9189	61.92	fil_he	6977	63.19	fil_he	5607	62.51	id_si	4487	68	az_nn	3753	53.99	el_nn	2818	72.06	so_sv	2249	51.43
ru_te	9168	58.25	bn_hi	6977	69.55	be_ca	5602	60.83	gl_uz	4486	56.56	az_bn	3752	62.27	arq_hu	2818	55.33	kmr_ml	2246	59.69
ko_te	9167	59.29	gl_ka	6977	66.59	fa_si	5600	65.24	da_gl	4483	64.64	ml_uz	3745	55.09	be_ur	2816	61.83	gl_srp	2239	63.48
kk_sv	9116	64.09	et_my	6974	60.52	en_uz	5598	67.37	et_nn	4473	52.69	so_uk	3737	52.07	en_is	2804	74.55	arq_cs	2239	49.84
bn_ca	9107	66.4	az_lv	6924	57.09	ms_sl	5598	62.21	fa_sh	4466	70.29	ca_ne	3729	65.6	sh_sq	2792	60.86	az_fil	2229	52.96
ca_ms	9106	61.66	hu_ne	6922	65.96	ml_th	5596	61.91	ml_sw	4465	63.82	ca_gu	3719	69.72	arq_fr	2779	53.37	fa_ps	2220	59.02
fi_ja	9066	62.97	ne_tr	6915	64.86	en_fil	5465	63.22	gl_ur	4454	67.26	sh_th	3652	66.5	fr_nn	2776	65.69	ar_kn	2199	62.73
ro_uz	9042	53.83	cs_ru	6914	68.6	mr_ur	5557	64.18	en_nb	4447	63.28	ca_tr	3708	64.66	arq_nl	2773	56.07	gu_nn	2197	60.96
bs_tl	9039	63.93	bn_nb	6908	72	pt-br_srp	5536	71.95	ko_so	4445	49.66	et_ta	3707	62.62	nn_zh-tw	2772	62.36	tl_tr	2196	58.2
hi_ms	9023	69.82	de_fil	6891	61.61	hy_kk	5524	61.23	so_zh-cn	4433	46.87	ckb_si	3707	63.92	eo_fil	2757	63.92	fil_nb	2193	64.34
te_zh-tw	9022	58.65	az_sl	6891	59.57	eo_kmr	5524	58.04	ckb_kk	4426	53.5	kn_pl	3701	59.52	eo_ta	2752	64.62	en_ps	2182	71.83
ml_nl	9013	63.18	be_da	6885	64.53	bs_lv	5496	64.56	cs_so	4415	47.38	so_sl	3701	51.25	is_sq	2745	57.39	hy_is	2181	63.05
ml_pt	9006	60.12	fa_ne	6885	63.45	sl_zh-tw	5496	63.22	fil_sk	4414	61.87	ca_gu	3699	66.13	arq_it	2743	54.76	fr-ca_ne	2180	64.12
et_pt	8999	65.64	ar_si	6880	64.73	az_ka	5461	58.72	fr_so	4403	47.75	ms_ta	3693	64.88	eo_et	2737	59.51	be_eo	2163	60.06
ml_sr	8940	64.36	fa_fil	6841	58.77	bs_sl	5456	67.93	bn-fr-ca	4388	68.25	be_kn	3671	58.27	sq_srp	2735	62.35	sl_sv	2158	68.37
cs_sw	8917	65.34	ms_nb	6836	66.83	az_nb	5423	60.32	gl_kk	4386	63.42	is_th	3664	62.56	be_et	2732	62.08	lt_so	2151	43.18
gu_th	8879	66.36	bn_nb	6812	63.28	eu_sl	5411	59.44	ckb_ms	4381	65.43	kn_kl	3644	63.03	be-fr-ca	2730	63.41	da_te	2150	61.59
my_zh	8849	54.59	es_ne	6810	67.61	kk_nb	5397	64.34	be_hy	4381	64.09	is_ja	3641	56.11	fil_pt	2719	63.56	mn_sw	2146	54.32
cs_gu	8826	67.74	ka_kmr	6798	54.37	sk_sw	5377	64.57	pl_tr	4381	64.74	kk_ms	3639	62.53	he_nn	2719	65.51	kn_hu	2146	60.96
it_te	8776	64.7	da_kk	6747	63.13	bn_uz	5371	66.49	ja_si	4375	61.36	ka_kk	3639	62.27	bg_nn	2713	69.08	nn-pt-br	2144	64.84
eu_tl	8759	56.03	fil_ko	6726	58.79	en-fr-ca	5299	62.61	de_js	4370	63.98	fa_is	3636	61.57	cs_nn	2700	63.58	is_lv	2132	59.08
bs_da	8743	69.14	kn_eo	6725	73.58	fr_si	5297	67.83	ru_so	4344	46.47	hr_si	3632	67.23	ta_te	2684	58.71	ja_ps	2125	61.54
ur_tr	8722	70.01	it_srp	6719	72.56	cs_te	5297	60.27	ru_sh	4328	70.13	sh_srp	3628	62.91	lt_zh-cn	2672	56.19	ko_tl	2119	58.26
be_my	8705	56.53	lv_ur	6708	66.35	fr-ca_mr	5293	59.47	cs_is	4323	59.35	ka_ms	3587	69.18	bn_ka	2658	67.08	pt-zh-tw	2114	60.79
sk_sq	8658	56.17	ka_ne	6705	64.24	hi_kk	5291	62.69	be_th	4311	55.08	th_ms	3587	62.29	sa_th	2658	54.36	arq_tr	2114	60.79
ca_eo	8657	60.68	hy_mr	6690	61.47	mr_uz	5271	61.27	hy_ur	4303	69.49	be_ml	3546	60.64	id_kn	2652	61.07	pt_br	2104	66.67
mk_nn	8647	50.65	fil_kmr	6689	54.67	kk_sl	5270	61.53	be_mm	4303	49.75	bn_sq	3546	55.79	arq_el	2643	56.61	da_nn	2103	68.84
es_te	8602	63.29	te_uk	6686	60.68	el_te	5261	60.41	ckb_uz	4297	52.92	si_uk	3542	65.88	fi_te	2641	60.67	ar_ps	2102	56.35
it_nl	8545	61.52	ko_ne	6672	62.13	be_hi	5256	62.45	ne_sk	4295	71.93	hu_kn	3534	62.1	id_tl	2636	61.04	lt_srp	2102	63.89
fr_te	8526	62.58	da_zh	6666	57.72	eu_lv	5248	56.42	ka_ur	4289	67.27	fr-ca_uz	3527	56.58	es_tl	2623	62.52	km_tr	2097	62.06
da_te	8506	65.29	ar_srp	6661	70.41	fr-ca_nn	5243	55.77	hu_uz	4289	62.21	bn_ms	3525	62.85	bn_ms	2618	69.58	ro_tl	2098	78.25
be_tr	8478	58.99	az_srp	6647	72.91	ar_kk	5213	57.22	it_zh	4257	58.12	be_zh	3512	60.2	sv_ta	2613	64.9	eo_km	2096	60.86
ms_my	8476	62.03	nb_ja	6642	66.56	bn_ja	5198	64.47	pl_so	4265	46.23	bg_si	3506	66.38	arq_ja	2609	51.17	bn_uz	2083	61.6
pl_te	8468	60.89	lt_te	6622	51.42	it_si	5187	67.98	ru_ru	4257	62.45	sv_te	3503	65.14	arq_zh-cn	2608	47.98	kn_kn	2083	63.91
ja_te	8463	68.84	da_uz	6553	58.93	ro_srp	5187	69.23	so_zh-tw	4251	50.09	fa_so	3501	49.75	bn_eo	2608	65.44	kn_kn	2083	66.38
kk_tl	8461	58.21	cs_fil	6533	59.86	ml_sk	5157	61.95	fil-fr-ca	4248	63.86	eu_kmr	3491	56.93	arq_uk	2605	53.34	kn_sw	2083	63.8
bg_ml	8458	61.88	fi_ms	6505	60.77	sl_uz	5150	67.65	ar_js	4239	60.34	et_eu	3455	52.62	fa_nn	2604	66.49	hr_tl	2083	61.47
te_vi	8370	60.57	ca_uz	6495	53.81	lv_sw	5150	59.09	az_eu	4239	54.42	eu_lv	3450	66.86	fil_lv	2601	56.91	cs_zh	2079	56.91
el_ml	8362	61.41	en-srp	6496	80.08	pt_uz	5144	57.65	lv_sw	4235	65.45	de_si	3450	66.87	fr_sw	2596	61.35	bo_vi	2076	41.92
be_sq	8345	58.57	sl_ta	6495	62.93	fil_sq	5140	54.7												

Lang Pair	#Sents	COMET-QE	Lang Pair	#Sents	COMET-QE	Lang Pair	#Sents	COMET-QE	Lang Pair	#Sents	COMET-QE	Lang Pair	#Sents	COMET-QE	Lang Pair	#Sents	COMET-QE	Lang Pair	#Sents	COMET-QE
fil_sl	1886	59.82	gu_ta	1557	66.39	ja_pa	1304	63.13	hy_srp	1082	65.79	bg_lo	854	64.75	pa_uk	665	62.48	ta_tg	532	44.58
km_ro	1881	67.57	ps_pt	1554	68.04	arq_my	1298	53.16	km_nm	1082	57.54	eu_lo	854	68.47	si_ur	665	57.54	lo_nl	530	69.82
kn_it	1878	58.29	th_tl	1553	61.61	km_nm	1296	70.71	ml_my	1079	59.78	et_lo	854	62.6	ja_sc	664	45.6	af_be	524	59.74
arq_pl	1870	46.11	ca_srp	1552	67.46	id_ug	1296	69.21	tg_uk	1079	37.69	lo_tr	854	62.66	bs_ne	663	67.47	srp_ur	523	66.31
bo_ru	1870	41.9	ms_srp	1549	66.72	af_da	1294	63.71	ht_lo	1076	46.13	lo_th	854	64.58	cs_pa	661	69.44	es_lo	523	65.34
bo_de	1868	41.21	eu_mr	1547	57.91	ms_so	1290	46.86	pa_uk	1073	67.43	ar_lo	854	61.37	mg_vi	661	48.29	lo_ru	523	59.54
hi_nn	1865	71.95	pa_zh-tw	1542	65.66	ckb_srp	1283	64.88	gl_nm	1073	65.89	da_lo	854	68.15	fr_mg	661	45.43	eo_kn	522	65.77
ta_uz	1864	60.22	am_de	1542	62.34	de_km	1283	68.27	km_kmr	1072	57.75	it_lo	854	67.64	mg_uk	661	47.49	km_nm	519	53.66
mk_te	1863	60.57	es_pa	1541	70.31	am_ro	1278	57.09	sr_tg	1071	39.21	id_lo	854	69.12	ko_mg	661	47.49	ky_uz	518	57.91
sl_te	1862	63.62	gl_ml	1538	64.06	lv_si	1277	62.72	hi_srp	1067	71.45	ca_lo	854	64.76	es_mg	661	45.23	eo_srp	515	64.9
gu_ms	1856	67.06	he_ps	1537	62.19	pt_so	1276	49.53	am_kmr	1066	59.48	hy_lo	854	63.15	ar_mg	661	44.97	fi_kn	514	71
nn_sv	1855	71.61	pa_zh-cn	1536	63.16	bs_fil	1272	61.05	fr-ca_ky	1065	64.75	eu_lo	854	64.05	mg_zh-tw	661	46.56	fi_tg	512	43.93
bs_eu	1847	59.46	ar_ug	1526	64.98	et_tl	1274	56.83	lt_tl	1064	54.01	hr_lo	854	66.7	it_mg	661	45.97	bn_so	512	53.43
en_tl	1845	72.59	it_ug	1521	67.45	fr_ht	1274	36.5	fil_my	1063	59.24	bn_lo	854	68.4	id_mg	661	45.08	ka_tg	511	43.56
am_he	1845	59.07	hu_ug	1518	65.35	kmr_ug	1272	52.61	ms_sw	1063	66.49	is_uz	853	53.24	mg_nl	661	51.62	kmr_ps	510	51.24
da_kn	1844	62.59	eo_ne	1512	62.26	he_ug	1271	65.85	hi_pa	1062	73.84	bn_si	845	69.27	he_mg	661	47.03	fi_ky	507	66.41
nn_ro	1843	67.09	af_hu	1510	64.07	ht_tr	1270	43.29	ckb_eo	1059	53.29	fr-ca_nm	843	65.41	ko_sc	661	44.9	az_so	505	46.07
ps_tr	1842	60.51	af_lo	1508	61.4	kmr_srp	1270	61.4	hy_srp	1058	44.51	hi_km	840	66.92	ru_sc	659	40.1	lv_mfe	503	40.8
am_ja	1840	61.1	fil_mk	1507	62.55	ja_ug	1269	64.66	el_pa	1052	66.66	arq_fi	839	53.46	sc_zh-cn	657	42.56	cs_sw	502	55.5
am_tr	1838	62.56	ru_ug	1505	65.89	nl_ug	1269	69.49	da_tl	1050	62.82	lo_zh-tw	839	58.96	mg_sr	655	46.18	ka_ky	501	64.55
bn_zh	1837	59.4	ko_ug	1504	65.48	de_ug	1267	68.1	ta_tl	1049	57.59	gn_lo	838	46.95	af_gl	654	63.25	da_ky	501	68.53
am_zh-tw	1835	57.88	be_eo	1504	55.38	ht_zh-cn	1266	41.45	ko_pa	1046	65.11	ky_sr	836	61.85	hy_te	653	62.91	cs_lo	500	64.42
eu_my	1832	59.79	fr_ps	1503	64.41	fa_ht	1263	43.64	arq_sl	1044	43.1	af_sr	831	61.81	et_nm	650	63.99	ps_te	498	65.11
bo_nl	1832	43.99	pa_ru	1501	66.84	gu_sw	1262	68.07	bs_ml	1041	65.4	lo_pt-br	830	68.13	eu_kk	649	56.46	mfe_ru	497	44.48
arq_ro	1831	47.37	af_lo	1495	65.58	ps_sv	1261	67.79	ne_uz	1038	63.79	en_nm	829	62.28	hy_nm	649	37.67	mr_pa	497	47.28
lt_si	1830	60.89	ar_ug	1494	66.21	cs_ug	1260	66.57	hy_tl	1036	65.97	bo_nm	829	68.11	bo_nm	649	42.53	tg_ur	493	47.92
km_pl	1830	67.67	bo_lt	1490	42.88	bg_ug	1260	63.15	fi_tl	1036	60.66	kmr_pa	826	60.09	ca_ps	648	61.2	ta_ug	492	64.15
is_sl	1827	62.6	af_ar	1484	63.07	af_sr	1257	67.67	mr_ps	1030	60.12	bo_mk	826	40.06	am_ka	647	60.07	fi_pa	490	69.41
sr_tl	1821	60.05	fr-ca_te	1484	62.77	fil_nm	1256	52.38	ml_sw	1026	61.88	lt_sr	824	39.05	fa_mg	647	49.02	ky_lv	489	63.51
kn_sr	1819	63.54	ps_sr	1483	67.91	mm_te	1254	57.32	si_sq	1022	60.27	sv_ug	824	68.7	bg_mg	647	42.36	fi_sl	489	62.48
am_ar	1817	59.12	id_kn	1482	70.45	ht_ru	1254	42.48	eu_ml	1022	62.17	ro_ug	823	66.3	ro_tg	647	47.67	et_si	488	62.93
gu_kk	1811	64.14	af_pt-br	1478	68.01	hi_tl	1252	64.59	nn_srp	1022	55.83	de_mg	823	41.69	de_mg	647	45.47	mr_ar	486	53.05
be_zh	1810	50.21	ml_nm	1478	53.69	lv_nm	1247	64.27	fa_ky	1019	64.39	lv_tl	822	57.41	mg_zh-cn	647	42.54	is_ms	486	57.87
it_km	1810	70.53	ps_zh-cn	1476	59.69	nb_srp	1243	73.56	bo_kmr	1019	45.66	hi_ps	822	61.93	mg_pt-br	647	46.99	fr_lo	486	65.89
ca_nn	1804	64.35	am_hu	1476	60.75	be_eu	1242	53.63	ky_uk	1018	63.79	ml_nb	822	65.94	et_te	646	62.95	arq_sh	483	36.33
hy_hm	1804	73.93	af_pl	1472	62.72	de_ps	1242	64.28	et_ug	1015	64.47	eo_tl	821	65.04	kn_mk	646	64.88	eu_sh	483	61.08
lt_nn	1801	61.55	bo_kn	1468	45.21	hi_so	1240	52.92	af_sv	1013	66.86	ne_uz	818	57.25	lv_ug	645	63.04	ckb_ps	482	49.54
fa_kn	1801	63.34	af_eo	1468	63.97	eu_ms	1240	56.45	nn_srp	1008	57.14	km_pt	814	68.83	hy_kn	645	61.5	mr_ar	482	45.73
am_vi	1800	63.6	ps_ro	1464	66.71	ht_zh-tw	1236	43.83	arq_gl	1007	37.38	lv_tg	810	40.13	fr-ca_srp	642	68.36	bs_is	478	63.01
gu_ka	1795	66.29	af_sk	1460	64.59	kk_ne	1234	62.26	ps_sq	1006	60.11	srp_zh	806	62.43	sc_tr	640	45.98	lo_vi	478	64.34
am_en	1782	74.17	af_th	1460	65	pa_sr	1227	70.8	lt_ug	1004	61.59	da_tg	805	43.92	be_sw	637	59.2	bs_lo	477	68.19
be_uz	1781	54.17	sv_tl	1459	64.73	fr_pa	1226	69.65	de_ht	1004	36.32	ko_tg	805	45.11	sc_zh-tw	636	46.67	ka_lo	477	59.07
arq_sv	1778	54.62	pa_ro	1458	70.16	hy_nm	1225	67.38	bg_ht	1004	36.77	ht_ja	801	44.38	eu_ur	629	64.06	lo_ta	477	65.35
sq_tl	1771	53.11	th_ug	1454	61.31	bo_hi	1224	46.88	ckb_ne	1003	57.14	de_zh-cn	801	68.93	hy_kn	614	61.5	arq_uz	477	64.62
am_sr	1770	62.3	sr_ug	1452	64.62	ar_ug	1218	66.24	arq_hi	999	53.98	kk_te	800	61.77	am_ar	628	59.93	hi_lo	477	66.27
az_be	1770	56.56	af_de	1449	64.69	ps_th	1215	60.13	el_km	999	65.89	sl_tg	799	39.38	bo_nb	628	41.73	eu_kn	474	60.86
am_fr	1767	63.61	af_mk	1445	41.2	ckb_te	1213	58.88	bo_my	998	53.05	am_sl	799	59.13	si_ta	627	64.87	eo_nm	474	68.21
ro_7666	1766	36	ar_pa	1444	67.02	pl_ps	1212	64.09	et_ml	998	58.92	ckb_pa	799	60.95	fr_sc	627	39.22	et_lo	473	62.13
ps_pt-br	1765	67.74	ta_te	1442	66.08	ca_so	1209	45.3	af_sq	987	61.27	nn_ta	795	67.44	ht_uk	626	44.82	ml_ms	468	64.07
ca_te	1756	72.68	af_lo	1441	46.07	ug_uk	1206	64.54	af_ro	985	67.08	my_ps	794	57.27	te_te	626	57.17	fr-ca_ug	468	47.28
arq_sq	1752	48.41	fr-ca_so	1440	47.35	af_tl	1205	59.86	pt_tl	980	60.33	fr_zh	793	61.9	ht_id	625	41.95	ar_uz	464	65.12
da_so	1751	51.29	es_ps	1439	67.16	ug_zh-cn	1203	62.89	cs_km	980	70.39	kk_sw	791	54.78	gu_uz	625	65.41	ky_sl	462	59.37
fr_km	1739	67.38	af_it	1437	68.37	ht_it	1203	42.06	fr-ca_tg	977	41.57	bo_sk	791	42.44	az_ky	624	59.16	ka_tl	461	62.4
bs_uz	1736	56.61	fr_pa	1437	56.64	kk_si	1203	56.64	am_hi	976	63.88	am_hi	791	65.52	fa_sc	622	45.43	ht_bt	461	36.67
am_nl	1717	65.48	af_el	1436	69.8	bo_sq	1202	39.96	de_pa	974	69.18	eu_zh	790	53.29	be_gu	620	61.7	lo_lv	460	65.67
nl_sl	1716	45.9	eo_kk	1433	55.54	bn_lo	1202	68.78	ly_pt-br	974	66.4	ja_lo	789	58.32	fr_sl	619	63.68	si_sl	459	69.43
ht_kk	1714	64.69	af_te	1432	68.21	ky_te	1199	65.32	fr_te	963	59.67	gl_lo	789	66.26	eu_nm	619	40.91	az_ne	458	69.56
am_pl	1713	60.77	bs_kk	1427	65.04	fil_ta	1198	63.48	km_pl	962	70.98	fr-ca_lo	786	63.33	am_ta	618	63.16	pa_sl	458	71.09
mk_ne	1709	61.55	pt_jk	1422	71.43	et_sh	1196	64.42	nl_pa	961	72	lo_sv	786	70.79	bn_tl	617	62.51	ca_bt	457	36.82
mk_sh	1709	70.52	af_hr	1422	70.12	id_pa	1193	71.73	ly_zh-cn	961	57.53	en_tg	785	51.12	is_mr	613	57.27	bo_gl	457	39.88
am_ru	1705	59.86	am_zh-cn	1420	57.84	kmr_si	1187	63.38	ky_sk	960	65.46	gu_zh	784	58.73	mr_sh	612	60.58	ne_tl	455	62.25
am_ko	1697	62.72	hr_kn	1416	70.21	pa_pt-br	1186	71.88	en_ht	957	51.36	hy_si	781	67.97	gu_kn	609	65.98	cs_mt	452	42.46
am_id	1695	62.92	mr_ne	1416	60.26	pl_ug	1184	65.24	ba_te	955	66.4	ne_ja	781	58.92	ky_mk	609	63.71	mr_si	452	43.65
am_eu	1692	62.28	km_zh-cn	1412	59.62	arq_ca	1182	38.95	ja_ky	952	62.43	af_fi	779	68.41	sv_tg	608	47.03	ht_mt	451	48.51
hy_ml	1691	60.19	ml_ta	1408	60.68	km_sv	1172	67.79	lt_tg	950	39.34	nb_tg	771	44.53	gu_te	605	65.17	bs_so	451	47.38
eo_uz	1689	53.17	am_uk	1404	62.72	arq_pt	1168	60.84												

Lang Pair	#Sents	COMET-QE	Lang Pair	#Sents	COMET-QE	Lang Pair	#Sents	COMET-QE	Lang Pair	#Sents	COMET-QE	Lang Pair	#Sents	COMET-QE	Lang Pair	#Sents	COMET-QE	Lang Pair	#Sents	COMET-QE
en_mg	411	57.44	id_mt	343	45.08	ga_ru	274	52.87	lb_nl	217	40.42	ug_zh	159	62.63	am_ur	128	66.7	km_sl	94	62.4
km_ky	411	62.79	mn_ug	343	56.64	kn_nb	272	63.34	hu_lb	217	38.76	tl_ug	159	63	ceb_uk	128	48.32	ckb_is	93	61.37
eu_ky	411	59.86	tg_tl	342	38.36	ga_ro	271	50.93	hr_lb	217	39.3	hy_pa	158	60.44	lv_ps	128	61.37	ckb_fil	93	53.04
eo_ky	411	60.48	en_mt	340	64.14	km_sk	271	70.12	he_lb	217	38.13	el_pt	157	55.9	ig_pt	127	38.76	fo_my	93	50.27
am_ky	411	59.28	et_mt	340	42.85	bn_ps	271	64.38	lb_pt-br	217	41.76	cs_it	156	57.62	mg_pt	125	49.39	fr_ig	92	36
ky_la	411	68.26	fr-ca_mt	340	45.25	ar_ga	270	51.87	lb_my	217	49.83	fr_it	156	55.26	am_ms	125	60.65	fo_fr	92	51.41
am_km	411	60.48	da_mt	340	46.18	ga_my	269	53.86	te_uz	217	59.9	fa_it	156	59.29	af_bo	122	36.83	fo_sr	92	48.98
km_ta	411	69.29	ca_mt	340	46.21	kmr_tg	269	44.45	ky_ur	213	62.34	bg_it	156	51.89	mg_ur	121	50.09	ky_zh	92	58.8
eo_tg	410	39.13	gl_ry	340	46.18	am_te	268	65.42	bn_ky	213	63.36	es_it	156	57.47	ms_tt	121	63.59	cs_ig	91	33.64
el_cs	410	43.82	ry_vi	340	43.83	af_my	266	64.68	be_ky	213	55.38	en_it	156	63.54	lo_ro	121	69.79	el_ig	91	40.48
am_sv	409	65.33	cs_ry	340	45.68	fil_ug	266	58.13	ky_ie	213	57.81	th_it	156	55.24	ckb_kn	120	58.58	fr-ca_it	91	49.6
lo_pl	409	64.39	lv_ry	340	43.11	kk_tg	264	44.78	af_az	212	61.65	pl_it	156	57.92	ckb_ig	120	44.92	nn_tl	91	56
lo_sr	409	72.54	lt_ry	340	42.06	gl_km	264	64.79	ga_uk	210	52.77	fr_zh-tw	156	55.94	fi_ig	120	40.05	bo_nl	91	44.72
lo_sq	409	65.53	fr_ry	340	42.75	arg_ug	260	69.16	tr_it	209	55.3	sr_it	156	57.84	ig_th	120	45.33	mt_zh	91	45.93
az_tg	408	46.01	fi_ry	340	45.17	ja_it	260	55.81	tl_ur	209	63.2	sq_it	156	49.28	hy_ig	120	45.74	gu_bt	90	56.37
it_mt	407	50.87	ry_uk	340	46.24	tl_zh-cn	260	50.97	af_fr-ca	208	67.69	sl_it	156	56.32	hr_ig	120	38.77	fo_pt-br	90	51.59
ne_te	407	65.71	fa_ry	340	42.65	is_ne	259	62.19	bn_bo	208	48.14	sk_it	156	57.21	bo_et	120	42.86	is_tg	89	40.67
km_sq	404	61.16	bg_ry	340	41.63	am_gl	259	63.35	af_it	208	60.6	it_it	156	56.32	sk_th	120	29.16	bo_ky	89	48.07
km_mn	404	54.45	ko_ry	340	42	ms_te	258	63.71	ky_pt	207	61.39	ru_it	156	55.68	be_mfe	119	42.25	mfe_uk	87	46.78
nn_sl	403	66.81	es_ry	340	45.8	ne_ps	258	60.89	en_ig	206	45.45	ca_it	156	56.2	ig_mn	119	44.55	mfe_sl	87	38.79
km_lv	403	65.04	en_ry	340	52.71	tl_uk	256	50.56	ar_jg	206	36.14	hy_it	156	56.95	ca_ig	118	38.67	fa_fo	86	50.82
arg_ps	403	54.75	el_ry	340	45.08	mn_pa	252	60.37	srp_uz	206	56.87	nl_it	156	61.72	cs_dz	118	38	fo_zh-tw	86	45.64
ml_so	400	50.8	ry_tr	340	44.06	mn_tl	251	52.44	af_mn	205	62.61	hu_it	156	57.34	dz_ko	118	46.66	ckb_ky	85	55.71
gu_si	399	68.7	az_ry	340	43.89	fil_ug	249	65.46	mt_uk	204	51.74	hr_it	156	58.33	dz_en	118	34.21	ht_sk	84	34.08
fr-ca_mg	397	44.07	ry_th	340	43.91	ga_tl	247	64.21	ga_pt	203	55.8	eb_it	156	60.72	dz_uk	116	45.61	af_fil	84	56.65
eo_ji	396	59.56	ar_ry	340	42.87	mt_sl	247	50.1	ce_vi	203	47.19	he_it	156	57.51	dz_ne	118	49.27	te_zh	84	55.34
gu_is	394	64.31	pl_ry	340	44.55	af_mn	246	60.52	ceb_cs	203	45.08	bn_km	156	62.32	ar_dz	117	38.43	am_srp	83	64.36
es_mt	394	46.83	ry_zh-tw	340	37.97	ht_sv	245	40.18	ceb_lv	203	44.33	bo_ckb	155	45.43	dz_sv	117	45.4	be_tl	83	62.58
eo_ug	392	58.28	ja_ry	340	40.97	arg_bn	245	60.64	ceb_fr	203	45.79	bo_ka	155	42.72	dz_sr	117	46.52	ceb_es	82	74.04
ky_lt	392	61.83	ry_sr	340	50.85	si_sw	244	65.21	ceb_fa	203	49.85	sc_ur	155	53.7	da_dz	117	38.45	ceb_sr	82	75.91
fr-ca_mg	390	46.06	de_ry	340	44.73	fa_ga	243	55.32	bg_ceb	203	43.82	fr_mfe	154	43.14	dz_uk	116	55.61	ht_sl	81	42.05
kmr_tl	386	54.59	ry_sl	340	44.93	bg_ga	243	54.46	ceb_es	203	45.42	kk_ps	154	59.21	dz_ja	116	47.77	mfe_nl	80	46.56
sw_uz	386	56.5	ry_sk	340	49.52	da_fa	243	52.08	ceb_en	203	60.16	ps_zh	154	56.72	dz_he	116	42.82	bg_ig	80	37.05
nb_so	385	51.03	da_ry	340	46.1	de_ga	243	54.13	ceb_th	203	48.41	ca_pa	153	73.6	ceb_gu	115	70.39	ceb_tr	80	44.6
bo_lv	384	43.53	it_ry	340	38.59	ga_nl	243	58.71	ar_ceb	203	43.8	he_ig	152	37.36	ceb_ckb	115	60.5	af_ne	80	66.31
ne_tg	384	50.69	ry_zh-cn	340	35.72	ga_hu	243	54.85	ceb_fr-ca	203	45.62	da_ga	152	70.59	ceb_ko	115	64.51	mik_it	80	53.69
bn_bs	384	67.99	id_ry	340	45.95	ga_hr	243	58.83	ceb_pt	203	44.98	da_ps	151	71.4	ceb_en	115	79.57	lv_it	80	60.86
my_s40	384	60.17	ru_ry	340	40.11	en_mfe	240	53.77	da_uz	203	45.35	arg_uz	150	68.07	zh-tw	115	68.07	ht_it	80	60.45
arg_zh	382	50.92	ro_ry	340	40.86	bs_gu	240	71.66	ceb_ja	203	40.31	ig_ja	150	45.72	ig_pl	114	37.2	tu_uz	80	62.09
mt_tr	381	47.02	nl_ry	340	47.36	am_si	240	64.09	ceb_it	203	48.23	az_si	150	63.08	id_ig	114	37.23	bn_it	80	65.96
de_sc	381	40.71	hu_ry	340	46.23	el_ga	239	54.9	ceb_zh-cn	203	41.58	lt_mfe	149	37.1	pt-br_tt	114	56.5	et_it	80	60.03
arg_fil	380	39.67	hr_ry	340	47.22	my_tg	239	44.19	ceb_it	203	43.73	fa_mfe	149	44.53	bn_pa	114	73.42	az_it	80	63.77
arg_eo	380	39.41	he_ry	340	43.4	et_ug	238	61.05	ceb_tr	203	44.83	bg_mfe	149	38.78	pa_ps	114	65.05	tg_it	80	46.58
fil_mtl	380	62.09	arg_bn	340	48.11	arg_bn	237	38.83	ceb_hu	203	46.77	mfe_tr	149	48.25	af_uz	113	45.37	ht_it	80	63.62
bs_ps	379	66.01	af_zb	339	66.76	arg_nn	237	37.52	ceb_hr	203	45.24	mt_th	149	50.1	ro_it	113	53.92	mfe_ro	79	35.24
az_srp	378	61.13	kk_so	336	48.94	arg_ne	237	48.96	ceb_hi	203	51.02	ar_mfe	149	40.85	dz_ro	113	38.39	hr_mfe	79	37
hr_mg	377	43.62	mk_si	334	68.9	am_nb	235	62.83	ceb_hu	203	47.07	da_mfe	149	37.5	ckb_mt	112	50.58	gu_ig	79	55.23
ps_si	377	69.01	bo_ms	334	41.23	tl_vi	235	50.8	nl_sc	202	41.68	ps_ur	149	65.01	mt_my	112	57.45	fl_mg	79	46.68
inh_ry	377	27.65	ky_nb	333	63.99	bo_kk	235	48.64	ig_ro	201	30.88	ceb_it	149	45.63	da_it	112	62.21	fl_ig	79	45.46
inh_lj	377	28.05	lv_pa	332	63.26	af_et	232	63.26	ast_tl	200	42.32	ceb_ko	149	47.74	dz_vi	112	42.32	bs_mg	79	45.46
inh_lo	377	32.15	ml_te	331	61.03	vt_it	232	59.3	ceb_el	199	47.71	ceb_sr	149	45.58	dz_sl	112	44.42	mg_tl	79	47.87
lo_ne	377	61.77	bo_uz	331	45.42	ast_gl	232	56.35	am_et	199	60.38	ceb_sq	149	39.11	aeb_ar	112	66.52	mg_sl	79	47.35
inh_uz	377	33.12	bo_fil	331	41.75	ast_vi	232	55.05	am_kk	198	59.38	ceb_de	149	44.79	dz_tr	111	46.89	mg_nb	79	49.89
fi_inh	377	30.2	ne_nn	328	70.52	ast_cs	232	55.24	ml_tg	197	44.26	ms_ps	149	63.6	aeb_zh-tw	111	64.53	lb_mk	79	44.67
inh_uk	377	29.9	az_tl	327	56.67	ast_lt	232	53.25	so_ig	197	40.17	mfe_zh-tw	148	48.32	bn_ug	110	68.13	gl_lb	79	36.54
fa_inh	377	31.56	az_ug	326	60.88	ar_tr	232	57.55	sc_sr	197	48.52	mfe_bn	148	38.22	dz_ig	110	43.01	lb_lv	79	40.75
be_inh	377	38.56	az_zb	323	34.63	ast_kk	232	58.02	tg_it	196	36.38	de_it	148	57.12	dz_en	109	41.8	lt_zh	79	39.68
bg_inh	377	25.27	bo_zh	322	37.56	ast_fa	232	55.09	ig_pt-br	195	37.5	mfe_zh-cn	147	43.99	it_mfe	108	40.37	fi_lb	79	40.46
be_inh	377	28.44	eu_te	318	59.75	ast_bn	232	60.32	gu_pa	195	70	kk_kn	147	60.39	ckb_so	108	49.4	kk_lb	79	44.8
fil_inh	377	28.73	hr_sc	318	41.12	ast_bg	232	60.4	af_eo	194	64.08	be_mfe	146	40.06	dz_th	107	47.3	lb_ta	79	48.63
eu_inh	377	30.93	am_mk	317	62.2	ast_es	232	57.94	ckb_ug	194	52.17	eu_mfe	146	40.6	hu_ig	106	40.65	fr-ca_lb	79	36.14
et_inh	377	29.38	ps_sl	317	69.41	ast_en	232	70.13	mt_ro	194	44.74	ig_it	146	39.85	be_mg	106	42.42	lb_sq	79	40.13
eo_inh	377	29.37	lv_pa	313	67.57	ast_tr	232	53.75	ig_tr	194	42.21	ig_zh-cn	146	37.64	mg_uz	106	44.22	fr_sl	79	40.17
inh_tr	377	32.09	arg_mr	312	55.92	ast_it	232	53.79	ast_pt-br	194	60.14	lv_mfe	145	39.58	ar_mg	105	46.8	is_lb	79	38.05
az_inh	377	32.66	pa_sq	311	63.1	ar_ast	232	48.66	sw_tl	194	54.14	bo_pt	145	41.4	gl_mt	105	49.66	ca_lb	79	37.12
inh_th	377	30.35	kk_ky	309	60.42	ast_pl	232	53.55	km_zh	191	53.4	gl_mfe	144	40.11	arg_ckb	104	60.96	lb_ne	79	49.69
arg_inh	377	29.46	ky_mn	309	58.43	ast_zh-tw	232	50.14	eo_ml	191	62.03	ig_sq	143	35.87	mg_my	103	54.91	bn_mg	78	49.17
fr-ca_inh	377	27.67	my_nn																	

Lang Pair	#Sents	COMET-QE	Lang Pair	#Sents	COMET-QE	Lang Pair	#Sents	COMET-QE	Lang Pair	#Sents	COMET-QE	Lang Pair	#Sents	COMET-QE	Lang Pair	#Sents	COMET-QE	Lang Pair	#Sents	COMET-QE
bo_ga	66	39.42	as_uz	57	56.78	ast_co	54	56.27	bi_fr	49	38.21	gl_la	33	46.56	ro_szl	23	38.62	ne_sw	4	77.92
ga_sv	66	62.57	as_fi	57	68.52	eo_oc	54	37.63	bi_li	49	39.17	ga_lv	33	54.87	kmr_mg	22	47.65	gu_hup	3	39.39
ga_sl	66	53.98	ax_uk	57	64.42	en_oc	54	45.12	bi_uk	49	43.11	ga_la	33	46.36	oc_ro	22	30.58	gu_mt	3	56.14
szl_vi	66	37.79	as_bg	57	63.52	oc_tr	54	42.44	bi_fa	49	45.28	ga_kk	33	50.93	lv_th	22	34.58	ckb_hup	3	38.44
cs_szl	66	34.42	as_ko	57	63.2	ast_az	54	53.9	bg_bi	49	33.37	eo_ga	33	53.9	ka_tlh	22	39.07	hup_vi	3	36.74
lt_szl	66	37.71	as_es	57	67.51	az_oc	54	41.61	bi_ko	49	48.18	ga_tl	33	48.45	ig_ta	22	38.14	cs_hup	3	29.79
lo_szl	66	47.98	as_en	57	77.97	oc_th	54	42.46	bi_kk	49	48.99	ga_ta	33	58.61	ha_ro	22	48.38	hup_lv	3	33.43
fr_szl	66	36.82	as_el	57	66.17	ar_oc	54	36.47	bi_et	49	38.02	fr-ca_ga	33	55.72	mfe_sk	21	31.81	hup_it	3	30.8
szl_uz	66	42.59	as_tr	57	55.52	am_ast	54	51.36	bi_es	49	38.23	ca_ga	33	56.03	af_so	21	45.13	fr_hup	3	31.36
fi_szl	66	40.53	as_th	57	57.17	am_oc	54	43.03	bi_en	49	51.24	ga_nn	33	60.46	mfe_mna	20	46.75	hup_uk	3	31.49
szl_uk	66	40.47	ar_as	57	57.84	ast_ta	54	55.77	bi_el	49	41.61	ga_nb	33	60.59	ga_zh	20	42.99	fa_hup	3	32.02
fa_szl	66	41.86	as_la	57	64.43	oc_ta	54	45.54	bi_ka	49	46.65	ga_hi	33	59.97	af_sw	20	69.21	bg_hup	3	33.05
bs_szl	66	39.55	as_pl	57	69.9	ast_fr-ca	54	59.54	bi_tr	49	42.66	la_vi	33	53.59	af_is	18	60.83	hup_ko	3	30.66
bn_szl	66	50.47	as_zh-tw	57	61.33	fr-ca_oc	54	36.55	az_bi	49	41.84	cs_la	33	44.39	am_fil	18	55.87	eu_hup	3	32.84
bg_szl	66	39.14	as_sv	57	70.23	af_ast	54	53.37	bi_th	49	49.53	la_lv	33	47.96	am_tl	18	55.66	eu_mt	3	54.81
fi_szl	66	41.31	as_ja	57	65.09	af_oc	54	35.75	ar_bi	49	37.41	la_li	33	48	af_ig	18	47.77	hup_kk	3	37.83
eu_szl	66	42.46	as_sr	57	72.55	oc_pl	54	37.4	am_bi	49	46.75	fr_la	33	43.81	kk_pa	18	68.09	kk_mt	3	43.48
et_szl	66	41.4	as_sq	57	58.7	oc_zh-tw	54	40.32	bi_fr-ca	49	38.44	fa_la	33	51.61	pt_uk	17	38.63	es_hup	3	30.86
cs_szl	66	38.67	as_de	57	66.92	ast_da	54	59.53	bi_pl	49	33.5	la_uk	33	54.17	pt_it	17	53.84	en_hup	3	29.72
eu_szl	66	39.76	as_sl	57	68.44	ast_oc	54	33.37	bi_zh-tw	49	45.55	ia_la	33	51.22	bi_id	17	36.83	hup_sr	3	33.82
ka_szl	66	41.77	as_kk	57	71.59	ast_ca	54	55.96	bi_sv	49	37.58	bg_la	33	47.54	af_pa	17	69.31	hup_ka	3	31.49
szl_tr	66	42.14	as_da	57	71.77	ast_hy	54	61.22	bi_ja	49	45.92	ko_la	33	46.07	gu_ps	17	79.79	ka_mt	3	50.52
az_szl	66	45.07	as_it	57	71.19	ast_nb	54	57.88	bi_sr	49	34.43	kk_la	33	51.07	ps_uz	17	55.63	hup_tr	3	36.8
szl_th	66	41.9	as_zh-cn	57	58.96	ast_hi	54	58.87	bi_sq	49	29.52	et_la	33	48.72	kk_lo	17	54.79	az_hup	3	38.06
ar_szl	66	38.4	as_id	57	69.72	ja_oc	54	39.1	bi_de	49	35.74	es_la	33	47	hi_ht	16	45.61	az_mt	3	55.83
szl_ta	66	45.4	as_ru	57	67.82	oc_sl	54	41.92	bi_sk	49	31.76	eo_la	33	48.21	oc_pt-br	16	35.98	hup_th	3	32.44
szl_zh-tw	66	39.07	as_ca	57	66.63	da_oc	54	34.6	bi_ji	49	36.19	en_ja	33	55.51	inh_pt	16	25.43	mt_th	3	70.32
de_szl	66	38.3	as_nl	57	71.8	it_oc	54	34.72	bi_zh-cn	49	37.64	la_tr	33	50.01	kk_szl	16	51.57	ar_hup	3	32.26
sl_szl	66	37.61	as_hu	57	63.94	oc_zh-cn	54	36.78	bi_kmr	49	48.81	la_tl	33	49.54	en_szl	16	48.04	hup_pl	3	28.72
da_szl	66	37.86	as_hr	57	71.5	id_oc	54	35.65	bi_ca	49	35.08	la_th	33	53.47	el_szl	16	46.91	hup_zh-tw	3	39.56
it_szl	66	37.53	as_nb	57	71.47	oc_ru	54	39.8	bi_hy	49	51.47	la_ta	33	57.38	aeb_nl	15	73.01	hup_sv	3	33.16
szl_zh-cn	66	36.88	as_nb	57	63.5	ca_oc	54	36.06	bi_nl	49	40.14	fr-ca_la	33	45.14	ig_uk	14	42.13	hup_ja	3	30.64
id_szl	66	38.52	as_pt-br	57	70.51	hy_oc	54	43.45	bi_hu	49	40.15	la_pt	33	49.28	el_la	14	46.06	hup_sr	3	31.32
ru_szl	66	37.62	as_mny	57	57.65	hu_oc	54	37.66	bi_lr	49	35.34	la_zh-tw	33	48.08	ht_uz	14	35.97	hup_sq	3	35.81
ca_szl	66	38.23	af_am	57	60.76	hr_oc	54	37.86	bi_nb	49	35.78	ja_la	33	47.59	bc_ps	14	68.61	mt_sq	3	62.59
hy_szl	66	46.79	te_tl	57	56.02	nb_oc	54	34.22	bi_ji	49	46.17	la_sr	33	54.11	ht_kk	14	46.7	de_hup	3	28.75
nn_szl	66	36.59	mg_zh	56	40.35	hi_oc	54	44.58	bi_he	49	40.82	la_sq	33	50.85	ht_zh	14	35.65	hup_sl	3	28.28
nl_szl	66	40.73	be_szl	56	37.63	he_oc	54	37.05	bi_pt-br	49	37.58	de_la	33	45.85	ht_si	14	42.03	mt_sl	3	71.04
hu_szl	66	42.47	ast_lv	55	53.6	fo_tl	53	40.98	bi_mny	48	54.55	da_la	33	46.43	sv_th	14	32.44	hup_sk	3	32.23
hr_szl	66	39.6	as_hi	54	58.57	inh_nl	51	28.13	inh_lv	48	28.13	li_la	33	46.1	af_te	13	65.93	hup_it	3	29.7
nb_szl	66	38.78	ceb_gl	54	41.93	as_fa	52	58.57	lo_mfe	48	49.16	la_zh-cn	33	49	en_inh	13	33.91	hup_zh-cn	3	29.81
hi_szl	66	47.68	gl_oc	54	34.8	as_mnr	52	54.57	mfe_uz	48	43.31	id_la	33	46.14	pa_sk	12	74.26	hup_kmr	3	36.9
he_szl	66	39.77	oc_vi	54	38.14	fo_pt	52	49.06	fi_mfe	48	40.34	la_ru	33	50.28	ht_lv	11	33.61	kmr_mt	3	58.67
tk_vi	65	42.63	cs_oc	54	31.64	te_ug	52	62.42	bn_mfe	48	49.83	az_pa	33	44.44	az_pa	11	63.17	hup_id	3	36.06
ru_ik	65	38.16	lv_oc	54	34.48	lo_pt	52	63.98	li_mfe	48	39.56	ca_la	33	47.97	pa_so	11	55.94	hup_ru	3	32.72
mg_mnk	65	36.82	ceb_tl	54	41.54	ha_vi	51	50.3	et_mfe	48	37.96	hy_la	33	52.77	arq_te	11	44.83	hup_ro	3	29.96
eo_mg	65	39.07	lt_oc	54	33.55	cs_ha	51	41.48	co_mfe	48	38.64	la_mn	33	50.38	am_arq	11	45.79	hup_mt	3	29.42
mg_ta	65	46.38	fr_oc	54	36.58	fr_ha	51	45.08	az_mfe	48	41.38	la_nl	33	51.85	so_te	11	52.72	hu_hup	3	25.78
ga_kmr	65	59.72	ceb_fi	54	44.43	ha_uk	51	51.56	inh_mfe	48	24.07	hu_la	33	50.29	am_so	11	50.97	hr_hup	3	33.12
es_tk	64	40.68	bn_ceb	54	32.97	fa_ha	51	48.88	ca_mfe	48	38.2	hr_la	33	50.17	bn_ht	11	46.82	hup_nb	3	27.2
it_ik	64	41.45	bn_ceb	54	52.26	bg_ha	51	41.27	hy_mfe	48	46.13	la_ne	33	59.65	ha_pt	11	50.07	ht_hup	3	32.54
dz_pt	64	40.46	be_ceb	54	41.51	ha_ko	51	55.3	mfe_ne	48	53.64	la_nb	33	49.3	as_pt	10	75.72	be_hup	3	36.23
aeb_fa	64	68.26	ceb_et	54	42.89	es_ha	51	46.62	el_inh	47	30.03	bi_la	33	56.9	la_pt	10	45.23	hup_pt-br	3	30.27
fo_vi	63	48.79	ceb_oc	54	46.88	en_ha	51	57.12	fo_id	46	49.58	be_la	33	45.19	dz_pl	10	44.34	hup_mny	3	46.37
nn_zh	63	60.29	az_ceb	54	44.36	el_ha	51	44.37	es_inh	46	28.37	la_pt-br	33	49.05	th_uk	10	42.32	hup_mt	3	31.76
bo_bs	63	43.99	am_ceb	54	45.32	ha_tr	51	44.45	inh_ru	46	25.83	la_mr	33	51.78	ro_uk	10	32.84	mt_nb	3	80.25
bs_km	63	62.47	ceb_ta	54	50.19	ar_ha	51	44.39	aeb_pt	45	76.27	af_ga	32	58.6	af_mfe	10	37.27	ht_mt	3	63.17
bo_km	63	47.71	af_ceb	54	43.24	ha_pl	51	47.55	af_gu	44	72.62	mfe_zh	31	39.96	la_zh	9	46.38	bi_ru	3	29.42
fr-ca_mfe	62	36.97	ast_ceb	54	37.96	ha_zh-tw	51	47.56	ckb_ha	44	46.68	ml_oc	30	33.26	pt_uk	9	34.85	lv_szl	3	55.01
ckb_uk	62	44.28	ceb_sl	54	44.1	ha_ja	51	55.31	ha_kmr	44	53.49	ro_tlh	30	24.69	sr_uk	9	37.66	ja_szl	3	39.65
tk_zh-cn	62	37.35	ceb_da	54	46.48	ha_sr	51	45.55	am_bs	43	65.81	af_lo	29	56.8	ht_ps	9	50.09	pt_szl	3	34.04
dz_el	62	45.12	ceb_oc	54	31.43	ha_sq	51	43.58	pt-br_szl	42	38.8	af_inh	29	31.05	gu_ug	9	66.01	arq_ht	2	40.03
fr_ik	61	38.95	ca_ceb	54	40.77	de_ha	51	43.72	mfe_sq	41	37.92	bs_ug	28	66.44	aeb_it	9	64.44	da_ht	2	41.83
dz_zh-tw	61	42.23	ceb_hy	54	50.01	ha_sk	51	47.48	af_nn	41	64.22	eu_ug	28	58.08	fr_inh	9	31.28	af_bi	2	32.58
dz_zh-cn	61	39.23	ceb_nb	54	46.76	ha_it	51	52.86	af_zh	40	59.43	kk_mg	27	47.26	fr-ca_pa	8	66.03	el_ik	2	45.76
fa_ik	60	44.95	ast_fi	54	56.77	ha_zh-cn	51	44.47	gl_ps	40	64.8	ceb_zh	27	44.84	be_it	8	36.26	ht_mnr	2	42.14
tk_zh-tw	60	41.8	fi_oc	54	36.88	ha_id	51	49.37	mfe_pt	39	44.45	ast_zh	27	44.92	af_ug	8	73.22	aeb_vi	2	55.35
ckb_ga	60	49.08	oc_uk	54	43	ha_ru	51	49.66	nn_te	39	57.31	oc_zh	27	32.03	ckb_sc	8	43.76	ckb_tt	1	88.13
gl_ig	59	40.04	fa_oc	54	41.87	ha_hy	51	51.5	bi_mn	39	44.58	gl_kn	27	66.78	ml_tl	8	51.02	as_ckb	1	47.01
id_ik	59	43.31	be_bo	54	37.13	ha_nl	51	54.74	af_arq	39	40.1									