Detecting Languages Unintelligible to Multilingual Models through Local Structure Probes

Louis Clouâtre[1,3] **Prasanna Parthasarathi**[2] **Amal Zouaq**[1] **and Sarath Chandar** [1,3,4]
[1] Polytechnique Montréal
[2] Huawei Noah's Ark Lab, Canada
[3] Quebec Artificial Intelligence Institute (Mila)
[4] Canada CIFAR AI Chair

Abstract

Providing better language tools for low-resource and endangered languages is imperative for equitable growth. Recent progress with massively multilingual pretrained models has proven surprisingly effective at performing zero-shot transfer to a wide variety of languages. However, this transfer is not universal, with many languages not currently understood by multilingual approaches. It is estimated that only 72 languages possess a "small set of labeled datasets" on which we could test a model's performance, the vast majority of languages not having the resources available to simply evaluate performances on. In this work, we attempt to clarify which languages *do* and *do not* currently benefit from such transfer. To that end, we develop a general approach that requires only unlabelled text to detect which languages are not well understood by a cross-lingual model. Our approach is derived from the hypothesis that if a model's understanding is insensitive to perturbations to text in a language, it is likely to have a limited understanding of that language. We construct a cross-lingual sentence similarity task to evaluate our approach empirically on 350, primarily low-resource, languages.

1 Introduction

Natural Language Processing (NLP) boasts of significant recent successes, largely driven by the introduction of different flavors of pretrained models (Devlin et al., 2019; Lan et al., 2019; Liu et al., 2019; Radford and Narasimhan, 2018; Radford et al., 2019; Brown et al., 2020). However, the rewards of those successes have been mostly reaped by high-resource languages. The existence of high-quality benchmarks and metrics, the abundance of readily available high-quality corpora, or the number of researchers speaking the language themselves (Blasi et al., 2022) are significant contributors to the disproportionate advances in high-resource languages. Although recent improvements in NLP have been shown to extend to several different languages, such as the progress to language understanding by BERT-style models (Cui et al., 2019; Le et al., 2019; Martin et al., 2019; Antoun et al., 2020; Carmo et al., 2020; de Vries et al., 2019; Malmsten et al., 2020; Polignano et al., 2019; Nguyen and Tuan Nguyen, 2020), many of those extensions have been limited to relatively high-resource languages. Such improvements are often perceived to extend to low-resource languages, but the lack of appropriate benchmarking in those languages curtails our ability to verify such perceptions.

The World Atlas of Language Structures (Haspelmath et al., 2014; Dryer and Haspelmath, 2013) categorizes over 2600 languages, and Ethnologue (Eberhard et al., 2022) estimates that there are currently over 7000 living languages (Hammarström, 2015); the most popular cross-lingual benchmarks (Liang et al., 2020; Hu et al., 2020) together cover less than 50 languages and Joshi et al. (2020) estimates that only 72 languages worldwide pass the threshold of having "a small set of labeled datasets", which could be used for evaluation. **Towards contributing to an equitable society with the development of language technologies, it is imperative that we ensure that no living languages are left behind.** Building automatic and cheap tools to provide better visibility into which languages are not currently well understood by cross-lingual NLP models then becomes essential.

To determine cheaply if a model understands text in a specific language or not, we first find behaviors that are consistently exhibited by models that do perform well on language understanding tasks. By finding when those behaviors are not exhibited, we can determine whether a model understands the text or not. Recent research trends have taken to evaluating

Findings of the Association for Computational Linguistics: EMNLP 2022, pages 5375–5396

well-known natural language understanding (NLU) models on perturbed text (Sinha et al., 2020, 2021; Pham et al., 2021; Gupta et al., 2021; O'Connor and Andreas, 2021; Taktasheva et al., 2021; Clouatre et al., 2022). Such works attempt to distill which aspects of a text are not necessary and which aspects are necessary for language models to understand it by selectively perturbing the text, such as by shuffling the order of words. It may be possible to use the sensitivity of models to perturbations to properties of text that are found to be essential for NLU as a proxy for model understanding. As an extreme example, if a model develops the same understanding of a text and the same text with its characters shuffled, it can be hypothesized that its original understanding of the text was limited. The texts "I will eat an apple" and "ln i plla wat Ieaep" contain the same characters. Yet, we would expect radically different representations of both from a model, assuming it correctly understood the unperturbed version.

We explore the following research questions and verify their corresponding hypotheses:

- **RQ1: Is there an aspect of text that is universally used by language models that perform well on understanding tasks?**

  H1: We hypothesize that the **local structure** (Clouatre et al., 2022) of text is one such aspect and that the performance of cross-lingual models on language understanding tasks in most languages should be highly sensitive to local structure perturbations. To verify this hypothesis, we use several cross-lingual tasks from popular benchmarks (Hu et al., 2020; Liang et al., 2020), on which we perform different local structure-altering perturbations and evaluate several cross-lingual models on said perturbed text.

- **RQ2: If there is such a universally relied upon aspect of text, can a model's performance sensitivity to that aspect be used as a proxy for understanding?**

  H2: We hypothesize that if such an aspect of text exists, a model that is *in*sensitive to perturbations to that aspect may be inferred to have a limited understanding of the original text. To verify this hypothesis, we construct a large-scale cross-lingual sentence representation task covering 350 languages which we use to measure the language understanding of several models in all the 350 languages. For each model and target language, we measure the sensitivity of perturbations to that aspect. We demonstrate that all languages for which our cross-lingual models are less sensitive to perturbations to the local structure of text are also not well understood by those models.

Our main contributions are:

- Across all tested languages, tasks, and models, we find that performance is directly correlated with the amount of local perturbations applied to the text.
- We develop the **monolingual local sensitivity** metric which measures the reliance of a model on the local structure to build text representations, only requiring unlabelled monolingual data.
- On a task covering 350 languages, we find that languages on which a model has low monolingual local sensitivity always has a poor representation of that language's text.

2 Related Work

Cross-Lingual Performance Prediction Predicting to what extent cross-lingual models' performances transfer to different languages and tasks has seen a fair amount of interest (Birch et al., 2008; Xia et al., 2020; Lauscher et al., 2020; Dolicki and Spanakis, 2021; Srinivasan et al., 2021; Ye et al., 2021; Ahuja et al., 2022). These works formulate the zero-shot transfer to different languages and tasks as regression problems. Linguistic features and model-specific features such as the size of the pretraining data and the models' performance in different languages and tasks serve as input. The performance of the model on a certain type of task and language is then used as the target of the regression.

All those approaches share a few limitations. They are evaluated on high-resource to medium-resource languages, as those languages all possess supervised learning datasets to be evaluated upon and generally rely upon linguistic features from the World Atlas of Language Structures (Haspelmath et al., 2014; Dryer and Haspelmath, 2013) which cover less than half of all estimated living languages (Hammarström, 2015). While

Languages	German	Chinese	Spanish	Turkish	Vietnamese	Arabic	Russian	Hindi
Appearances	20	20	19	18	17	17	17	16
Native Speaker (Millions)	95	1300	493	80	76	400	150	260
Languages	French	Greek	Thai	Bulgarian	Japanese	Korean	Indonesian	Italian
Appearances	16	16	15	13	12	12	12	12
Native Speaker (Millions) (Eberhard et al., 2022)	77	13	28	8	128	80	43	67

Table 1: Statistics on languages making the most appearances in the cited cross-lingual performance prediction work.

those approaches are tremendously valuable for optimizing transfer learning, they provide limited utility in predicting the performance of a model on a very low-resource language.

All cited studies (Birch et al., 2008; Xia et al., 2020; Lauscher et al., 2020; Dolicki and Spanakis, 2021; Srinivasan et al., 2021; Ye et al., 2021; Ahuja et al., 2022) predicting cross-lingual performance cover a total of 75 languages. It may seem like a large selection, but we observe that high-resource languages dominate these works. By counting the frequencies of appearances of every language used in those works, we find that most of the evaluations were made on some of the world's highest resource languages, in terms of native speakers, as illustrated in Table 1. Taking an average of the number of native speakers in all languages surveyed, weighted by their appearances in the cited literature, we observe that evaluations were made on languages with, on average, 127 million native speakers.

Text Perturbations and Structure Probing Several text perturbation schemes have been explored in the context of probing model performances. Sankar et al. (2019) shuffles and reverses utterances and words in a generative dialogue setting, highlighting insensitivity to the order of conversational history. Pham et al. (2021) shuffles n-grams for different values of n , highlighting the insensitivity of pretrained Transformer models. Sinha et al. (2020) performs perturbations on the position of the words on textual entailment tasks, with the added criterion that all words' positions must have changed. Taktasheva et al. (2021) extend perturbation studies to Swedish and Russian and performs perturbations by shuffling syntactic phrases, rotating sub-trees around the root of the syntactic tree of a sentence, or simply shuffling the words of the text.

These approaches work well to provide insight into many languages with automatic parsing tools or well-developed tokenizers. However, low-resource languages cannot be assumed to possess those automatic linguistic tools that permit grammatical perturbations. Language-agnostic tools and measures will need to be prioritized to evaluate the importance of the different aspects of text in low-resource languages. Priors regarding the form of the text, such as the presence of white-space delimited words, will have to be kept to a minimum.

Clouatre et al. (2022) proposes a suite of controllable perturbations on characters, which should be compatible with almost any written language, as well as a metric quantifying perturbations to the *local* structure that measures perturbations on a character-level. The findings of Clouatre et al. (2022) in regards to the ubiquitous nature of **local sensitivity** as it relates to language understanding and the compatibility of both the metric and perturbations with any text make their work particularly well suited to a massively multilingual setting.

Canine and General Tokenization Some of the language scripts used in this work, such as Inuktitut Syllabics, are not covered by the tokenization scheme of most pretrained cross-lingual models such as XLM-R (Lample and Conneau, 2019) and multilingual-BERT (Devlin et al., 2019), which rely on a learned vocabulary of subwords (Sennrich et al., 2015; Wu et al., 2016). The Canine model (Clark et al., 2021) offers a tokenization scheme that covers every Unicode character, allowing it to have representations for scripts that were not part of the pretraining dataset. This permits us to evaluate low-resource languages in previously unseen scripts that would otherwise have to be ignored. Has evidence exists that transfer can occur even in languages written in different scripts (Pires et al., 2019), the use of universal tokenization will be necessary to evaluate cross-lingual transfer properly. Canine also uses character-level tokenization instead of explicitly modeling subwords, which should be more resilient to perturbations to the order of

characters and control for the confounder of vocabulary destruction.

Cross-Lingual Sentence Similarity Cross-lingual sentence retrieval tasks, such as Tatoeba (Artetxe and Schwenk, 2018), rely on the presence of language-agnostic sentence embeddings. By comparing the cosine distance between the embeddings of a text in a target language with the same text in English or another high-resource language, we can obtain a relative idea of the quality of the representations of said target language when compared to its understanding of English. As models evaluated on English NLU tasks obtain, at times, super-human performances (Wang et al., 2019b,a), a model having a similar representation to English sentences in a low-resource language would imply at least *some* level of understanding of that text. Cross-lingual sentence retrieval is also particularly interesting as, compared to other NLU tasks, obtaining a broad coverage of languages is relatively simple.

3 Multilingual Local Sensitivity

To answer **RQ1**, we borrow some of the perturbation schemes and metrics from Clouatre et al. (2022) and apply them to a multilingual setting. We aim to demonstrate empirically that neural models generally make some use of local structure to perform understanding tasks, irrespective of language. This can be demonstrated by progressively removing local structure from text through order altering perturbations and observing a similar decline in understanding (as measured by performance metrics) of that text from models. Such results will motivate using low local structure sensitivity as a proxy for lack of ability to perform language understanding tasks. We perform those experiments on seven popular cross-lingual tasks covering 44 unique languages.

3.1 Metric and Perturbations

The **CHRF-2** (chrF) (Popović, 2015) metric measures the amount of character bi-gram overlap between a perturbed text and the original text and is used to represent the amount of local structure that has not been perturbed in a text.

We perform perturbations by altering the order of **characters** present in the text. This is done by using the **neighbor flipping** (Clouatre et al., 2022) perturbations, which, with a controllable probability ρ, flips a character with its neighbor, thus providing an arbitrary amount of local perturbations. This perturbation is illustrated in 1. [1]

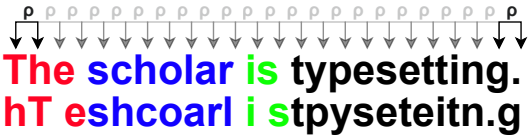


Figure 1: From top to bottom: Unperturbed Text, Neighbor Flipping with $\rho = 0.5$

3.2 Experimental Details

All experiments are conducted with the pretrained cross-lingual models Canine-S (Clark et al., 2021), XLM-RoBERTa-base (XLM-R) (Lample and Conneau, 2019) and multilingual-BERT-base-cased (mBERT) (Devlin et al., 2019).

A total of 7 cross-lingual tasks selected from the most popular cross-lingual benchmarks (Hu et al., 2020; Liang et al., 2020) covering 44 languages are used for evaluation (see Table 2). [2]

Task	*n* Languages	Task Type	Metric
PAWS-X	7	Paraphrase Detection	ACC
XNLI	15	NLI	ACC
QAM	3	Text Classification	ACC
QADSM	3	Text Classification	ACC
WPR	7	Page Ranking	nDCG
BUCC	5	Sentence Retrieval	F1
Tatoeba	33	Sentence Retrieval	ACC

Table 2: Summary information of the different tasks used.

The zero-shot cross-lingual setting (Hu et al., 2020) is used for all experiments, meaning that the cross-lingual model is finetuned on the English version of the dataset and evaluated without further tuning on all target languages. [3]

No finetuning is performed on the cross-lingual sentence retrieval tasks, defaulting to simple cosine similarity of the mean of the final hidden representations of the model for every input token, as described in Hu et al. (2020).

The English version on which the model is finetuned is kept unperturbed, while the target

[1] Pseudocode of the perturbation is present in the Appendix D

[2] Extractive tasks such as extractive QA are not compatible with our perturbations, as the answer would also be perturbed and were not considered.

[3] Detailed training and testing hyperparameters and process are present in the Appendix A.

language text on which the model is evaluated goes through several perturbations. We perform a total of 12 different perturbations on every task and language and obtain their performance, thus evaluating the sensitivity of the target languages to the perturbations. [4] All models are finetuned on five different random seeds, and all perturbations are performed on five different random seeds, for a total of 25 evaluations for every model on every task, every language present in the tasks, and every perturbation setting.

3.3 Results and Discussion

We observe that, in an aggregate, local structure perturbations almost perfectly correlate with the degradation of the ability of a model to perform language understanding tasks in a cross-lingual setting. A Pearson's r of 0.99 is found between our measure of the perturbations and the performance obtained. We call this correlation between degradation in performance and the amount of local perturbation the **local sensitivity**. Figure 2 shows the results averaged across all tasks, all random seeds, and all languages. We can observe an almost perfect linear relationship between the amount of local structure remaining in the text on which a model is evaluated, as measured by the character bigram F-score, and the average score of our models when evaluated on that text.

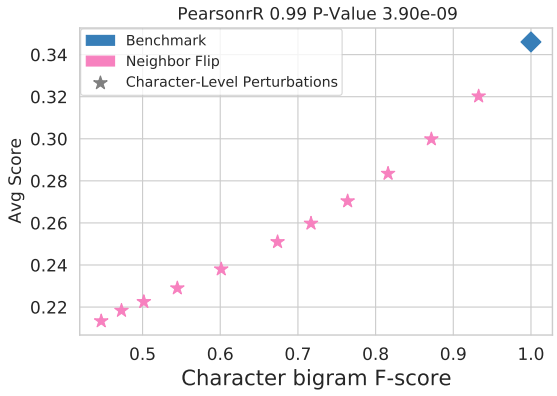


Figure 2: Plotted is the relation between local structure perturbations and average performance on all tested datasets and languages, averaged across all models. The local sensitivity, measured by the correlation between local perturbations and performance degradations, is reported at the top of the figure.

The local sensitivities of the different models on the various tasks are also very consistent,

[4] Details perturbations used are present in the Appendix A

Figure 3: Local sensitivity matrix of the different models tested on the various tasks averaged across all random seeds. The higher the value, the more sensitive a model's performance is to perturbations to local structure.

performances being either perfectly or highly correlated to the amount of local structure remaining, as pictured in Figure 3. This is consistent across all models, including the tokenization-free Canine, which lets us control for the vocabulary destruction brought by perturbing the order of characters.

Finally, we can observe whether or not languages with lower local sensitivity tend to underperform their locally sensitive counterparts. In Figure 4, we observe that while high local sensitivity does not guarantee good performance, none of the languages that posses low local sensitivity do much better then chance on the task of Natural Language Inference. Those results are consistent across all tasks and present in Appendix B.

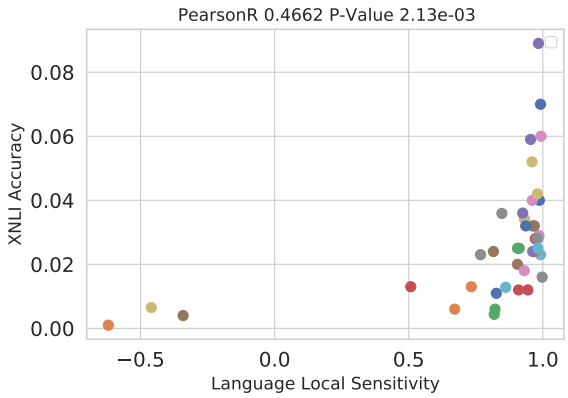


Figure 4: Plotted are the individual language's local sensitivity plotted against their performance on the unperturbed text on the XNLI task, averaged across all models.

From our results, we cannot find a dimension in which a model's performance is not extremely sensitive to local structure perturbations, lending credence that local structure is an aspect of text that is always, at least, relied upon to

perform understanding tasks. Those results support **H1**, demonstrating that it is likely that language models universally make some use of local structure to perform understanding tasks, irrespective of language, task, or the specific pretrained model. Further, in our tested tasks, languages on which a model has low local sensitivity tend to underperform those with high local sensitivity.

4 Low Monolingual Local Sensitivity as a Proxy for Lack of Understanding

This section explores using an insensitivity to local perturbations as a proxy for lack of understanding to address **RQ2**. To find a proxy for understanding that will provide greater visibility in the performance of very low-resource languages where evaluation is not possible, we cannot measure the local sensitivity by evaluating a language on a labeled task. Therefore, we will explore **monolingual local sensitivity** as a proxy for lack of language understanding, with unlabelled monolingual data in the target language as its only requirement.

4.1 Monolingual Local Sensitivity

We previously defined local sensitivity as the correlation between the degradation of performance of a model on a task and the local perturbations applied to its text. To calculate the local sensitivity of a model on a specific task, we evaluate the model's performance on that task with all 12 of our perturbations and calculate the Pearson's r between the performance on the perturbed text and the local structure as measured by CHRF-2. However, this process has the limitation of requiring a labeled dataset on which to evaluate performance.

To obtain a measure of local sensitivity while bypassing the requirement for a supervised learning dataset, we turn to the **monolingual local sensitivity**. First, we build a corpus in the target language containing 1000 unique texts. We then formulate the problem as a sentence similarity between two copies of the same corpus, initially resulting in a perfect similarity between sentence pairs. We apply our perturbations to one copy of the corpus while keeping the other copy unperturbed. As more of the local structure is destroyed, the representation of the different pieces of text should also drift apart, assuming that the model considers local structure. We can then obtain a measure of local sensitivity based on the task of sentence retrieval between the same corpus, one of which is perturbed. A toy example comparing cross-lingual sentence similarity and monolingual sentence similarity is pictured in Figure 8.

4.2 MTData Sentence Retrieval

We first require a simple task covering many low-resource languages to evaluate low monolingual local sensitivity as an indicator of lack of understanding in a meaningful way. From the MTData (Gowda et al., 2021) dataset, which is composed of millions of sentence pairs between English and over 500 target languages, we build an English-to-language cross-lingual sentence similarity task covering 350 different language-to-English pairs containing 1000 text pairs per language. The dataset is built using the same process and filtering as was used to construct the Tatoeba cross-lingual sentence similarity dataset (Artetxe and Schwenk, 2018). [5] We will use the normalized cosine similarity between the sentence representation and its target representation to evaluate performance. We normalize by removing the mean and scaling it by the standard deviation of cosine similarity between the text and all other potential texts. This evaluation metric should control for the different models' behaviors, the different quality of corpora for the different languages, and the diversity of examples for every language, making comparisons more uniform than a simple cosine distance and less sparse than an absolute hit rate. Under this scoring system, a score of 1.0 would mean that the representation of a text with its counterpart would be 1.0 standard deviation closer than its distance to all other texts, as measured by the cosine distance. We will refer to this metric as the **similarity Z-Score**.

4.3 Results and Discussion

From our MTData cross-lingual sentence similarity task, we can obtain and compare two measures for the 350 languages.

The first is the model's performance on the task of cross-lingual similarity between an English representation, which is assumed to be of reasonable quality as the model performs well on English understanding tasks, and a target language

[5] Specific details, dataset statistics, and evaluation methods are expanded upon in the Appendix C.

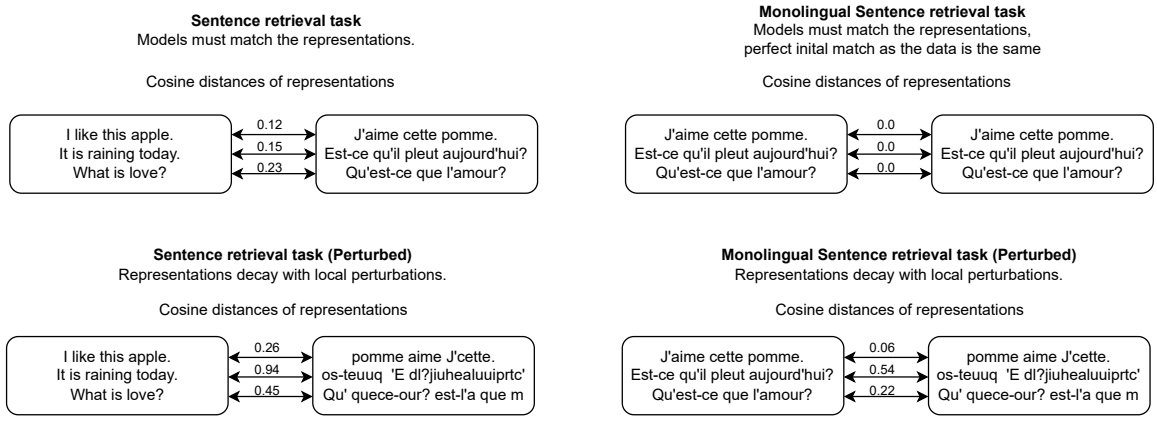


Figure 5: Toy example of sentence similarity and monolingual sentence similarity with and without perturbations.

representation, as measured by the similarity Z-Score. The closer the representation target language's text to its English representation, the closer the abilities of the model to represent that language are to the ability of the model to represent English.

The second is the monolingual local sensitivity, which we obtain by performing sentence retrieval using two copies of the target language side of the MTData cross-lingual sentence retrieval dataset, as illustrated in the left side of Figure 8.

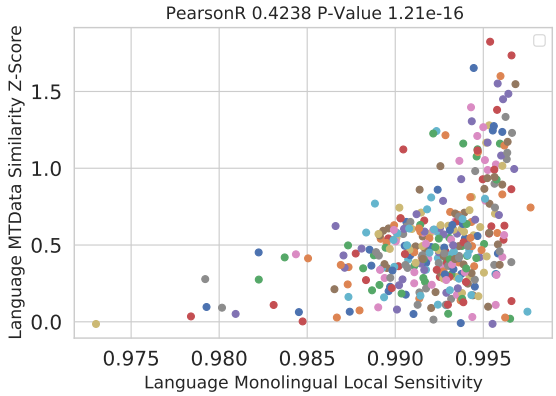


Figure 6: Scatter plot of all 350 languages comparing their degree of monolingual local sensitivity and their similarity Z-Score on the MTData cross-lingual sentence similarity task. Left to right is less sensitive to more sensitive to monolingual local perturbations. The Pearson's r between a language's monolingual local sensitivity and its cross-lingual similarity Z-Score is reported.

We compare the performance of our pretrained models against the monolingual local sensitivity of all 350 tested languages, pictured in Figure 6. Languages with high local sensitivity may often have poor unperturbed performance, meaning that relying on local structure does not imply good language understanding. The opposite, however, seems broadly true. Specifically, languages with low monolingual local sensitivity have universally poor unperturbed performance. To build representations roughly in line with the quality of an English representation, a model must rely, at least somewhat, on that text's local structure.

All languages that obtained a monolingual local sensitivity of under 0.99 did not have representations that were very close to their English counterparts. Assuming normality, a similarity Z-Score of 0.8 implies that over 21% of representations outputted by the model for that language were closer to the representation of the English counterpart than the target text pair. None of the languages with monolingual local sensitivity under 0.99 clear that hurdle. Surprisingly, from the 350 languages surveyed, only a few could truly be said not to be understood by the models. The probability of having an average score of even 0.10 on this task through a random process is vanishingly small, and only 23 of the 350 languages do not cross that threshold, an encouraging result for the current multilingual pretraining approaches.

To provide greater context on those results, we have plotted all 350 surveyed languages on their estimated geographical centers (Haspelmath et al., 2014; Dryer and Haspelmath, 2013), in Figure 12. We can observe several languages that both underperform, as indicated by the color, and are predicted to underperform, as indicated by the size of the circle. Further analysis of our results is provided in the Appendix B.

From our results, we find that a low monolingual

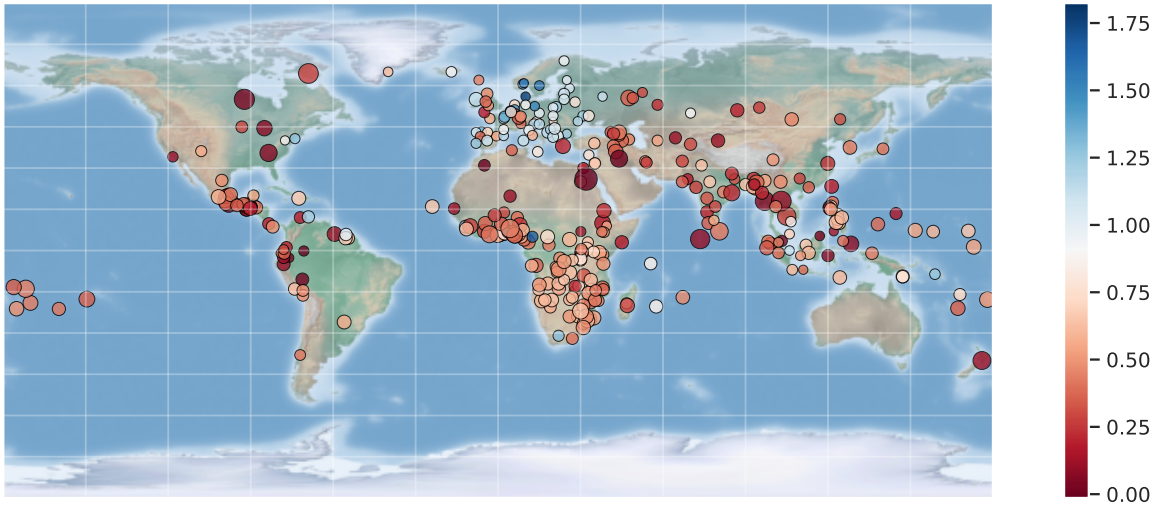


Figure 7: All 350 languages are plotted on their estimated geographical centers. The color of the dots are scaled by the cross-lingual similarity Z-Score while their size are scaled by how low the monolingual local sensitivity is for that language. Large red dots represent languages that were both has poor performance and low monolingual local sensitivity. Small blue dots represent languages that both had good performance and high monolingual local sensitivity.

sensitivity is indicative of a limited ability to represent text. Those results support **H2**, demonstrating that it is likely that a model's inability to properly represent a certain language can be detected through monolingual local structure probes.

5 Limitations and Ethical Considerations

There are several limitations to our approach that may have an ethical impact.

The first one is the poor recall. While languages with low monolingual local sensitivity, from our experiments, always are poor performers, many poor performers are also sensitive to local perturbations. Our approach can successfully find some languages that do require additional attention but will miss many other languages. If we rely on automatic tools to detect where to put efforts, there is a possibility that no efforts are put on languages that are not detected by those tools.

The second one is the data requirement. Obtaining a sufficient sample of text to calculate monolingual local sensitivity for some low-resource languages may still be too high of a hurdle. Some living languages, especially those from oral traditions, may have a limited pool of written text available.

It is crucial that if we use automatic tools to detect which languages requires further efforts, we do not forget of the languages that might not be detected or are incompatible with those tools.

6 Conclusion

Regardless of the language, task, or model used, the use of local structure seems to be relied upon by neural models to build an understanding of text. Local structure sensitivity does not seem to be an artifact of the English language and broadly applies to written text in most languages.

We explore monolingual local sensitivity to automatically detect unintelligible languages to cross-lingual models, the only requirement being access to monolingual unlabelled text. If local structure is essential to building understanding, not relying on the local structure would imply a limited understanding. We demonstrate a high correlation between monolingual local sensitivity and the ability of a model to perform cross-lingual sentence similarity in 350 diverse languages. Specifically, all languages with low monolingual local sensitivity performed poorly on that task. Those results indicate that with the measure of monolingual local sensitivity alone, it is possible to estimate the performance of a certain language on a certain model without access to any supervised learning datasets.

Our contribution will be useful to direct further efforts, such as unlabelled data gathering for pretraining, to expand the coverage of cross-lingual models in the most efficient way.

Acknowledgements

This research has been funded by the NSERC Discovery Grant Program.

References

Kabir Ahuja, Shanu Kumar, Sandipan Dandapat, and Monojit Choudhury. 2022. Multi task learning for zero shot performance prediction of multilingual models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5454–5467, Dublin, Ireland. Association for Computational Linguistics.

Wissam Antoun, Fady Baly, and Hazem M. Hajj. 2020. Arabert: Transformer-based model for arabic language understanding. *ArXiv*, abs/2003.00104.

Mikel Artetxe and Holger Schwenk. 2018. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *CoRR*, abs/1812.10464.

Alexandra Birch, Miles Osborne, and Philipp Koehn. 2008. Predicting success in machine translation. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 745–754, Honolulu, Hawaii. Association for Computational Linguistics.

Damian Blasi, Antonios Anastasopoulos, and Graham Neubig. 2022. Systematic inequalities in language technology performance across the world's languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5486–5505, Dublin, Ireland. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Diedre Carmo, Marcos Piau, Israel Campiotti, Rodrigo Nogueira, and Roberto de Alencar Lotufo. 2020. PTT5: pretraining and validating the T5 model on brazilian portuguese data. *CoRR*, abs/2008.09144.

Jonathan H. Clark, Dan Garrette, Iulia Turc, and John Wieting. 2021. Canine: Pre-training an efficient tokenization-free encoder for language representation.

Louis Clouatre, Prasanna Parthasarathi, Amal Zouaq, and Sarath Chandar. 2022. Local structure matters most: Perturbation study in NLU. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3712–3731, Dublin, Ireland. Association for Computational Linguistics.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Ziqing Yang, Shijin Wang, and Guoping Hu. 2019. Pre-training with whole word masking for chinese BERT. *CoRR*, abs/1906.08101.

Wietse de Vries, Andreas van Cranenburgh, Arianna Bisazza, Tommaso Caselli, Gertjan van Noord, and Malvina Nissim. 2019. Bertje: A dutch BERT model. *CoRR*, abs/1912.09582.

J. Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*.

Blazej Dolicki and Gerasimos Spanakis. 2021. Analysing the impact of linguistic features on cross-lingual transfer. *CoRR*, abs/2105.05975.

Matthew S. Dryer and Martin Haspelmath, editors. 2013. *WALS Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig. 2022. *Ethnologue: Languages of the World*, twenty-fifth edition. SIL International.

Thamme Gowda, Zhao Zhang, Chris A Mattmann, and Jonathan May. 2021. Many-to-english machine translation tools, data, and pretrained models.

Ashim Gupta, Giorgi Kvernadze, and Vivek Srikumar. 2021. Bert & family eat word salad: Experiments with text understanding. *arXiv preprint arXiv:2101.03453*.

Harald Hammarström. 2015. Ethnologue 16/17/18th editions: A comprehensive review. *Language*, 91:723 – 737.

Martin Haspelmath, Hans-Jörg Bibiko, and Claudia Schmidt. 2014. *The World Atlas of Language Structures*. Oxford University Press.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *CoRR*, abs/1901.07291.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. ALBERT: A lite BERT for self-supervised learning of language representations. *CoRR*, abs/1909.11942.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From zero to hero: On the limitations of zero-shot language transfer with multilingual Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4483–4499, Online. Association for Computational Linguistics.

Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoît Crabbé, Laurent Besacier, and Didier Schwab. 2019. Flaubert: Unsupervised language model pre-training for french. *CoRR*, abs/1912.05372.

Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, Xiaodong Fan, Bruce Zhang, Rahul Agrawal, Edward Cui, Sining Wei, Taroon Bharti, Ying Qiao, Jiun-Hung Chen, Winnie Wu, Shuguang Liu, Fan Yang, Rangan Majumder, and Ming Zhou. 2020. XGLUE: A new benchmark dataset for cross-lingual pre-training, understanding and generation. *CoRR*, abs/2004.01401.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

Martin Malmsten, Love Börjeson, and Chris Haffenden. 2020. Playing with words at the national library of sweden - making a swedish BERT. *CoRR*, abs/2007.01658.

Louis Martin, Benjamin Müller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Villemonte de la Clergerie, Djamé Seddah, and Benoît Sagot. 2019. Camembert: a tasty french language model. *CoRR*, abs/1911.03894.

Dat Quoc Nguyen and Anh Tuan Nguyen. 2020. PhoBERT: Pre-trained language models for Vietnamese. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1037–1042, Online. Association for Computational Linguistics.

Joe O'Connor and Jacob Andreas. 2021. What context features can transformer language models use? In *ACL/IJCNLP*.

Thang M. Pham, Trung Bui, Long Mai, and Anh M Nguyen. 2021. Out of order: How important is the sequential order of words in a sentence in natural language understanding tasks? *ArXiv*, abs/2012.15180.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Marco Polignano, Pierpaolo Basile, Marco Degemmis, Giovanni Semeraro, and Valerio Basile. 2019. Alberto: Italian bert language understanding model for nlp challenging tasks based on tweets. In *CLiC-it*.

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, Lisbon, Portugal. Association for Computational Linguistics.

Alec Radford and Karthik Narasimhan. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Chinnadhurai Sankar, Sandeep Subramanian, Chris Pal, Sarath Chandar, and Yoshua Bengio. 2019. Do neural dialog systems use the conversation history effectively? an empirical study. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 32–37, Florence, Italy. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. *CoRR*, abs/1508.07909.

Koustuv Sinha, Robin Jia, Dieuwke Hupkes, Joelle Pineau, Adina Williams, and Douwe Kiela. 2021. Masked language modeling and the distributional hypothesis: Order word matters pre-training for little. *arXiv preprint arXiv:2104.06644*.

Koustuv Sinha, Prasanna Parthasarathi, Joelle Pineau, and Adina Williams. 2020. Unnatural language inference. *arXiv preprint arXiv:2101.00010*.

Anirudh Srinivasan, Sunayana Sitaram, Tanuja Ganu, Sandipan Dandapat, Kalika Bali, and Monojit Choudhury. 2021. Predicting the performance of multilingual NLP models. *CoRR*, abs/2110.08875.

Ekaterina Taktasheva, Vladislav Mikhailov, and Ekaterina Artemova. 2021. Shaking syntactic trees on the sesame street: Multilingual probing with controllable perturbations. *CoRR*, abs/2109.14017.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019a. Superglue: A stickier benchmark for general-purpose language understanding systems. *CoRR*, abs/1905.00537.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019b. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144.

Mengzhou Xia, Antonios Anastasopoulos, Ruochen Xu, Yiming Yang, and Graham Neubig. 2020. Predicting performance for natural language processing tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8625–8646, Online. Association for Computational Linguistics.

Zihuiwen Ye, Pengfei Liu, Jinlan Fu, and Graham Neubig. 2021. Towards more fine-grained and reliable NLP performance prediction. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3703–3714, Online. Association for Computational Linguistics.

A Experiment Details

Model Hyperparameters and Training We finetune each pretrained models on the English version of each dataset for a total of 10 epochs, checkpointing the model after each epochs. The English version is never perturbed, the finetuning is done on unperturbed data. This finetuning is done 5 times with different random seeds for each model and each datasets. For 7 datasets and 3 models we have a total of $3 * 7 * 5 = 105$ finetuning and 1050 checkpoints, one for each epoch. A learning rate of 2e-5, a batch size of 32 and a weight decay of 0.01 is used in all finetuning. All experiments used a warmup ratio of 0.06, as described in Liu et al. (2019). A maximum sequence length of 512 for the mBERT and XLM-R model and a maximum sequence length of 2048 for the Canine model are used.

For the evaluation, we perform the same perturbations on the validation and testing data of the different target languages. We evaluate the perturbed validation data on each of the 10 checkpoints, chose the best checkpoint on the perturbed validation data, and evaluate that checkpoint on the perturbed test data. This process is repeated for each perturbations, each of the 5 random seed and 5 times with different perturbation random seeds for each finetuned models. In total, for each language in each task on each model for each perturbation setup we average results over 25 random seeds.

For the sentence retrieval tasks, such as Tatoeba and BUCC, we do not perform any finetuning. We obtain the representation by averaging the output of the final hidden layer of the model. (Hu et al., 2020) First, we obtain the representation of the unperturbed English side of the dataset. This is done by feeding the English text through the model and averaging the final layers hidden representation of the text. We then perform our perturbations on the target language text, feed those perturbed text through the same pretrained cross-lingual model and obtain it's representation through the same process. We now have a set of English representation and a set of target language representation, on which we can obtain the cosine distances. We can either find the nearest neighbours (Tatoeba, BUCC) or use the Z-Score of those representations (MTData). If the nearest neighbour is the sentence that was to be retrieved, we consider this an hit, else it is a miss. The reported results are over the average of 5 random seeds of those perturbations.

Monolingual Local Sensitivity The monolingual sentence retrieval task is performed in the exact same process as for the sentence retrieval task described in Appendix A. The only difference is that the unperturbed English text is replaced by the target language corpus. Pictured in Figure 8 is a toy example representing the monolingual sentence retrieval tasks compared to the crosslingual one. We calculate monolingual local sensitivity by taking the correlation of the degradation in performance on the monolingual sentence retrieval task with the amount of local perturbations applied to the right side of the dataset.

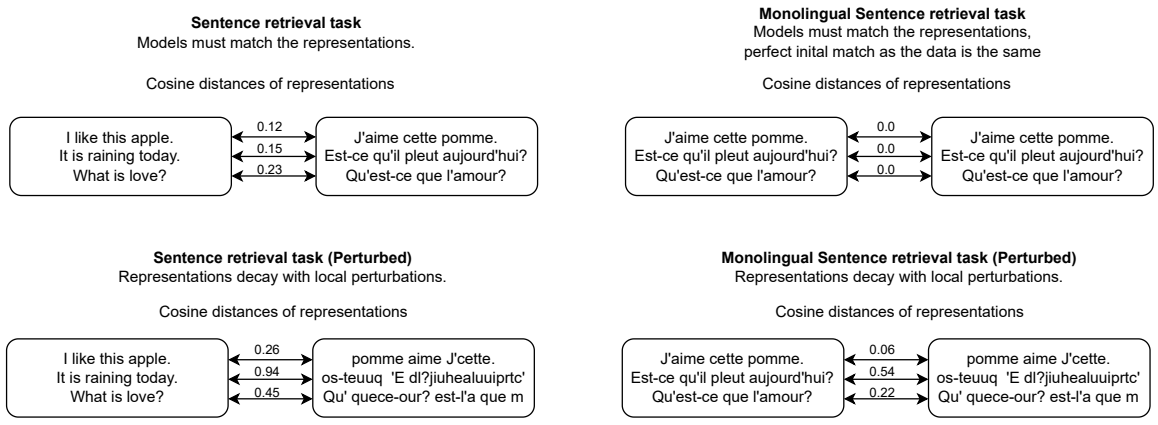


Figure 8: Toy example of sentence retrieval and monolingual sentence retrieval with and without perturbations.

Perturbations A total of 13 evaluations, containing 12 perturbations are used for all experiments. The first one is the Benchmark, which is simply the unperturbed text. On a character-level perturbations we

perform neighbour-flip shuffling with ρ values of: [0.025, 0.05, 0.075, 0.1, 0.125, 0.15, 0.175, 0.2, 0.25, 0.3, 0.35, 0.45]. No neighbor-flip with ρ over 0.5 or over are performed, as they would ultimately shuffle the text *less*. Unlike Clouatre et al. (2022), we focus purely on local structure perturbations, as we are not interested in the relative importance of local structure compared to other structures, but simply that local structure is important at all.

B Additional Results

Cross-Lingual Local Sensitivity Additional Results In this section we present additional results on the first set of experiments on the cross-lingual zero-shot local sensitivity tasks.

The trend of extremely high correlation between performance and perturbations also holds when grouping results by script and language family, as shown in Figure 9 and Figure 10.

Figure 9: Local sensitivity matrix between the different languages families with at least 3 tested languages in our tasks, averaged across all tasks and models.

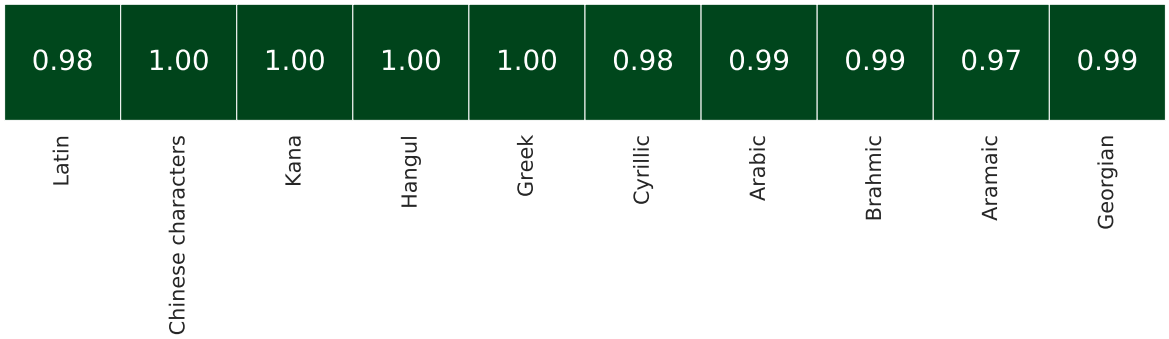
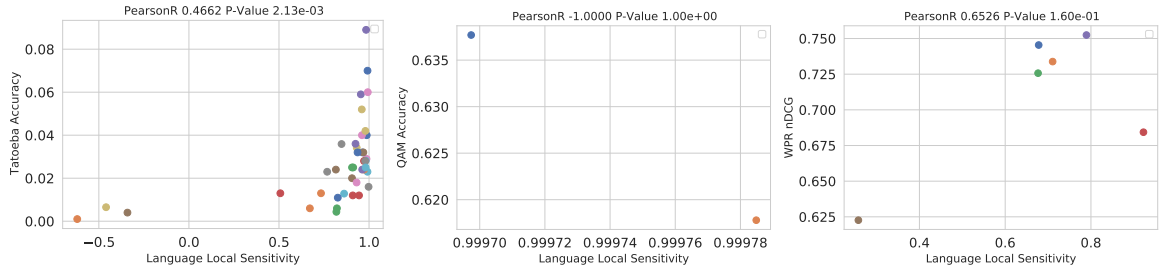
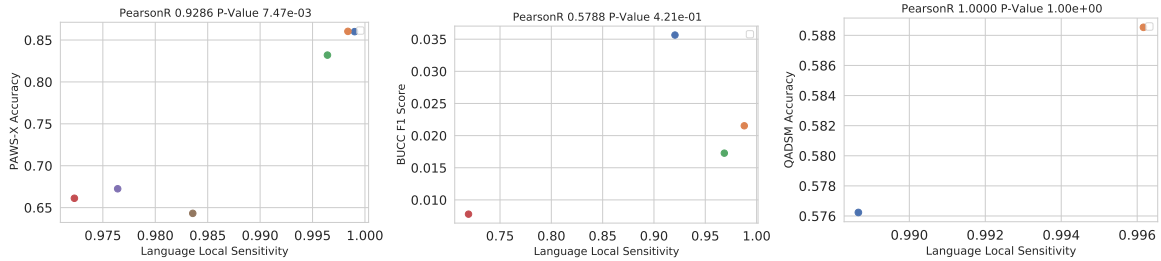


Figure 10: Local sensitivity matrix between the different scripts with at least 3 tested languages in our tasks, averaged across all tasks and models.

Further, using low local sensitivity to predict low performance on a particular language seem to be consistent across tested tasks, as seen in Figure 11.



(a) Tatoeba, QAM, WPR



(b) Paws-X, BUCC, QADSM

Figure 11: Plotted are the individual language's local sensitivity plotted against their performance on the unperturbed text on all tasks, averaged across all models. We can observe that with the exception of QAM, which only contains two language with very high local sensitivity, all language and tasks exhibit the same overall behaviour. Languages with low local sensitivity invariably have low performance.

Languages	Coptic	Northwestern Ojibwa	Inuktitut	Lao	Dhivehi	S'gaw Karen	Yoruba	Khmer
Sensitivity	0.973	0.978	0.979	0.979	0.980	0.981	0.982	0.73
Family	Afro-Asiatic	Algic	Eskimo–Aleut	Kra-Dai	IE: Indo-Iranian	Sino-Tibetan	Niger-Congo	Austroasiatic
Script	Coptic	Latin	Inuktitut syllabics	Brahmic	Thaana	Brahmic	Latin	Brahmic
Native Speaker (Millions)	0.0	0.02	0.04	30	0.34	3	43	16
Languages	Maori	Sinhala	Samoan	Cherokee	Syriac	Nzima	Oriya	Venda
Sensitivity	0.983	0.984	0.984	0.985	0.985	0.985	0.987	0.987
Family	Austronesian	IE: Indo-Iranian	Autronesian	Iroquoian	Afro-Asiatic	Niger-Congo	IE: Indo-Iranian	Niger-Congo
Script	Latin	Brahmic	Latin	Latin	Aramaic	Latin	Brahmic	Latin
Native Speaker (Millions) [6]	0.05	17	0.51	0.002	0.24	0.41	35	1.3

Table 3: Statistics on the language containing the lowest monolingual local sensitivity of all 350 languages.

Low-Performance Languages In Figure 12, we have plotted the monolingual local sensitivity of all 350 languages on a world map at their geographical centers (Haspelmath et al., 2014; Dryer and Haspelmath, 2013), with their size scaled by amount of native speakers in those specific languages. Many statements can be made about low-performance languages from this study.

It seems that languages that are geographically close to Europe or South-East Asia are generally well understood by our cross-lingual models. The majority of poorly understood languages seem to either be concentrated in Sub-Saharan Africa or Central America, as well as island-specific languages across the Pacific ocean.

The languages that have the lowest monolingual local sensitivity are reported in Table 3. Some of those languages, like Coptic, a now long-dead language in an unseen script, are fairly obvious low-performers. Our approach, however, seems able to detect low-performance in languages that would not be that obvious and would be quite important to detect, like Lao with its over 30 million native speakers.

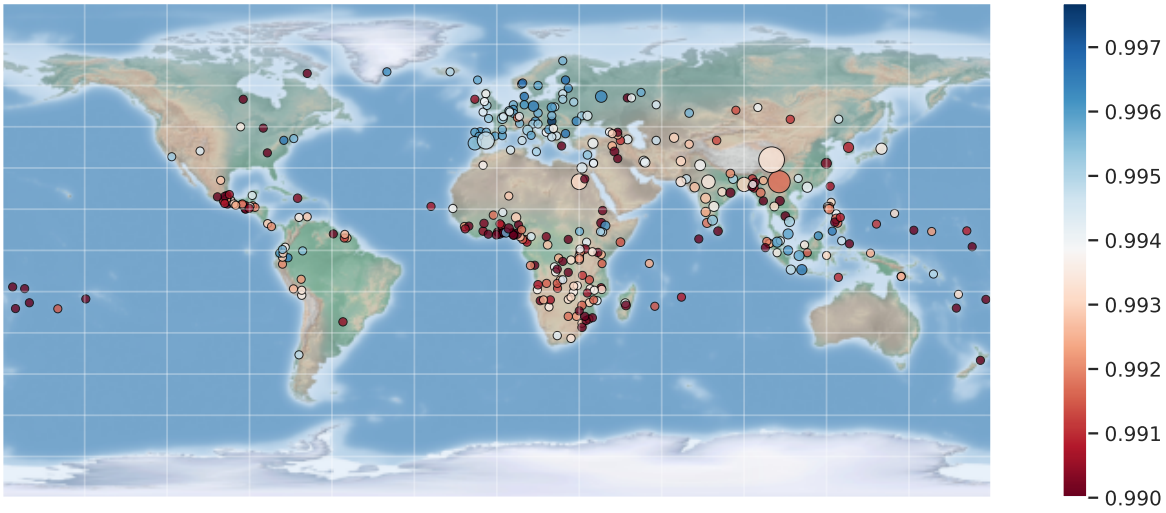


Figure 12: Monolingual local sensitivity of 350 languages on the task of cross-lingual similarity on our MTData cross-lingual sentence similarity dataset, scaled by the estimated amount of native speakers.

C MTData Sentence Similarity Task

From the MTData dataset (Gowda et al., 2021) we build a sentence similarity dataset covering a total of 350 languages. We use and adapt the approach used to build the Tatoeba sentence retrieval dataset (Artetxe and Schwenk, 2018). Specifically, from the MTData dataset containing english aligned sentences in over 500 languages, we remove all sentences containing either "@", "http" or "%", remove any English sentence containing less than 3 words, and remove any duplicate. We randomly sample 1000 sentence pairs per language, removing languages with less then 1000 sentence pair present after filtering. We also remove text of sign languages, as their written form is almost exactly the same as the original language. In total, 350 languages remain after that point. Table 4 to Table 9 contains statistics on every single language present in our MTData Sentence Retrieval Task.

D Pseudocode for Perturbation

Function `NeighborFlip`($\rho \leftarrow 0.5$, *text*$\leftarrow$`list`)**:**
    perturbed_tokens $\leftarrow$ `list()`;
    held_token $\leftarrow$ `list(text[0])`
    for *token* **in** text$[1:]$ **do**
        p $\sim$ *Unif* $([0, 1])$;
        if $p < \rho$ **then**
            perturbed_tokens.append(held_token);
            held_token $\leftarrow$ `list`(token)
        else
            perturbed_tokens $\leftarrow$ [perturbed_tokens, token];
        end
    end
    perturbed_tokens.append(held_token);
    perturbed_text $\leftarrow$ ''.`join`(perturbed_tokens)
return perturbed_text

Algorithm 1: Pseudocode for NeighborFlip.

Language	ISO-639-3	Language Family	Language Script	Num Native Speakers
Chinese	zho	Sino-Tibetan	Chinese characters and derivatives	1300000000
Mandarin Chinese	cmn	Sino-Tibetan	Chinese characters and derivatives	920000000
Spanish	spa	IE: Italic	Latin	493000000
Arabic	ara	Afro-Asiatic	Arabic	400000000
Bengali	ben	IE: Indo-Iranian	Brahmic	300000000
Hindi	hin	IE: Indo-Iranian	Brahmic	260000000
Portuguese	por	IE: Italic	Latin	250000000
Russian	rus	IE: Balto-Slavic	Cyrillic	150000000
Japanese	jpn	Japonic	Kana	128000000
Panjabi	pan	IE: Indo-Iranian	Arabic	113000000
German	deu	IE: Germanic	Latin	95000000
Yue Chinese	yue	Sino-Tibetan	Chinese characters and derivatives	84000000
Egyptian Arabic	arz	Afro-Asiatic	Arabic	83000000
Javanese	jav	Autronesian	Brahmic	82000000
Korean	kor	Koreanic	Hangul	80400000
Turkish	tur	Turkic	Latin	80000000
Wu Chinese	wuu	Sino-Tibetan	Chinese characters and derivatives	80000000
Malay (individual language)	zlm	Autronesian	Arabic	77000000
Malay (macrolanguage)	msa	Austronesian	Latin	77000000
Standard Malay	zsm	Autronesian	Arabic	77000000
French	fra	IE: Italic	Latin	76800000
Vietnamese	vie	Austroasiatic	Latin	76000000
Telugu	tel	Dravidian	Brahmic	75000000
Marathi	mar	IE: Indo-Iranian	Brahmic	73000000
Persian	fas	IE: Indo-Iranian	Arabic	70000000
Tamil	tam	Dravidian	Brahmic	70000000
Urdu	urd	IE: Indo-Iranian	Arabic	70000000
Italian	ita	IE: Italic	Latin	67000000
Iranian Persian	pes	IE: Indo-Iranian	Arabic	55600000
Gujarati	guj	IE: Indo-Iranian	Brahmic	50000000
Hausa	hau	Afro-Asiatic	Latin	50000000
Pushto	pus	IE: Indo-Iranian	Arabic	50000000
Tagalog	tgl	Autronesian	Latin	45000000
Polish	pol	IE: Balto-Slavic	Latin	45000000
Filipino	fil	Austronesian	Latin	45000000
Uzbek	uzb	Turkic	Latin	44000000
Indonesian	ind	Autronesian	Latin	43000000
Yoruba	yor	Niger-Congo	Latin	43000000
Kannada	kan	Dravidian	Brahmic	43000000
Sundanese	sun	Austronesian	Latin	42000000
Ukrainian	ukr	IE: Balto-Slavic	Cyrillic	40000000
Nigerian Pidgin	pcm	English Creole	Latin	40000000
Oromo	orm	Afro-Asiatic	Latin	37400000
Oriya (macrolanguage)	ori	IE: Indo-Iranian	Brahmic	35000000
Malayalam	mal	Dravidian	Brahmic	35000000
Maithili	mai	IE: Indo-Iranian	Brahmic	33900000
Burmese	mya	Sino-Tibetan	Brahmic	33000000
Amharic	amh	Afro-Asiatic	Ge'ez	32000000
Azerbaijani	aze	Turkic	Arabic	30000000
Lao	lao	Kra-Dai	Brahmic	30000000
Igbo	ibo	Niger-Congo	Latin	30000000
Thai	tha	Kra-Dai	Brahmic	28000000
Sindhi	snd	IE: Indo-Iranian	Arabic	25000000
Malagasy	mlg	Austronesian	Latin	25000000
Plateau Malagasy	plt	Austronesian	Latin	25000000
Dutch	nld	IE: Germanic	Latin	25000000
Kurdish	kur	IE: Indo-Iranian	Arabic	25000000
Romanian	ron	IE: Italic	Latin	23800000
Cebuano	ceb	Autronesian	Latin	22000000
Somali	som	Afro-Asiatic	Latin	21807730
Croatian	hrv	IE: Balto-Slavic	Cyrillic	21000000
Ganda	lug	Niger-Congo	Latin	20000000
Ewe	ewe	Niger-Congo	Latin	20000000
Swahili (macrolanguage)	swa	Niger-Congo	Latin	18000000
Chhattisgarhi	hne	IE: Indo-Iranian	Brahmic	18000000
Kazakh	kaz	Turkic	Cyrillic	17800000
Lingala	lin	Niger-Congo	Latin	17500000
Sinhala	sin	IE: Indo-Iranian	Brahmic	17000000
Nepali (macrolanguage)	nep	IE: Indo-Iranian	Brahmic	16000000

Table 4: Statistics on all 350 languages present in the MTData sentence retrieval dataset. (1 of 6)

Language	ISO-639-3	Language Family	Language Script	Num Native Speakers
Khmer	khm	Austroasiatic	Brahmic	16000000
Assamese	asm	IE: Indo-Iranian	Brahmic	15311351
Northern Kurdish	kmr	IE: Indo-Iranian	Arabic	15000000
Bavarian	bar	IE: Germanic	Latin	14000000
Modern Greek (1453-)	ell	IE: Hellenic	Greek	13400000
Hungarian	hun	Uralic	Latin	13000000
Umbundu	umb	Niger-Congo	Latin	12740000
Haitian	hat	IE: Italic	Latin	12000000
Shona	sna	Niger-Congo	Latin	12000000
Zulu	zul	Niger-Congo	Latin	12000000
Serbian	srp	IE: Balto-Slavic	Cyrillic	12000000
Nyanja	nya	Niger-Congo	Latin	12000000
Rundi	run	Niger-Congo	Latin	11244750
Turkmen	tuk	Turkic	Latin	11000000
Czech	ces	IE: Balto-Slavic	Latin	10700000
Swedish	swe	IE: Germanic	Latin	10000000
Uighur	uig	Turkic	Arabic	10000000
Tigrinya	tir	Afro-Asiatic	Ge'ez	9850000
Kinyarwanda	kin	Niger-Congo	Latin	9800000
Congo Swahili	swc	Niger-Congo	Latin	9000000
Xhosa	xho	Niger-Congo	Latin	8700000
Ga	gaa	Niger-Congo	Latin	8500000
Iloko	ilo	Austronesian	Latin	8100000
Tajik	tgk	IE: Indo-Iranian	Cyrillic	8100000
Bulgarian	bul	IE: Balto-Slavic	Cyrillic	8000000
Quechua	que	Quechuan	Latin	8000000
Mossi	mos	Niger-Congo	Latin	7830000
Hiligaynon	hil	Austronesian	Latin	7800000
Makhuwa	vmw	Niger-Congo	Latin	7400000
Afrikaans	afr	IE: Germanic	Arabic	7200000
Dyula	dyu	Mande	Latin	6852620
Kikuyu	kik	Niger-Congo	Latin	6600000
Paraguayan Guaraní	gug	Tupian	Latin	6500000
San Salvador Kongo	kwy	Niger-Congo	Latin	6500000
Kongo	kon	Niger-Congo	Latin	6500000
Luba-Lulua	lua	Niger-Congo	Latin	6300000
Low German	nds	IE: Germanic	Latin	6000000
Armenian	hye	IE: Armenian	Armenian	6000000
Albanian	sqi	IE: Albanian	Latin	6000000
Danish	dan	IE: Germanic	Latin	6000000
Kabyle	kab	Afro-Asiatic	Arabic	6000000
Finnish	fin	Uralic	Latin	5800000
Wolof	wol	Niger-Congo	Latin	5454000
Norwegian	nor	IE: Germanic	Latin	5320000
Slovak	slk	IE: Balto-Slavic	Latin	5200000
Tatar	tat	Turkic	Cyrillic	5200000
Tswana	tsn	Niger-Congo	Latin	5200000
Mongolian	mon	Mongolic	Cyrillic	5200000
Belarusian	bel	IE: Balto-Slavic	Cyrillic	5100000
Tiv	tiv	Niger-Congo	Latin	5000000
Hebrew	heb	Afro-Asiatic	Aramaic	5000000
Pedi	nso	Niger-Congo	Latin	4700000
Baoulé	bci	Niger-Congo	Latin	4700000
Kirghiz	kir	Turkic	Cyrillic	4500000
Luo (Kenya and Tanzania)	luo	Nilo-Saharan	Latin	4200000
Bemba (Zambia)	bem	Niger-Congo	Latin	4100000
Kamba (Kenya)	kam	Niger-Congo	Latin	3900000
Tachelhit	shi	Afro-Asiatic	Arabic	3900000

Table 5: Statistics on all 350 languages present in the MTData sentence retrieval dataset. (2 of 6)

Language	ISO-639-3	Language Family	Language Script	Num Native Speakers
Lombard	lmo	IE: Italic	Latin	3800000
Georgian	kat	Kartvelian	Georgian	3700000
Hmong	hmn	Hmong–Mien	Latin	3700000
Tsonga	tso	Niger-Congo	Latin	3700000
Waray (Philippines)	war	Autronesian	Latin	3600000
Zarma	dje	Nilo-Saharan	Latin	3600000
Tumbuka	tum	Niger-Congo	Latin	3546000
Romany	rom	IE: Indo-Iranian	Latin	3500000
Nyankole	nyn	Niger-Congo	Latin	3400000
Yao	yao	Niger-Congo	Latin	3100000
Lithuanian	lit	IE: Balto-Slavic	Latin	3000000
S'gaw Karen	ksw	Sino-Tibetan	Brahmic	3000000
Sidamo	sid	Afro-Asiatic	Latin	3000000
Pampanga	pam	Autronesian	Brahmic	2800000
Slovenian	slv	IE: Balto-Slavic	Latin	2500000
Macedonian	mkd	IE: Balto-Slavic	Cyrillic	2500000
Bosnian	bos	IE: Balto-Slavic	Cyrillic	2500000
Central Bikol	bcl	Austronesian	Latin	2500000
Galician	glg	IE: Italic	Latin	2400000
Ndau	ndc	Niger-Congo	Latin	2400000
Iban	iba	Autronesian	Latin	2300000
Swati	ssw	Niger-Congo	Latin	2300000
Fon	fon	Niger-Congo	Latin	2200000
Kimbundu	kmb	Niger-Congo	Latin	2100000
Acoli	ach	Nilo-Saharan	Latin	2100000
Cameroon Pidgin	wes	English Creole	Latin	2000000
Urhobo	urh	Niger-Congo	Latin	2000000
Lomwe	ngl	Niger-Congo	Latin	1850000
Pangasinan	pag	Austronesian	Latin	1800000
Latvian	lav	IE: Balto-Slavic	Latin	1750000
Alur	alz	Nilo-Saharan	Latin	1700000
Aymara	aym	Aymaran	Latin	1700000
Batak Toba	bbc	Austronesian	Latin	1610000
Wolaytta	wal	Afro-Asiatic	Latin	1600000
Sena	seh	Niger-Congo	Latin	1600000
Bini	bin	Niger-Congo	Latin	1600000
Luba-Katanga	lub	Niger-Congo	Latin	1505000
Mende (Sierra Leone)	men	Mande	Latin	1500000
Yiddish	yid	IE: Germanic	Aramaic	1500000
Cusco Quechua	quz	Quechuan	Latin	1500000
Tonga (Zambia)	toi	Niger-Congo	Latin	1500000
Kuanyama	kua	Niger-Congo	Latin	1441000
Bashkir	bak	Turkic	Cyrillic	1400000
Limburgan	lim	IE: Germanic	Latin	1300000
Southwestern Dinka	dik	Nilo-Saharan	Latin	1300000
Venda	ven	Niger-Congo	Latin	1300000
Manipuri	mni	Sino-Tibetan	Brahmic	1250000
Tswa	tsc	Niger-Congo	Latin	1200000
Batak Simalungun	bts	Austronesian	Latin	1200000
Sardinian	srd	IE: Italic	Latin	1175000

Table 6: Statistics on all 350 languages present in the MTData sentence retrieval dataset. (3 of 6)

Language	ISO-639-3	Language Family	Language Script	Num Native Speakers
Gun	guw	Niger-Congo	Latin	1162000
Kekchí	kek	Mayan	Latin	1100000
Estonian	est	Uralic	Latin	1100000
Zande (individual language)	zne	Niger-Congo	Latin	1100000
K'iche'	quc	Mayan	Latin	1100000
Morisyen	mfe	French Creole	Latin	1090000
Chuvash	chv	Turkic	Cyrillic	1042989
Kabiyè	kbp	Niger-Congo	Latin	1000000
Songe	sop	Niger-Congo	Latin	1000000
Central Huasteca Nahuatl	nch	Uto-Aztecan	Latin	1000000
Chokwe	cjk	Niger-Congo	Latin	980000
Chuwabu	chw	Niger-Congo	Latin	970000
Kachin	kac	Sino-Tibetan	Latin	940000
Ayacucho Quechua	quy	Quechuan	Latin	918200
Welsh	cym	IE: Celtic	Latin	892200
Ngaju	nij	Austronesian	Latin	890000
Kabuverdianu	kea	English Creole	Latin	871000
Bulu (Cameroon)	bum	Niger-Congo	Latin	860000
Lushai	lus	Sino-Tibetan	Brahmic	843750
Ndonga	ndo	Niger-Congo	Latin	810000
Adangme	ada	Niger-Congo	Latin	800000
Yucateco	yua	Mayan	Latin	770000
Nias	nia	Austronesian	Latin	770000
Chopi	cce	Niger-Congo	Latin	760000
Tetela	tll	Niger-Congo	Latin	760000
Basque	eus	Basque	Latin	750000
Nyaneka	nyk	Niger-Congo	Latin	750000
Lozi	loz	Niger-Congo	Latin	725000
Chavacano	cbk	IE: Italic	Latin	700000
Luvale	lue	Niger-Congo	Latin	640000
Konzo	koo	Niger-Congo	Latin	610000
Walloon	wln	IE: Italic	Latin	600000
Mam	mam	Mayan	Latin	600000
Batak Karo	btx	Austronesian	Latin	600000
Luxembourgish	ltz	IE: Germanic	Latin	600000
Ossetian	oss	IE: Indo-Iranian	Cyrillic	597450
Tzeltal	tzh	Mayan	Latin	590000
Balkan Romani	rmn	IE: Indo-Iranian	Latin	563670
Udmurt	udm	Uralic	Cyrillic	554000
Tzotzil	tzo	Mayan	Latin	550000
Norwegian Nynorsk	nno	IE: Germanic	Latin	532000
Southern Kisi	kss	Niger-Congo	Latin	530000
Maltese	mlt	Afro-Asiatic	Latin	520000
Samoan	smo	Autronesian	Latin	510000
Mambwe-Lungu	mgr	Niger-Congo	Latin	500000
Tamashek	tmh	Afro-Asiatic	Latin	500000
Krio	kri	English Creole	Latin	500000
Imbabura Highland Quichua	qvi	Quechuan	Latin	500000
Tooro	ttj	Niger-Congo	Latin	490000

Table 7: Statistics on all 350 languages present in the MTData sentence retrieval dataset. (4 of 6)

Language	ISO-639-3	Language Family	Language Script	Num Native Speakers
Western Frisian	fry	IE: Germanic	Latin	470000
Sango	sag	French Creole	Latin	450000
Plautdietsch	pdt	IE: Germanic	Latin	450000
Occitan (post 1500)	oci	IE: Italic	Latin	450000
Chimborazo Highland Quichua	qug	Quechuan	Latin	450000
Hakha Chin	cnh	Sino-Tibetan	Latin	446264
Nyungwe	nyu	Niger-Congo	Latin	440000
Friulian	fur	IE: Italic	Latin	420000
Isoko	iso	Niger-Congo	Latin	420000
Nzima	nzi	Niger-Congo	Latin	412000
Catalan	cat	IE: Italic	Latin	410000
Kaqchikel	cak	Mayan	Latin	410000
Efik	efi	Niger-Congo	Latin	400000
Ibanag	ibg	Autronesian	Latin	400000
Lunda	lun	Niger-Congo	Latin	400000
Tetun Dili	tdt	Autronesian	Latin	390000
Gitonga	toh	Niger-Congo	Latin	380000
Mingrelian	xmf	Kartvelian	Georgian	344000
Papiamento	pap	IE: Italic	Latin	341300
Dhivehi	div	IE: Indo-Iranian	Thaana	340000
Fijian	fij	Austronesian	Latin	339210
Icelandic	isl	IE: Germanic	Latin	314000
Wayuu	guc	Arawakan	Latin	305000
Esan	ish	Niger-Congo	Latin	300000
Basa (Cameroon)	bas	Niger-Congo	Latin	300000
Tuvinian	tyv	Turkic	Cyrillic	280000
Mapudungun	arn	Araucanian	Latin	260000
Ruund	rnd	Niger-Congo	Latin	250000
Syriac	syr	Afro-Asiatic	Aramaic	240000
Kaonde	kqn	Niger-Congo	Latin	240000
Huautla Mazatec	mau	Oto-Manguean	Latin	240000
Nyemba	nba	Niger-Congo	Latin	232000
Herero	her	Niger-Congo	Latin	211700
Breton	bre	IE: Celtic	Latin	210000
Amis	ami	Austronesian	Latin	200000
Garifuna	cab	Arawakan	Latin	200000
Sangir	sxn	Austronesian	Latin	200000
Northern Puebla Nahuatl	ncj	Uto-Aztecan	Latin	200000
Lamba	lam	Niger-Congo	Latin	200000
Abkhazian	abk	Northwest Caucasian	Cyrillic	190000
Tonga (Tonga Islands)	ton	Austronesian	Latin	187000
Tahitian	tah	Austronesian	Latin	185000
Navajo	nav	Dené-Yeniseian	Latin	170000
Ngäbere	gym	Chibchan	Latin	170000
Irish	gle	IE: Celtic	Latin	170000
Tonga (Nyasa)	tog	Niger-Congo	Latin	170000
Kwangali	kwn	Niger-Congo	Latin	152000
Malinaltepec Me'phaa	tcf	Oto-Manguean	Latin	150000
Belize Kriol English	bzj	English Creole	Latin	150000
Metlatónoc Mixtec	mxv	Oto-Manguean	Latin	150000
Guerrero Nahuatl	ngu	Uto-Aztecan	Latin	150000
Purepecha	tsz	Purepecha	Latin	140000
Kadazan Dusun	dtp	Autronesian	Latin	140000
Sranan Tongo	srn	English Creole	Latin	130000
Tok Pisin	tpi	English Creole	Latin	120000
Gilbertese	gil	Austronesian	Latin	120000

Table 8: Statistics on all 350 languages present in the MTData sentence retrieval dataset. (5 of 6)

Language	ISO-639-3	Language Family	Language Script	Num Native Speakers
Paite Chin	pck	Sino-Tibetan	Latin	100000
Saramaccan	srm	English Creole	Latin	90000
Duala	dua	Niger-Congo	Latin	87700
Isthmus Zapotec	zai	Oto-Manguean	Latin	85000
Galela	gbi	West Papuan	Latin	80000
Papantla Totonac	top	Mayan	Latin	80000
Seselwa Creole French	crs	French Creole	Latin	73000
Faroese	fao	IE: Germanic	Latin	72000
Lukpa	dop	Niger-Congo	Latin	70000
Biak	bhw	Austronesian	Latin	70000
Tojolabal	toj	Mayan	Latin	67000
Eastern Maroon Creole	djk	English Creole	Latin	67000
Guerrero Amuzgo	amu	Oto-Manguean	Latin	60000
Chamorro	cha	Autronesian	Latin	58000
Scottish Gaelic	gla	IE: Celtic	Latin	57000
Kalaallisut	kal	Eskimo–Aleut	Latin	56000
Southern Altai	alt	Turkic	Cyrillic	55720
Marshallese	mah	Austronesian	Latin	55000
Aguaruna	agr	Chicham	Latin	53400
Chuukese	chk	Austronesian	Latin	51330
Aragonese	arg	IE: Italic	Latin	50000
Maori	mri	Austronesian	Latin	50000
Coatlán Mixe	mco	Mixe–Zoque	Latin	45000
Chol	ctu	Mayan	Latin	43870
Inuktitut	iku	Eskimo–Aleut	Inuktitut syllabics	39770
Asháninka	cni	Arawakan	Latin	35000
Shuar	jiv	Chicham	Latin	35000
Pohnpeian	pon	Austronesian	Latin	29000
Jakun	jak	Austronesian	Latin	28000
Northern Sami	sme	Uralic	Latin	25000
Okpe (Southwestern Edo)	oke	Niger-Congo	Latin	25000
Pijin	pis	English Creole	Latin	24000
Uma	ppk	Austronesian	Latin	20000
Northwestern Ojibwa	ojb	Algic	Latin	20000
Tena Lowland Quichua	quw	Quechuan	Latin	17855
Central Puebla Nahuatl	ncx	Uto-Aztecan	Latin	16000
Mirandese	mwl	IE: Italic	Latin	15000
Dehu	dhv	Austronesian	Latin	13000
Wallisian	wls	Austronesian	Latin	10400
Bislama	bis	IE: Germanic	Latin	10000
Akawaio	ake	Cariban	Latin	10000
Quiotepec Chinantec	chq	Oto-Manguean	Latin	10000
Cabécar	cjp	Chibchan	Latin	8800
Yapese	yap	Austronesian	Latin	5130
Uspanteco	usp	Mayan	Latin	5100
Camsá	kbh	Oto-Manguean	Camsa	4000
Achuar-Shiwiar	acu	Chicham	Latin	4000
Tetelcingo Nahuatl	nhg	Uto-Aztecan	Latin	3500
Cherokee	chr	Iroquoian	Latin	2100
Asturian	ast	IE: Italic	Latin	2000
Niuean	niu	Austronesian	Latin	2000
Barasana-Eduria	bsn	Tucanoan	Latin	1900
Interlingua (International Auxiliary Language Association)	ina	Constructed	Latin	1500
Esperanto	epo	Constructed	Latin	1000
Cornish	cor	IE: Celtic	Latin	557
Rarotongan	rar	Austronesian	Latin	450
Hiri Motu	hmo	Austronesian	Latin	100
Potawatomi	pot	Algic	Latin	100
Manx	glv	IE: Celtic	Latin	53
Klingon	tlh	Constructed	Latin	25
Ido	ido	Constructed	Latin	25
Volapük	vol	Constructed	Latin	20
Latin	lat	IE: Italic	Latin	1
Interlingue	ile	Constructed	Latin	1
Coptic	cop	Afro-Asiatic	Coptic	1
Lojban	jbo	Constructed	Latin	1
Lingua Franca Nova	lfn	Constructed	Latin	1

Table 9: Statistics on all 350 languages present in the MTData sentence retrieval dataset. (6 of 6)