Understanding Translationese Effects in Multilingual Machine Translation

Anonymous ACL submission

Abstract

This study explores the impact of translationese on multilingual machine translation (MT). Using a newly curated directed "one-way" parallel corpora from Global Voices (MSGV), featuring original texts in diverse languages and explicit anotation of actual translation directions, we evaluated the NLLB and TOWER models on MT tasks between English and five other languages. Our results reveal that translationese inputs are easier to translate into English but not out of English. Additionally, machine translations of translationese are lexically richer than those of original texts when translating into English. These findings suggest that multilingual MT systems experience different translationese effects compared to dedicated bilingual systems, underscoring the need for diverse test beds in MT evaluations. We contribute our dataset to enhance future research.

1 Introduction

Multilingual machine translation (MT) models and large language models (LLMs) have displayed great potential in enhancing global communication across language barriers by scaling MT to many language pairs through transfer learning (Johnson et al., 2017; Arivazhagan et al., 2019; Team et al., 2022) and leveraging multilingual models pre-trained on vast amounts of monolingual data (Alves et al., 2024). For example, mT5 (Xue et al., 2021) is a multilingual variant of T5 model pre-trained on a Common Crawl-based dataset covering 101 languages. Llama 3 (Meta AI, 2024) is an open source LLM with enhanced performance, energy efficiency, and robust safety measures for versatile NLP applications.

However, massively multilingual systems are typically evaluated on the FLORES test bed, created by translation from English into 101 other languages (Goyal et al., 2021). While this enables valuable controlled evaluations across many language pairs, MT from any source language other than English is evaluated on so called “translationese” – inputs that are translations – which is easier to translate by dedicated bilingual MT systems (Toral et al., 2018; Graham et al., 2020).

At the same time, properties of the output text might not be captured by quality ratings alone. For instance, English grammatical structures have been found to influence the fluency of multilingual models in lower resource languages (Papadimitriou et al., 2023). Furthermore, translated language presents distinct features than original texts whether they are written by humans (Volansky et al., 2015) or bilingual machine translation (Vanmassenhove et al., 2021), and that distinguishing original from translated text benefits multilingual MT (Riley et al., 2020).

In this paper, we ask how multilingual MT systems are affected by translationese effects, both in terms of evaluation results and the nature of their outputs. We construct a directed translation evaluation corpus[1] from the Global Voices[2] website, featuring original texts in diverse languages and explicit labeling of translation direction. For example, the Spanish → English corpus in the corpora includes original texts written in Spanish and their corresponding English translations. Unlike FLORES, our test sets are directed "one-way" datasets. For instance, "Spanish → English" and "English → Spanish" are two distinct datasets with distinct contents. We use the corpus to test two hypotheses with the NLLB (Team et al., 2022) and TOWER (Alves et al., 2024) MT systems, on translation between English and five other languages:

H1 Translationese inputs are easier to translate by multilingual MT systems.

H2 The lexical diversity of MT translationese is impacted by translationese inputs.

[1] The dataset will be released upon publication.
[2] https://globalvoices.org/

Our findings suggest that translationese impacts massively multilingual MT and LLMs differently than dedicated bilingual systems.

2 Background

Translated text has been shown to have distinct linguistic features from texts originally written in the same language (Toury, 1979; Baker, 2019). Computational analysis has identified the translationese patterns found in parallel corpora (Volansky et al., 2015) and has made it possible to detect translation direction in parallel text with high accuracy (Baroni and Bernardini, 2006; Kurokawa et al., 2009; Lembersky et al., 2011; Koppel and Ordan, 2011).

The differences between original (O) and translationese (T) texts impact the evaluation of machine translation systems. Suppose a MT system is given a translation task $X \rightarrow Y$. If the parallel test set has original texts in language X and translated texts from X to Y, we say the translation is in *actual direction* (O (original) $\rightarrow$ T (translated)). By contrast, if the parallel test set has original texts in language Y and translated texts from Y to X, we say the translation is in *reverse direction* (T (translated) $\rightarrow$ O (original)). Studies comparing the translation quality obtained with the same system on test sets created in the actual vs. reverse direction have found that MT systems produce better translations in the reverse direction, suggesting that translationese is easier to translate (Toral et al., 2018; Zhang and Toral, 2019; Graham et al., 2020; Läubli et al., 2020). Toral et al. (2018) observed this effect on MT between Chinese and English. Zhang and Toral (2019) revealed that the use of translationese in test sets can result in inflated scores for MT systems through experiments on 17 translation directions, while Graham et al. (2020) studied WMT systems on news translation tasks between English and 9 other languages.

Hence, it is generally recommended to evaluate MT tasks on the actual translation direction ($O \rightarrow T$). However, recent results suggest that actual and reverse test sets capture complementary aspects of translation quality (Freitag et al., 2019), and a causal analysis on Europarl data (Ni et al., 2022) suggests that the inflation of MT scores on the reverse translation direction at test time depends on whether the training and test data directions match.

However, these studies are all based on dedicated statistical or neural systems, often trained for a specific language pair and translation direction. This paper asks whether massively multilingual MT systems and LLM-based MT are impacted by translationese effects. To address this question, we present a "directed" multilingual parallel corpus, including diverse source languages and explicit labeling of actual translation direction, and use it to evaluate recent multilingual MT systems on $O \rightarrow T$ and $T \rightarrow O$ directions.

3 A Directed Parallel Corpora for MT Evaluation

We present Multilingual Source Global Voices (MSGV), a directed parallel corpora for MT Evaluation featuring diverse source languages and explicit labeling of actual translation direction.

Data collection. We draw original texts and their translations from Global Voices, a multilingual platform that features voices from diverse communities and translates these stories into multiple languages. Global Voices provides local perspectives to a global audience, ensuring that the translation direction and MT task align with the intention of writers, who want their articles shared in other languages. Articles are translated by volunteers from the Lingua community[3] through a process ensuring quality control. We initially collected articles from 2016 across all languages before curating a directed parallel corpus for all language pairs between English and one of the following five languages: ES, PT, FR, AR and BN, in both directions.

Sentence alignment and filtering. After crawling document-level aligned original texts and their translations, we segment documents into sentences using NLTK (Bird and Loper, 2004), and run the Vecalign (Thompson and Koehn, 2019, 2020) sentence aligner using LASER embeddings (Artetxe and Schwenk, 2019) to align sentences between the original and translated documents. We further filter out the resulting sentence pairs using a set of rules based on language identification tools, LASER similarity scores, and regular expressions.

Test Sets We constructed 10 test sets by sampling $n = 500$ of data points from the most recently published articles from each of the 10 following parallel corpus: English v.s. (Spanish, Portuguese, French, Arabic, Bengali) in both directions. We select these languages as they are among the highest resource languages with translations on the Global

[3] https://globalvoices.org/lingua/

Voices website, while being spoken by large populations across the globe, and presenting diverse typological properties. For instance, Bengali follows a subject-object-verb order, while English, French, Portuguese and Spanish follow a subject-verb-object oder, and Arabic exhibits both.

Details of the entire data selection and preparation process can be found in Appendix A.

4 Experimental Setup

MT Models. We consider two models in our experiment: (1) NLLB 3.3B (Team et al., 2022), a dedicated MT model trained to translate between any pair of more than 200 languages, including low-resource ones, and (2) TowerInstruct-7B (Alves et al., 2024), a multilingual LLM instruction-tuned for translation related tasks. It was fine-tuned on a wide range of languages. For example, high-quality samples for all language pairs were sampled from OPUS (Tiedemann, 2012), where 744 languages are available in total, and included in the fine-tuning set for TowerInstruct-7B.

Metrics. We evaluate translation quality using (1) COMET (Rei et al., 2020), a state-of-the-art reference-based metric trained to mimic direct assessment scores from human judges, and (2) the NLTK implementation of the chrF metric (Popović, 2015; Bird and Loper, 2004), a character n-gram F-score which has proven to robustly correlate with human judgments in many languages.

5 Results

Each system translates from $X \rightarrow Y$ (where one of X and Y is English, and the other is selected from ES, PT, FR, AR, and BN) in actual ($O \rightarrow T$) and reverse ($T \rightarrow O$) directions. We first discuss the impact of translationese data on evaluation (Section 5.1), before analyzing the properties of MT translationese in multilingual systems (Section 5.2)

5.1 Impact of Translating Translationese

The COMET and chrF for all models and evaluation settings are plotted in Figure 1.[4] We reported both metrics as they follow similar trends.

When translating into English, both models exhibit a statistically significant advantage in the reverse $T \rightarrow O$ direction compared to the $O \rightarrow T$ direction. The paired t-test was used to evaluate the significance of these differences, with p-values less than 0.05 indicating strong evidence against the null hypothesis. For NLLB, the $T \rightarrow O$ direction significantly outperforms the $O \rightarrow T$ direction across both evaluation metrics in all 5 comparisons ($p < 0.05$). Similarly, for TOWER, the $T \rightarrow O$ direction significantly outperforms the $O \rightarrow T$ direction in 4 out of 5 comparisons ($p < 0.05$). This is consistent with translationese effects observed in prior work with older MT models.

However, this trend surprisingly does not hold when translating out of English. For NLLB, "$O \rightarrow T$" beats "$T \rightarrow O$" on both metrics for 4 out of 5 times ($p < 0.05$), while it is 3 out of 5 times for TOWER ($p < 0.05$), suggesting that translating original English text is easier than translating English translationese. We hypothesize that the make-up of the training data of these multilingual systems eliminates the expected translationese effect for English, in line with Ni et al. (2022)'s finding that the inflation of scores in the reverse direction is influenced by the direction of the training data with bilingual Transformer models. While the complete make-up of their (pre-)training data is unknown, Tower/LLaMA-2 have been exposed to vast amounts of original monolingual English text, while NLLB training data included a seed corpus curated by translating English sources into other languages (Team et al., 2022), and the majority of the parallel text can be assumed to have one English side.

In summary, our results suggest that hypothesis H1 holds true only for translation into English, but not for translation out of English when utilizing multilingual MT or LLM systems.

5.2 Linguistic Diversity of Translationese

We turn to assessing the linguistic diversity of machine translationese, compared to that of our various human-written test sets. Following Vanmassenhove et al. (2021), to measure the repetitiveness of vocabulary, we use Yule's I (Yule, 1944)

$$I = \frac{\sum_{i=1}^{N} i^2 \cdot V_i - N}{N^2}$$

where N is the total number of words in the text. V_i is the number of vocabulary items (types) that occur exactly i times in the text. Figure 2 summarizes the Yule's I scores.[5]

[4]Raw scores can be found in Appendix Table 3.

[5]We also measured the Shannon Entropy (Shannon, 2001) of word surface forms given lemma to measure grammatical diversity as manifested in morphology, but did not find any patterns of grammatical diversity with the languages and trans-

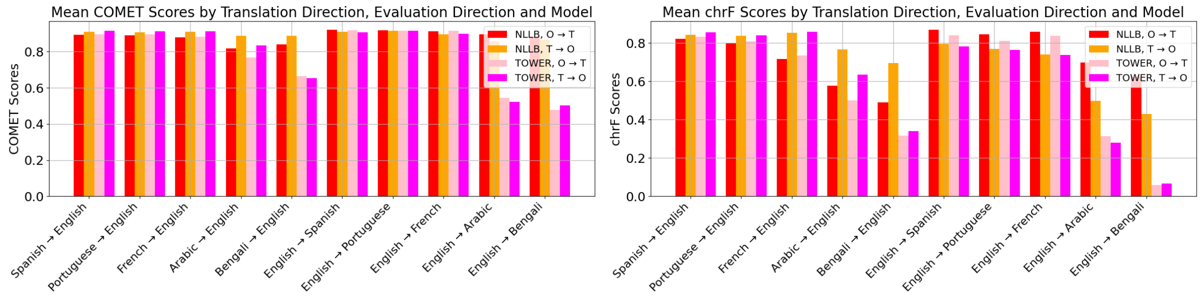

Figure 1: MT evaluation results for NLLB-3.3B and TowerInstruct-7B on 10 translation directions in both $O \rightarrow T$ and $T \rightarrow O$ settings.

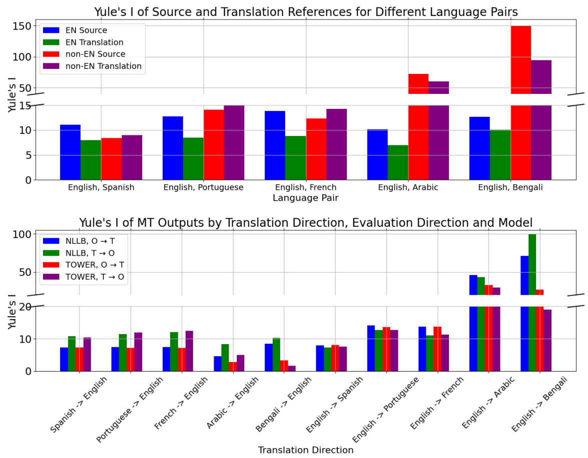

Figure 2: Yule's I score of source and translation references for difference language pairs and Yule's I score of MT outputs on 10 translation directions in both $O \rightarrow T$ and $T \rightarrow O$ settings for NLLB-3.3B and TowerInstruct-7B.

For English, original texts always have a higher Yule's I score than translated text, which indicates that original texts are lexically richer than translations, as expected. However, this may not hold true for non-English languages. Similar to MT evaluation, linguistic diversity of MT outputs displays different trends when translating into and out of English. When translating into English, the $T \rightarrow O$ outputs yield a higher Yule's I score than the corresponding $O \rightarrow T$ evaluation 5 out of 5 times for NLLB, and 4 out of 5 times for TOWER, suggesting that machine translations of human translationese are more lexically diverse than machine translations of original text. When translating out of English, it is quite the opposite, with $O \rightarrow T$ outputs yielding a higher Yule's I score than the corresponding $T \rightarrow O$ evaluation 4 out of 5 times for NLLB, and 5 out of 5 times for TOWER.

In sum, these results suggest that H2 holds: the lexical diversity of MT translationese is impacted by translationese inputs.

6 Conclusion

We curated a multilingual parallel corpora from Global Voices, which explicitly labels the translation direction. Using test sets extracted from the corpora, we evaluated NLLB-3.3B and TowerInstruct-7B on 10 translation directions in both actual $O \rightarrow T$ and reverse $T \rightarrow O$ settings. We found that $T \rightarrow O$ evaluation inflates MT performance when translating into English, while opposite trend can be observed when translating out of English. Additionally, we measured the linguistic diversity of source, target references and the MT outputs. We found that English original texts are lexically richer than translationese, and that evaluation in the reverse $T \rightarrow O$ inflates the lexical diversity of MT outputs compared to the actual direction when translating into English.

These results show that massively multilingual MT and LLMs do not suffer from the exact same translationese effects as dedicated bilingual systems. Translationese is easier to translate for these systems when it is in non-English languages, suggesting that the FLORES test bed artificially amplifies MT quality for translation out of non-English languages. Lexical diversity analysis suggests that machine translating translationese gives artificially more diverse outputs when translating into English.

These findings motivate the use of more diverse test beds when evaluating multilingual machine translation, including text originally written in non-English languages. To that end, we release the test sets used in this paper along with all the parallel data extracted from Global Voices with translation direction annotation.

lation directions studied. All scores for the human-written data and MT outputs can be found in Appendix Tables 4 and 5 respectively.

7 Limitations

Despite the findings, this study has several limitations that should be considered.

First, the number of languages involved in the experiment is limited. Besides English, only five languages are included: Spanish, Portuguese, French, Arabic and Bengali. This restriction may affect the generalizability of the results to a broader range of languages present in global translation.

Second, the translation direction in this study always involves English. It is unknown whether the trends observed in this study still hold for translation between non-English languages. The limitations mentioned above are largely due to the lack of non-English data, particularly original texts. For example, Malagasy is a linguistically distinct, low-resource language that we were interested in including in our experiment at first due to its high availability on the Global Voices website. However, we ultimately had to drop it because nearly all the Malagasy texts available are translations, not original texts. The discrepancy in data availability among different languages is still significant, even on a multilingual citizen media website like Global Voices.

Third, the number of models evaluated in this study is relatively small, as only two models, NLLB and TOWER, were included. This limitation can impact the comprehensiveness of the findings. Future research may explore whether these trends are applicable to a broader range of models.

References

Duarte M. Alves, José Pombal, Nuno M. Guerreiro, Pedro H. Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, José G. C. de Souza, and André F. T. Martins. 2024. Tower: An open multilingual large language model for translation-related tasks. *Preprint*, arXiv:2402.17733.

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, Melvin Johnson, Maxim Krikun, Mia Xu Chen, Yuan Cao, George Foster, Colin Cherry, Wolfgang Macherey, Zhifeng Chen, and Yonghui Wu. 2019. Massively Multilingual Neural Machine Translation in the Wild: Findings and Challenges. *Preprint*, arxiv:1907.05019.

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.

Mona Baker. 2019. Corpus linguistics and translation studies*: Implications and applications. In *Researching translation in the age of technology and global conflict*, pages 9–24. Routledge.

Marco Baroni and Silvia Bernardini. 2006. A New Approach to the Study of Translationese: Machine-learning the Difference between Original and Translated Text. *Literary and Linguistic Computing*, 21(3):259–274.

Steven Bird and Edward Loper. 2004. NLTK: The natural language toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 214–217, Barcelona, Spain. Association for Computational Linguistics.

Markus Freitag, Isaac Caswell, and Scott Roy. 2019. APE at Scale and Its Implications on MT Evaluation Biases. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 34–44, Florence, Italy. Association for Computational Linguistics.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzman, and Angela Fan. 2021. The flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Preprint*, arXiv:2106.03193.

Yvette Graham, Barry Haddow, and Philipp Koehn. 2020. Statistical Power and Translationese in Machine Translation Evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 72–81, Online. Association for Computational Linguistics.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Moshe Koppel and Noam Ordan. 2011. Translationese and Its Dialects. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1318–1326, Portland, Oregon, USA. Association for Computational Linguistics.

David Kurokawa, Cyril Goutte, and Pierre Isabelle. 2009. Automatic Detection of Translated Text and its Impact on Machine Translation. In *Proceedings of Machine Translation Summit XII: Papers*, Ottawa, Canada.

Samuel Läubli, Sheila Castilho, Graham Neubig, Rico Sennrich, Qinlan Shen, and Antonio Toral. 2020. A Set of Recommendations for Assessing Human–Machine Parity in Language Translation. *Journal of Artificial Intelligence Research*, 67:653–672–653–672.

Gennadi Lembersky, Noam Ordan, and Shuly Wintner. 2011. Language Models for Machine Translation: Original vs. Translated Texts. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 363–374, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Meta AI. 2024. Introducing llama 3: Advancing state-of-the-art language models. `https://ai.meta.com/blog/meta-llama-3/`. Accessed: 2024-06-15.

Jingwei Ni, Zhijing Jin, Markus Freitag, Mrinmaya Sachan, and Bernhard Schölkopf. 2022. Original or Translated? A Causal Analysis of the Impact of Translationese on Machine Translation Performance. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5303–5320, Seattle, United States. Association for Computational Linguistics.

Isabel Papadimitriou, Kezia Lopez, and Dan Jurafsky. 2023. Multilingual BERT has an accent: Evaluating English influences on fluency in multilingual models. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1194–1200, Dubrovnik, Croatia. Association for Computational Linguistics.

Maja Popović. 2015. chrf: character n-gram f-score for automatic mt evaluation. In *Proceedings of the tenth workshop on statistical machine translation*, pages 392–395.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. Comet: A neural framework for mt evaluation. *Preprint*, arXiv:2009.09025.

Parker Riley, Isaac Caswell, Markus Freitag, and David Grangier. 2020. Translationese as a Language in “Multilingual” NMT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7737–7746, Online. Association for Computational Linguistics.

Claude Elwood Shannon. 2001. A mathematical theory of communication. *ACM SIGMOBILE mobile computing and communications review*, 5(1):3–55.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. *Preprint*, arXiv:2207.04672. [link].

Brian Thompson and Philipp Koehn. 2019. Vecalign: Improved sentence alignment in linear time and space. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1342–1348, Hong Kong, China. Association for Computational Linguistics.

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

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Lrec*, volume 2012, pages 2214–2218. Citeseer.

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

Gideon Toury. 1979. Interlanguage and its Manifestations in Translation. *Meta: Journal des traducteurs*, 24(2):223.

Eva Vanmassenhove, Dimitar Shterionov, and Matthew Gwilliam. 2021. Machine Translationese: Effects of Algorithmic Bias on Linguistic Complexity in Machine Translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2203–2213, Online. Association for Computational Linguistics.

Vered Volansky, Noam Ordan, and Shuly Wintner. 2015. On the features of translationese. *Digital Scholarship in the Humanities*, 30(1):98–118.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

G. Udnv Yule. 1944. *The Statistical Study of Literary Vocabulary*. Cambridge University Press.

Mike Zhang and Antonio Toral. 2019. The effect of translationese in machine translation test sets. *Preprint*, arXiv:1906.08069.

A Corpus construction

A.1 Data Collection

Using Scrapy[6], we crawled the HTML files of over 100k source articles and their corresponding translations from Global Voices[7], spanning the years from 2004 to 2024. We then parsed the HTML files and extracted the main content into plain text. We discarded articles published before 2016, keeping only those from 2016 onwards for the following reasons: (1) Recent articles are preferred over older ones. (2) Articles from 2016 onwards display reduced English-dominance. (3) Articles from 2016 onwards includes more diverse languages. Table 1 gives an overview of data statistics before 2016 and from 2016 onwards to illustrate these points. Figure 3 shows the language distribution in source articles from 2016 to 2024. Figure 4 shows the language distribution in all (source and translation) articles from 2016 to 2024. While a significant percentage of translation articles are written in non-English, non-English source articles still remain relatively low-resource compared to the vast number of English source articles.

	Before 2016	2016 and later
% of non-English articles (source)	0.28%	11.92%
% of non-English articles (source+translation)	10.77%	81.17%
# of languages in articles (source)	11	15
# of languages in articles (source+translation)	44	50
% of articles in top 5 high-resource languages (source)	99.98%	99.17%
% of articles in top 5 high-resource languages (source+translation)	96.80%	62.46%

Table 1: Statistics of data before 2016 and from 2016 onwards, respectively.

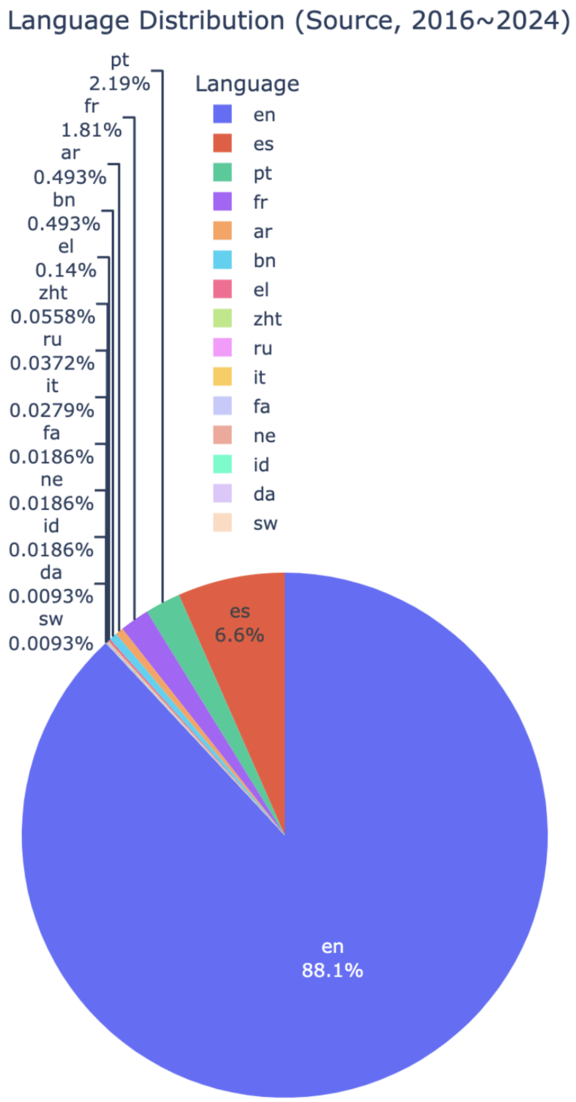

Figure 3: Language Distribution of Source Articles from 2016 to 2024.

A.2 Sentence Alignment.

After initial data collection, we first tokenized each article into individual sentences using the NLTK (Bird and Loper, 2004) package. Then, we used LASER embedding (Artetxe and Schwenk, 2019), a language-agnostic embedding framework, to encode multilingual sentences into shared-space vectors. After this, we proceeded to apply Vecalign (Thompson and Koehn, 2019, 2020) to evaluate the similarity among multilingual sentence embeddings. Sentences with analogous meanings are aligned together due to their closeness in the vector space. In the end, we curated a directed parallel corpus for all language pairs between English and one of the following five languages: Spanish, Portuguese, French, Arabic and Bengali, in both direction. We explicitly retained the temporal information by associating the publishing date of each article with each of its constituent sentences after segmenting the article.

[6] https://scrapy.org/

[7] Unless otherwise stated, all content created by Global Voices is published under a Creative Commons Attribution-Only license. This means that anyone, anywhere has the right to share — copy and redistribute the material in any medium or format.

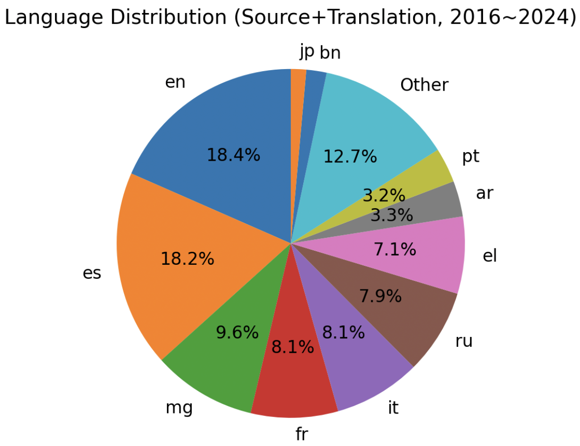

Figure 4: Language Distribution of All Articles (Source+Translation) from 2016 to 2024.

Source → Translation	# of Sentences
English → Spanish	239,709
Spanish → English	8,438
English → Portuguese	37,733
Portuguese → English	3,344
English → French	92,583
French → English	2,047
English → Arabic	25,314
Arabic → English	773
English → Bengali	4,615
Bengali → English	500

Table 2: Number of sentences in the refined parallel corpus between English and each of the next five high-resource languages in both translation directions when the threshold of LASER cosine similarity is set to 0.85 (for Bengali → English, the threshold was set to 0.63; extensive manual processing was involved to ensure data quality).

A.3 Data Filtering and Quality Assurance.

We filtered out noisy samples to ensure high data quality. First, for each Source → Translation parallel corpus, we used the langdetect[8] package to to detect and filter out language-text mismatches. Second, we used python regex[9] to filter out texts containing unwanted patterns, including emojis, certain special symbols, etc. Third, we applied LASER embedding (Artetxe and Schwenk, 2019) to encode each pair of aligned sentences and compute the cosine similarity between them. Data points with low cosine similarities are filtered out. Empirically we found out that 0.85 is a good threshold to ensure both enough amount of data and good data quality. Fourth, if needed, other heuristics like texts length, or manual processing may be involved. After data filtering, Table 2 presents the amount of data in the parallel corpus between English and each of the next five high-resource languages ('Spanish', 'Portuguese', 'French', 'Arabic', 'Bengali') in source texts in both translation directions when the threshold of LASER cosine similarity is set to 0.85, i.e. all samples with LASER cosine similarity below 0.85 are not included (the threshold was set to 0.63 for Bengali → English, due to limited data and a lower cosine similarity distribution; extensive manual processing was involved to ensure data quality).

A.4 Test Sets Sampling

For better data quality, we mostly sample data points from a subset of the parallel corpus where a LASER cosine similarity threshold must be met. The thresholds are 0.98, 0.95, 0.97, 0.95, 0.98, 0.94, 0.95, 0.85 and 0.86 for EN → ES, ES → EN, EN → PT, PT → EN, EN → FR, FR → EN, EN → AR, AR → EN and EN → BN, respectively. We want the threshold to be as high as possible for better quality, but ensure that there is enough amount of data points left. Due to a preference of more recent data, we select the most recent 500 samples from all the X → English. We recorded the time range of selected samples from X → English and randomly selected 500 samples in with corresponding English → X corpus within the same time range. This had never been a problem, as English → X always has a much larger size than X → English given the same language X. Hence, when passing in a time frame where X → English contains 500 samples, the corresponding English → X always has a pool containing more than 500 data points to sample from under this time frame.

[8] https://github.com/Mimino666/langdetect?tab=readme-ov-file

[9] https://docs.python.org/3/library/re.html, https://github.com/mrabarnett/mrab-regex

B MT Evaluation Results

NLLB-3.3B

Translation Direction	COMET (↑)			chrF (↑)		
	O → T	T → O	p_value	O → T	T → O	p_value
English → Spanish	**0.921**±0.041	0.91±0.042	7.305e-05	**0.869**±0.083	0.795±0.106	2.974e-32
Spanish → English	0.892±0.053	**0.91**±0.042	3.307e-09	0.821±0.107	**0.843**±0.109	1.388e-03
English → Portuguese	**0.918**±0.047	0.914±0.045	1.126e-01	**0.845**±0.099	0.769±0.119	1.323e-26
Portuguese → English	0.89±0.049	**0.908**±0.042	2.211e-10	0.798±0.113	**0.836**±0.099	1.834e-08
English → French	**0.912**±0.059	0.896±0.052	1.314e-05	**0.859**±0.123	0.74±0.114	3.271e-50
French → English	0.88±0.054	**0.909**±0.051	4.438e-17	0.717±0.118	**0.852**±0.105	3.264e-69
English → Arabic	**0.895**±0.066	0.857±0.059	2.607e-21	**0.698**±0.145	0.497±0.132	2.651e-93
Arabic → English	0.818±0.071	**0.888**±0.047	4.690e-66	0.576±0.144	**0.767**±0.109	6.413e-99
English → Bengali	**0.882**±0.046	0.865±0.072	1.143e-05	**0.624**±0.116	0.431±0.157	9.395e-89
Bengali → English	0.84±0.074	**0.886**±0.041	3.982e-32	0.489±0.169	**0.696**±0.107	1.584e-95

TowerInstruct-7B

Translation Direction	COMET (↑)			chrF (↑)		
	O → T	T → O	p_value	O → T	T → O	p_value
English → Spanish	**0.918**±0.041	0.908±0.046	1.887e-04	**0.841**±0.099	0.781±0.11	1.259e-18
Spanish → English	0.896±0.047	**0.914**±0.04	4.570e-10	0.832±0.097	**0.856**±0.095	7.627e-05
English → Portuguese	0.915±0.04	**0.916**±0.037	7.319e-01	**0.81**±0.099	0.763±0.11	1.993e-12
Portuguese → English	0.896±0.043	**0.912**±0.037	3.614e-10	0.809±0.102	**0.84**±0.095	8.012e-07
English → French	**0.914**±0.052	0.898±0.052	8.452e-07	**0.838**±0.11	0.736±0.112	1.041e-43
French → English	0.883±0.054	**0.912**±0.053	7.294e-18	0.735±0.11	**0.859**±0.114	1.563e-59
English → Arabic	**0.545**±0.147	0.521±0.113	3.956e-03	**0.315**±0.129	0.279±0.1	8.347e-07
Arabic → English	0.768±0.082	**0.834**±0.068	9.003e-41	0.5±0.125	**0.636**±0.124	9.245e-59
English → Bengali	0.477±0.117	**0.503**±0.125	7.921e-04	0.06±0.085	**0.066**±0.1	3.363e-01
Bengali → English	**0.664**±0.116	0.652±0.122	1.243e-01	0.316±0.117	**0.341**±0.1	3.232e-04

Table 3: COMET Score and chrF Score of NLLB-3.3B and TowerInstruct-7B evaluated on 10 translation directions in both $O \rightarrow T$ and $T \rightarrow O$ settings. Both scores are reported in $0 \sim 1$ scale.

C Linguistic Diversity

Language Pair		Yule's I (↑)		Shannon Entropy (↑)	
		Source	Translation	Source	Translation
English, Spanish	English	**11.107**	7.955	0.09	**0.101**
	Spanish	8.444	**9.023**	**0.142**	0.128
English, Portuguese	English	**12.755**	8.522	0.096	**0.101**
	Portuguese	14.071	**15.651**	0.131	**0.138**
English, French	English	**13.862**	8.83	**0.081**	0.077
	French	12.357	**14.299**	0.118	0.118
English, Arabic	English	**10.198**	6.966	0.111	**0.127**
	Arabic	**72.282**	60.095	**0.367**	0.332
English, Bengali	English	**12.701**	10.094	0.09	**0.1**
	Bengali	**149.768**	94.325	0.234	**0.254**

Table 4: . Linguistic diversity of source references and target references for each language pair. Within block language pair (X, Y) or (Y, X), grid (X, Source) denotes the linguistic diversity of the source side of test set $X \rightarrow Y$, while (X, Translation) denotes the linguistic diversity of the target side of test set $Y \rightarrow X$. Yule's I scores by 10,000 for ease of readability. Shannon Entropy is reported in $0 \sim 1$ scale.

Translation Direction	Yule's I (↑)				Shannon Entropy (↑)			
	NLLB		TOWER		NLLB		TOWER	
	O → T	T → O	O → T	T → O	O → T	T → O	O → T	T → O
English → Spanish	**7.92**	7.314	**8.105**	7.533	0.13	**0.145**	0.129	**0.148**
Spanish → English	7.243	**10.716**	7.294	**10.416**	**0.099**	0.088	**0.103**	0.093
English → Portuguese	**14.071**	12.711	**13.634**	12.667	**0.135**	0.128	**0.134**	0.132
Portuguese → English	7.455	**11.381**	7.214	**11.988**	**0.097**	0.096	**0.098**	0.095
English → French	**13.735**	11.066	**13.769**	11.344	0.12	0.12	0.12	**0.121**
French → English	7.388	**12.046**	7.192	**12.491**	0.075	**0.081**	0.077	**0.08**
English → Arabic	**46.284**	43.1	**32.854**	29.717	0.331	**0.352**	0.29	**0.35**
Arabic → English	4.537	**8.345**	2.824	**4.95**	**0.126**	0.115	**0.12**	0.11
English → Bengali	71.073	**99.916**	**27.181**	19.07	**0.23**	0.214	0.123	**0.124**
Bengali → English	8.46	**10.214**	**3.297**	1.655	**0.095**	0.09	**0.059**	0.053

Table 5: Yule's I Score and Shannon Entropy of the MT outputs of NLLB-3.3B and TowerInstruct-7B evaluated on 10 translation directions in both $O \rightarrow T$ and $T \rightarrow O$ settings. For ease of readability and comparison, we multiplied Yule's I scores by 10,000. Shannon Entropy is reported in $0 \sim 1$ scale. (For English → Bengali translation, TowerInstruct-7B output texts that are not in Bengali frequently. Therefore, the diversity of English → Bengali MT outputs by TowerInstruct-7B was calculated only based on outputs in Bengali, i.e. after all the non-Bengali MT outputs were removed.)