Zero-shot Cross-lingual POS Tagging for Filipino

Jimson Paulo Layacan, Isaiah Edri W. Flores, Katrina Bernice M. Tan,
Ma. Regina E. Estuar, Jann Railey E. Montalan, Marlene M. De Leon
Ateneo Social Computing Lab, Department of Information Systems and Computer Science
Ateneo de Manila University
Quezon City, Philippines

Abstract

Supervised learning approaches in NLP, exemplified by POS tagging, rely heavily on the presence of large amounts of annotated data. However, acquiring such data often requires significant amount of resources and incurs high costs. In this work, we explore zero-shot cross-lingual transfer learning to address data scarcity issues in Filipino POS tagging, particularly focusing on optimizing source language selection. Our zero-shot approach demonstrates superior performance compared to previous studies, with top-performing fine-tuned PLMs achieving F1 scores as high as 79.10%. The analysis reveals moderate correlations between cross-lingual transfer performance and specific linguistic distances–featural, inventory, and syntactic–suggesting that source languages with these features closer to Filipino provide better results. We identify tokenizer optimization as a key challenge, as PLM tokenization sometimes fails to align with meaningful representations, thus hindering POS tagging performance.

1 Introduction

The rise of pretrained language models (PLMs) has revolutionized the landscape of natural language processing (NLP). While these models demonstrably address data scarcity in under-resource languages by learning universal language representations (Qiu et al., 2020), many languages, including Filipino, a widely spoken under-resource language in the Philippines (Lewis, 2009), continue to face significant challenges. Building robust NLP pipelines for Filipino remains difficult despite the abundance of textual resources like literary works, linguistic references, and social media data.

Filipino lacks dedicated resources for a range of language processing tasks (Aquino and de Leon, 2020; Cruz and Cheng, 2021; Miranda, 2023). Robust and reliable part-of-speech (POS) taggers could significantly improve the performance of such tasks by accurately classifying words into their grammatical categories. This disambiguation is essential because many words can have multiple meanings based on context. For example, the Filipino word "buhay" can be a "pangngalan" (noun) meaning "life" or a "pang-uri" (adjective) meaning "lively" or "vibrant." By clearing up word confusion, POS tagging helps in performing higher-level NLP tasks such as machine translation, information extraction, text-to-speech conversion, speech recognition, etc.

However, annotating datasets for POS tagging is complex and resource-intensive. One potential solution is cross-lingual transfer learning, which involves using the knowledge gained from training a model in one language to address tasks in another language (Kim et al., 2017). In this paradigm, a language model acquires representations from a source language and then undergoes fine-tuning to execute tasks in a target language with limited labeled data. Furthermore, zero-shot learning, a specific form of cross-lingual transfer learning, presents a solution in scenarios with a complete absence of annotated data (de Vries et al., 2022).

One crucial factor in enhancing zero-shot cross-lingual transfer learning is the selection of the source language. This selection process involves identifying and analyzing language similarity metrics that can improve the success of cross-lingual transfer learning (Eronen et al., 2023). These metrics quantify and compare linguistic and structural correspondences between languages.

Linguists often use intuitive notions of structure to compare languages (Stabler and Keenan, 2003), and source language selection tends to follow similar intuitive approaches. However, quantified language similarity metrics provide a more objective basis for these comparisons, suggesting that higher similarity between a source-target language pair generally results in improved cross-lingual transfer learning performance. The challenge, however, lies in selecting the most appropriate similarity metric,

Proceedings of the 3rd Workshop on NLP Applications to Field Linguistics (Field Matters 2024), pages 69–77

given the wide array of available options. Identifying which metrics are most indicative of successful cross-lingual transfer learning could streamline the source language selection process, thereby enhancing adaptability for under-resource languages such as Filipino.

Prior studies have explored the impact of several linguistic features on cross-lingual transfer performance. One study emphasized the correlation between linguistic similarity and transfer performance, advocating for selecting source languages based on rigorous linguistic assessments rather than defaulting to English (Eronen et al., 2023). In contrast, another study proposed exploring syntactic and morphological similarities across languages to improve model transfer capabilities (Philippy et al., 2023). Additionally, another study emphasized the importance of including linguistically similar languages in pre-training for improved transfer learning outcomes (de Vries et al., 2022). Our paper extends this line of research by examining linguistic similarity distances between Filipino and source languages and within the context of zero-shot learning for POS tagging.

More specifically, we examined how measures of linguistic distances across multiple dimensions contributed to the effectiveness of POS tagging. While a study (Philippy et al., 2023) investigated this aspect for the Natural Language Inference (NLI) task across all 15 languages in the XNLI dataset (Conneau et al., 2018) individually, our focus is on POS tagging and Filipino as the target language. Futhermore, we investigated how the choice of PLM influenced the outcome and effectiveness of source language selection. We also explored which source language and combination of source languages yielded the highest F1 scores for Filipino POS tagging.

2 Language Similarity

Lang2vec (Littell et al., 2017) is a versatile tool for linguistic analysis that provides readily available pre-computed distances between languages represented as vectors of featural, syntactic, geographic, inventory, genetic, and phonological dimensions from multiple databases including the World Atlas of Language Structures (WALS) (Dryer and Haspelmath, 2013), Syntactic Structures of World Languages (SSWL) (Collins and Kayne, 2009), PHOIBLE (Moran and McCloy, 2019), Glottolog (Hammarström et al., 2018) tree of language families, and Ethnologue (Lewis, 2009). These dimensions enable comparisons of various linguistic features across different languages. Understanding cross-lingual transfer performance in Filipino POS tagging will benefit an investigation of language similarity metrics.

- **Featural Distance** is the cosine distance between vectors defined by features across multiple databases. If a feature value is unknown in one of the languages, it is excluded from the calculation.
- **Genetic Distance** is based on the Glottolog tree of language families, calculated as the distance between two languages in the tree.
- **Geographic Distance** is the shortest distance between two languages on the Earth's sphere, also known as orthodromic distance.
- **Syntactic, Phonological, and Inventory distances** are computed based on specific features identified in the databases, distinguishing between syntactic, phonological, and inventory features.

3 Methods

We used a selection of PLMs, including XLM-R (Conneau et al., 2019), a multilingual variant of the RoBERTa model, and RoBERTa-Tagalog (Cruz and Cheng, 2021), a RoBERTa model pretrained using a Filipino-language pretraining corpus. In this study, both models were finetuned and assessed in a zero-shot cross-lingual scenario, tasked with performing POS tagging for Filipino texts using their base configurations. XLM-R was selected for its well-established performance in multilingual contexts and its robustness in handling large-scale text datasets across various sequence-labeling tasks (Qiu et al., 2020). RoBERTa-Tagalog, on the other hand, was chosen because it is an improvement over the previous Tagalog pretrained Transformer models (Cruz and Cheng, 2021).

3.1 PLM Fine-tuning

Two modeling approaches were employed. First, each PLM was finetuned on data from a single source language and then used to predict POS tags for Filipino text without any further training. This approach assesses the models' ability to generalize to a new language based on their knowledge of the source language. Second, the better-performing

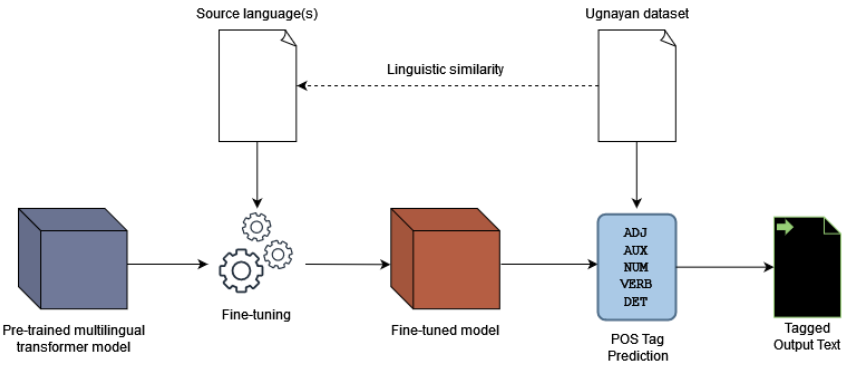

Figure 1: Methodological pipeline for developing POS tagging models (Eronen et al., 2023)

PLM was finetuned on data from several source languages using a progressive approach inspired by curriculum learning (Bengio et al., 2009), adding languages one at a time, starting from the top-performing source language in the monolingual training. This strategy leverages information from related languages, potentially improving the generalizability of the PLM by exposing it to a broader training data.

All models were trained with the same hyperparameter settings. Specifically, the models were trained for 1,000 batches, each containing 10 samples, using a linearly decreasing learning rate starting at 5e-5. These hyperparameters were chosen based on De Vries's configuration (de Vries et al., 2022), which employed a comprehensive transfer learning setup with multiple source and target languages for POS tagging.

3.2 Training and Testing Data

The training dataset for the PLMs was sourced from the Universal Dependencies (UD) 2.13 dataset (De Marneffe et al., 2021). This dataset is designed to facilitate cross-lingual learning and parsing projects by providing a consistent annotation framework across multiple languages. Only languages with available training data were included in this study, with no additional eliminations, as the focus was on establishing a comprehensive setup for a single target language: Filipino.

The UD framework is built on linguistic typology and supports comparisons across languages through consistent annotation. It includes 17 Universal POS (UPOS) tags and comprises 259 treebanks for 148 languages. Below is a list of the UPOS tags used in the dataset (see Table 1).

Note that the varying quality of UD datasets is a limitation. Some corpora lack diversity in writing styles, and UD updates are inconsistent across languages, with some shifting towards language-specific features and augmented dependencies while fundamental syntactic structures remain problematic (Iwamoto et al., 2021). This may have impacted our cross-lingual transfer learning results, as model performance is sensitive to training data quality.

The finetuned models were evaluated on the Ugnayan dataset (Aquino and de Leon, 2020), which is a standard benchmark for Filipino POS tagging. The performance of these models was measured using the F1 score. This dataset includes 94 sentences with 1011 manually annotated tokens. The Ugnayan dataset, sourced from resources on the Philippines' Department of Education Learning Resource Portal, provides a broad range of sentence

Table 1: Universal POS (UPOS) Tags

Tag	Description
ADJ	Adjective
ADP	Adposition
ADV	Adverb
AUX	Auxiliary
CCONJ	Coordinating Conjunction
DET	Determiner
INTJ	Interjection
NOUN	Noun
NUM	Numeral
PART	Particle
PRON	Pronoun
PROPN	Proper Noun
PUNCT	Punctuation
SCONJ	Subordinating Conjunction
SYM	Symbol
VERB	Verb
X	Other

structures and syntactic phenomena, utilizing 14 out of the 17 UPOS tags.

3.3 Language Similarity and Learning Performance

The linguistic distances between Filipino and source languages were extracted across various dimensions. These distances were represented as normalized values, creating lists of distances between Filipino and each respective source language. For instance, syntactic distances quantified the similarity between syntax features of Filipino and other languages, with values ranging from 0 to 1.

Each of these lists was then subjected to correlation analysis with the F1 scores obtained from the finetuned models, both XLM-R and RoBERTa-Tagalog. The correlation analysis involved computing Pearson's correlation coefficients to quantify the relationship between language distances and cross-lingual transfer performance. Significance testing was conducted to assess the statistical significance of the observed correlations.

4 Results

The results of the top-performing finetuned PLMs outperform all previously presented zero-shot learning methods listed in Table 2. Specifically, the approach utilizing single-source language fine-tuning achieved the highest F1 score of 79.10%, representing a significant improvement over the highest score achieved by previous methods (Aquino and de Leon, 2022). This improvement demonstrates the effectiveness of the fine-tuning methodology for PLMs, particularly for Filipino POS tagging.

Table 2: Previous zero-shot methods (Aquino and de Leon, 2022) and their corresponding F1 scores for POS tagging on the Ugnayan dataset

Zero-shot Method	F1
UDify (zero-shot baseline)	59.80
POS tag conversion (MGNN)	68.19
POS projection (en)	61.17
POS projection (en+id+it+pl)	61.90

Table 3 shows that, for XLM-R, Afrikaans emerged as the top-performing source language, despite its distant relation to Filipino. Afrikaans is a Germanic language, while Filipino is Austronesian, placing them in very different language families. However, this unexpected result suggests that the two seemingly different languages share some linguistic features.

Table 3: Top 10 best-performing source languages for XLM-R monolingual fine-tuning

Rank	XLM-R	F1
1	Afrikaans	79.10
2	Hebrew	77.02
3	Bulgarian	77.00
4	Vietnamese	76.78
5	Norwegian	75.83
6	Urdu	75.47
7	Czech	75.40
8	Persian	75.36
9	Faroese	75.36
10	English	75.33

Table 4: Top 10 best-performing source languages for RoBERTa-Tagalog monolingual fine-tuning

Rank	RoBERTa-Tagalog	F1
1	English	71.63
2	Naija/Nigerian Pidgin	45.94
3	Serbian	42.47
4	Manx-Cadhan	42.04
5	Slovenian	41.22
6	Spanish	41.20
7	Dutch	41.19
8	Croatian	41.12
9	Polish	40.76
10	Irish	40.35

One potential similarity is their flexible word order, which allows for both subject-verb-object (SVO) and verb-subject-object (VSO) constructions. Additionally, both Afrikaans and Filipino utilize the Latin writing system, albeit with distinct orthographic conventions and phonetic representations. Furthermore, they share the use of affixes to denote verb tense and lack subject-verb agreement (Lewis, 2009; Comrie, 1989). While Afrikaans does exhibit some cognates with Malay, another Austronesian language akin to Filipino, these similarities are still insufficient to claim a structural relationship.

In contrast, Table 4 shows that RoBERTa-Tagalog's top performers are English and Naija. English, as a global lingua franca, shares a rich history with Filipino, likely resulting in lexical borrowings and syntactic influences. Similarly, Naija/Nigerian Pidgin, though distinct, shares linguistic features with English, particularly simplified verb conjugation systems (Lewis, 2009; Comrie, 1989).

Despite these similarities, descriptive observa-

tions alone are insufficient to suggest a meaningful structural connection between Filipino and the source languages. The similarities are also not easily generalizable with the other top-performing source languages. Therefore, an examination of quantitative linguistic distances is crucial for optimal source language selection.

4.1 Analysis of Language Similarity Metrics

Correlation analysis was conducted to investigate the relationship between the zero-shot cross-lingual transfer F1 scores of XLM-R and RoBERTa-Tagalog models and various linguistic similarity distances. Pearson's correlation coefficients and their corresponding p-values were calculated to assess the strength and significance of these relationships.

Table 5: Correlation analysis for XLM-R with various linguistic distances

Distances	ρ	**p-value**
Featural	-0.319	0.005
Genetic	-0.089	0.448
Geographic	0.106	0.365
Inventory	-0.236	0.042
Phonological	-0.106	0.368
Syntactic	-0.365	0.001

Table 6: Correlation analysis for RoBERTa-Tagalog with various linguistic distances

Distances	ρ	**p-value**
Featural	-0.233	0.044
Genetic	-0.094	0.421
Geographic	0.304	0.008
Inventory	-0.316	0.006
Phonological	-0.138	0.237
Syntactic	-0.204	0.079

The analysis revealed a relationship between linguistic similarity and the zero-shot cross-lingual transfer performance of both models. Negative correlations, typically between -0.2 and -0.3, were observed with featural, inventory, and syntactic distances. This suggests that as these distances increase, indicating that languages are becoming less similar, the cross-lingual performance of both models tends to decline. These correlations were statistically significant, with p-values below 0.05. Notably, RoBERTa-Tagalog exhibited a weak but statistically significant positive correlation (0.304) with geographic distance, while this correlation for XLM-R was not significant. The genetic and phonological correlations with both models were weaker and not statistically significant.

These findings highlight the importance of considering linguistic similarity when choosing source languages for zero-shot transfer learning. Languages with closer features, inventory, and syntax tend to show better transfer performance for both XLM-R and RoBERTa-Tagalog. Interestingly, RoBERTa-Tagalog seems to benefit, to some extent, from geographic proximity, although higher performance is observed with source languages farther apart from Filipino.

Understanding which linguistic distances significantly correlate with cross-lingual transfer performance is strategic for source language selection. This can be done by prioritizing languages with favorable distances that positively impact transfer learning success.

4.2 Impact of PLM Selection

The experiments highlight the importance of PLM selection in influencing the performance of cross-lingual transfer learning. Since the target language in this study is Filipino, it might be reasonable to expect that RoBERTa-Tagalog would perform competitively. However, the results show that XLM-R outperforms RoBERTa-Tagalog based on F1 scores.

The superior performance of XLM-R may be due to the fact that while RoBERTa-Tagalog is specifically tailored for Tagalog, XLM-R's multilingual pretraining exposed it to a wider range of languages. This diversity of languages enabled XLM-R to recognize a greater variety of linguistic patterns. The architecture of XLM-R may have provided it with a stronger ability to adapt to new languages compared to RoBERTa-Tagalog.

Moreover, there is a notable difference in the top 10 source languages between XLM-R and RoBERTa-Tagalog. This divergence likely reflects how each model adapted distinct linguistic information during fine-tuning, which influenced their performance in transferring knowledge to a new language. Despite RoBERTa-Tagalog's specialization for Tagalog, the specific linguistic characteristics that XLM-R excelled with may not have optimally aligned with Tagalog's features, leading to its lower performance.

4.3 Investigating Multilingual Source Languages

This study also investigated the implementation of a multilingual source language approach for both PLMs. The methodology employed a progressive strategy, beginning with the single best-performing source language and sequentially including additional languages from the top ten performers into the training dataset.

This approach helped us isolate the impact of each additional language on POS tagging performance. Sequentially adding languages can be seen as a blocking strategy akin to curriculum learning (Lee et al., 2023). However, this top-down approach may not always be optimal. Selecting examples and their order can significantly accelerate learning in curriculum learning (Bengio et al., 2009). In this study, we use the monolingual performance of source languages as a measure of how easy it is for the model to "learn" a language.

While the multilingual source language approach did not surpass the highest F1 score achieved by monolingual source training, the results demonstrate promising performance. This setup suggests the potential benefits of simultaneously learning from multiple languages, which allows for the learning of diverse linguistic patterns and structures. Notably, adding more and more languages did not lead to drastic changes in performance. For both XLM-R and RoBERTa-Tagalog, multilingual source training achieved F1 scores in the range of 70% to 80%.

Table 7: F1 scores of XLM-R and RoBERTa-Tagalog with multilingual source languages (top-down approach)

Combination	XLM-R	RoBERTa-Tagalog
1 language	79.10	71.63
2 languages	79.06	71.08
3 languages	76.14	74.49
4 languages	77.55	75.68
5 languages	76.33	73.11

We also tested a random addition of source languages instead of the top-down approach starting from the top source language in terms of performance. We observed that systematically adding sources is slightly better, but the difference is not substantial. At this point, the difference between the two approaches is minimal. Therefore, other approaches can be experimented with in the future.

Table 8: F1 scores of XLM-R and RoBERTa-Tagalog with multilingual source languages (random addition)

Combination	XLM-R	RoBERTa-Tagalog
1 language	79.10	71.63
2 languages	75.58	73.99
3 languages	75.29	71.91
4 languages	77.55	72.97
5 languages	79.01	72.51

4.4 PLM Tokenization

Although zero-shot learning using PLMs has shown promising results for Filipino POS tagging, one main challenge in refining PLMs is optimizing tokenizers. These tokenizers are often inadequate when confronted with previously unseen data variations (Blaschke et al., 2023). This issue is evident when Filipino input texts make model output erroneous parsing, automatically causing incorrect tags.

For instance, upon analyzing the tokenization of the sample input sentence "Tila ang bango ng bulaklak dahil napapikit siya at napangiti." using the RoBERTa-Tagalog model trained on English, an instance of incorrect tokenization was observed. Specifically, the word "napapikit" was split into "napapik" and "it," mistakenly labeled as a verb and adjective, rather than recognizing its actual function as a verb alone.

In another example sentence, "Sa pagpataw ng suspension laban sa Noveras, inamin naman ng Ombudsman na walang matibay na ebidensiya," tokens are incorrectly split and merged. "Sa pagpataw" should be split into "Sa" (adposition) and "pagpataw" (noun), but they have been tokenized as "sa pagp" and "ataw," due to the model's limited exposure to variations in Filipino text. These tokenization errors indicate a lack of sensitivity to the morphological structure of Filipino words. Note that similar problems occur with other source languages and with the XLM-R model.

Despite linguistic similarities from the source languages, Filipino text tokenization using PLMs sometimes fails to align with meaningful representations, leading to poor performance in POS tagging. These errors in tokenization indicate limitations in processing the linguistic nuances of Filipino text.

Another note is that there is variability in the fertility scores across different languages when evaluated. The average tokenizer fertility for each

training dataset is reported in Appendix C. This variability suggests the importance of using controlled training data to achieve reliable model performance across languages, as it can significantly affect the performance of the source languages. Future works should consider these variations when selecting and preparing datasets for transfer learning tasks, as they may have an impact on model training and evaluation.

5 Conclusion

This study implements zero-shot fine-tuning using PLMs for Filipino POS tagging, exploring the role of linguistic distances in source language selection. Correlation analysis between linguistic similarity distances and PLM performance suggests that featural, inventory, and syntactic distances between source languages and Filipino, impact cross-lingual transfer learning outcomes.

The study also explored the role of PLM selection in influencing cross-lingual transfer learning performance. While RoBERTa-Tagalog is specifically designed for Tagalog, the multilingual language model XLM-R outperformed it. Furthermore, the exploration of a multilingual source language approach shows good results, though slightly lower than monolingual fine-tuning, suggesting potential benefits of using multiple languages simultaneously for cross-lingual transfer learning tasks.

Despite promising results, challenges in tokenization were observed, particularly in accurately tokenizing Filipino text. Errors in tokenization underscore the need for improved tokenization processes for PLMs, especially for under-resourced languages like Filipino.

Future research should address these challenges by creating new treebanks and expanding existing ones to further enhance model performance. Using top-performing models from this study to annotate unannotated datasets can serve as a foundation for future researches. These annotations, once manually refined, can produce gold-standard annotations for improved training and evaluation of NLP models.

Acknowledgments

We wish to express our sincere gratitude to those who provided support and guidance throughout the development of this paper. We would like to thank the Ateneo Social Computing Lab for providing a supportive academic environment for conducting this research.

References

Angelina Aquino and Franz de Leon. 2020. Parsing in the absence of related languages: Evaluating low-resource dependency parsers on Tagalog. In *Proceedings of the Fourth Workshop on Universal Dependencies (UDW 2020)*, pages 8–15.

Angelina Aquino and Franz de Leon. 2022. Zero-shot and few-shot approaches for tokenization, tagging, and dependency parsing of Tagalog text. In *Proceedings of the 36th Pacific Asia Conference on Language, Information and Computation*, pages 190–202.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 41–48.

Verena Blaschke, Hinrich Schütze, and Barbara Plank. 2023. Does manipulating tokenization aid cross-lingual transfer? A study on POS tagging for non-standardized languages. *arXiv preprint arXiv:2304.10158*.

Chris Collins and Richard Kayne. 2009. Syntactic structures of the world's languages. *New York: New York University*.

Bernard Comrie. 1989. *Language universals and linguistic typology: Syntax and morphology*. University of Chicago press.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Alexis Conneau, Guillaume Lample, Ruty Rinott, Adina Williams, Samuel R Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. *arXiv preprint arXiv:1809.05053*.

Jan Christian Blaise Cruz and Charibeth Cheng. 2021. Improving large-scale language models and resources for Filipino. *arXiv preprint arXiv:2111.06053*.

Marie-Catherine De Marneffe, Christopher D Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal dependencies. *Computational linguistics*, 47(2):255–308.

Wietse de Vries, Martijn Wieling, and Malvina Nissim. 2022. Make the best of cross-lingual transfer: Evidence from POS tagging with over 100 languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7676–7685, Dublin, Ireland. Association for Computational Linguistics.

Matthew S Dryer and Martin Haspelmath. 2013. The world atlas of language structures online.

Juuso Eronen, Michal Ptaszynski, and Fumito Masui. 2023. Zero-shot cross-lingual transfer language selection using linguistic similarity. *Information Processing & Management*, 60(3):103250.

Harald Hammarström, Robert Forkel, and Martin Haspelmath. 2018. Glottolog 3.0. *Max Planck Institute for the Science of Human History*.

Ran Iwamoto, Hiroshi Kanayama, Alexandre Rademaker, and Takuya Ohko. 2021. A universal dependencies corpora maintenance methodology using downstream application. In *Proceedings of the Third Workshop on Computational Typology and Multilingual NLP*, pages 23–31.

Joo-Kyung Kim, Young-Bum Kim, Ruhi Sarikaya, and Eric Fosler-Lussier. 2017. Cross-lingual transfer learning for POS tagging without cross-lingual resources. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2832–2838, Copenhagen, Denmark. Association for Computational Linguistics.

Bruce W Lee, Hyunsoo Cho, and Kang Min Yoo. 2023. Instruction tuning with human curriculum. *arXiv preprint arXiv:2310.09518*.

Paul Lewis. 2009. Ethnologue: Languages of the world.

Patrick Littell, David R Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. 2017. Uriel and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, volume 2, pages 8–14.

LJ Miranda. 2023. Towards a Tagalog NLP pipeline.

Steven Moran and Daniel McCloy. 2019. Phoible 2.0. *Jena: Max Planck Institute for the Science of Human History*.

Fred Philippy, Siwen Guo, and Shohreh Haddadan. 2023. Identifying the correlation between language distance and cross-lingual transfer in a multilingual representation space. *arXiv preprint arXiv:2305.02151*.

Xipeng Qiu, Tianxiang Sun, Yige Xu, Yunfan Shao, Ning Dai, and Xuanjing Huang. 2020. Pre-trained models for natural language processing: A survey. *Science China Technological Sciences*, 63(10):1872–1897.

Edward P Stabler and Edward L Keenan. 2003. Structural similarity within and among languages. *Theoretical Computer Science*, 293(2):345–363.

A Appendix: Linguistic Distances from Filipino of the Top Performing Source Languages for XLM-R

Table 9: Linguistic distances from Filipino for the top 10 performing source languages, as determined by the XLM-R model's F1 score.

Lang	Fea	Gen	Geo	Inv	Pho	Synt
afr	0.63	1	0.54	0.49	0.59	0.75
heb	0.57	1	0.44	0.52	0.59	0.53
bul	0.55	1	0.47	0.55	0.36	0.60
vie	0.54	1	0.08	0.49	0.39	0.64
nor	0.80	1	0.49	0.66	0.59	0.68
urd	0.63	1	0.29	0.49	0.59	0.76
ces	0.62	1	0.50	0.47	0.59	0.72
pes	0.54	1	0.35	0.45	0.41	0.68
fao	0.80	1	0.52	0.66	0.59	0.68
eng	0.53	1	0.54	0.46	0.34	0.66

B Appendix: Linguistic Distances from Filipino of the Top Performing Source Languages for RoBERTa-Tagalog

Table 10: Linguistic distances from Filipino of the top 10 performing source languages, as determined by the RoBERTa-Tagalog model's F1 score.

Lang	Fea	Gen	Geo	Inv	Pho	Synt
eng	0.53	1	0.54	0.46	0.34	0.66
pcm	0.64	1	0.63	0.43	0.59	0.59
srp	0.78	1	0.48	0.66	0.86	0.65
glv	0.86	1	0.54	0.66	0.59	0.78
slv	0.58	1	0.51	0.47	0.59	0.63
spa	0.50	1	0.58	0.46	0.51	0.53
nld	0.63	1	0.52	0.53	0.59	0.71
hrv	0.65	1	0.50	0.46	0.59	0.89
pol	0.49	1	0.48	0.44	0.36	0.58
gle	0.53	1	0.56	0.45	0.59	0.54

C Appendix: Fertility Scores for UD Training Datasets

Table 11: Fertility scores for the training datasets of UD using XLM-R and RoBERTa-Tagalog as tokenizers (Part 1 of 2).

Language	XLM-R	RoBERTa-Tagalog
af	1.54	2.22
ar	1.13	2.58
be	2.16	6.17
bg	1.54	5.41
bxr	2.44	6.36
ca	1.38	1.93
cop	1.96	10.26
cs	1.72	3.33
cu	3.12	7.28
cy	1.56	2.38
da	1.47	2.34
de	1.56	2.68
el	1.65	9.24
en	1.32	1.63
es	1.34	1.94
et	1.82	2.83
eu	1.78	2.62
fa	1.36	6.60
fi	1.91	3.27
fo	1.58	2.25
fr	1.44	2.04
gd	1.67	2.26
gl	1.31	2.00
got	2.25	2.98
grc	3.27	10.36
gv	1.85	1.97
hbo	4.99	9.96
hi	1.30	8.49
hr	1.58	2.81
hsb	2.27	3.33
hu	1.75	3.41
hy	1.85	9.72
hyw	2.35	9.85
id	1.39	2.33
is	1.58	2.87
it	1.41	2.01

Table 12: Fertility scores for the training datasets of UD using XLM-R and RoBERTa-Tagalog as tokenizers (Part 2 of 2).

Language	XLM-R	RoBERTa-Tagalog
ja	1.20	1.49
kk	1.87	5.91
kmr	1.65	3.02
ko	2.12	8.11
koi	2.49	5.13
kpv	2.66	5.59
ky	1.83	7.12
la	1.61	2.22
lij	1.59	1.89
lt	1.82	3.32
lzh	1.96	3.06
mdf	2.35	5.13
mr	1.68	8.59
mt	2.29	2.77
myv	2.54	5.64
nl	1.48	2.23
no	1.48	2.36
olo	1.93	2.62
orv	2.44	5.77
pcm	1.22	1.41
pl	1.74	3.25
pt	1.38	2.08
qaf	1.93	2.30
qpm	2.03	2.68
qtd	1.40	2.45
ro	1.68	2.69
ru	1.63	5.68
sa	2.73	4.33
sk	1.75	2.84
sl	1.58	2.53
sme	2.55	3.25
sms	3.11	4.41
sr	1.60	2.73
sv	1.49	2.56
ta	2.10	20.86
te	1.94	13.50
tr	1.89	3.46
ug	2.19	9.77
uk	1.74	5.56
ur	1.32	6.28
vi	1.44	3.89
wo	1.81	2.05
zh	2.09	4.21