Untangling the Influence of Typology, Data and Model Architecture on Ranking Transfer Languages for Cross-Lingual POS Tagging

Enora Rice[1] **Ali Marashian**[1] **Hannah Haynie**[1] **Katharina von der Wense**[1,2]
Alexis Palmer[1]
[1]University of Colorado Boulder [2] Johannes Gutenberg University Mainz
enora.rice@colorado.edu

Abstract

Cross-lingual transfer learning is an invaluable tool for overcoming data scarcity, yet selecting a suitable transfer language remains a challenge. The precise roles of linguistic typology, training data, and model architecture in transfer language choice are not fully understood. We take a holistic approach, examining how both dataset-specific and fine-grained typological features influence transfer language selection for part-of-speech tagging, considering two different sources for morphosyntactic features. While previous work examines these dynamics in the context of bilingual biLSTMS, we extend our analysis to a more modern transfer learning pipeline: zero-shot prediction with **pretrained multilingual models**. We train a series of transfer language ranking systems and examine how different feature inputs influence ranker performance across architectures. Word overlap, type-token ratio, and genealogical distance emerge as top features across all architectures. Our findings reveal that a combination of **typological** and **dataset-dependent** features leads to the best rankings, and that good performance can be obtained with either feature group on its own.

1 Introduction

Despite being trained on 100+ languages, pretrained multilingual language models (MLMs) fail to cover the vast majority of the world's languages. Finetuning MLMs for zero-shot cross-lingual transfer is a useful technique to extend their reach by circumventing the lack of task-specific labeled data in low-resource languages. Effective zero-shot transfer hinges on choosing an appropriate source language (Eronen et al., 2023, 2022; Layacan et al., 2024), but it is still not well understood how to make this selection. Most analyses of successful source/target pairs fall into one of two categories: typological or dataset-dependent. The typological view investigates the role of linguistic similarity, with studies showing that more "similar" languages tend to form better source/target pairs (Eronen et al., 2023; de Vries et al., 2022; Lauscher et al., 2020). Much of this typological analysis is coarse-grained, focusing on features like language family or abstract distance measures. The dataset-dependent view focuses on comparing source and target datasets based on features like sub-word overlap (Wu and Dredze, 2019; Pires et al., 2019; K et al., 2020). Few papers consider both views, and those that do focus on older methods of crosslingual transfer like bilingual LSTMS (Lin et al., 2019). Additionally, previous analyses shed little light on the linguistic question of which fine-grained typological features are especially relevant for the task.

This primary goal of this paper is to offer a deeper understanding of effective transfer language selection across architectures, comparing crosslingual transfer with biLSTMs to XLM-R (Conneau et al., 2020) and M-BERT (Devlin et al., 2019). We aim to identify which features contribute to selecting a successful source/target pair for part-of-speech (POS) tagging. We focus on POS tagging because it directly reflects typological features such as word order. Our analysis addresses the following key questions:

Q1. Which features are most important for cross-lingual transfer?

Q2. Do these features differ between biLSTMs and MLMs?

Q3. How does the granularity of typological features—whether fine or coarse—affect transfer language selection?

Q4. Is it necessary to consider data set features in selecting a transfer language?

We train a series of gradient-boosted decision tree models to rank transfer languages for POS

tagging, with separate rankers for the two architectures. During training, we generate feature importance scores and identify the most salient features for each architecture (Q1, Q2). To examine the role of fine-grained typological features, we compare two typological inputs: source/target distance measures, and full finegrained feature vectors (Q3). We also evaluate how the source and quality of typological data affects ranker performance by swapping between URIEL (Littell et al., 2017) and Grambank (Skirgård et al., 2023a) feature vectors. Last, we investigate whether typological information alone can effectively determine suitable source/target language pairs by experimenting with the exclusion of dataset-specific features (Q4).

We find that impressive performance can be achieved when relying primarily on either feature category, without the need for the other, indicating that both "typological" and "dataset-dependent" views of transfer language choice represent independently viable strategies. However, peak performance is achieved by combining dataset-dependent and fine-grained typological features. Crucially, our analysis reveals that key features such as word overlap, type-token ratio, and genealogical distance remain consistently important across architectures, suggesting that the relevance of these features may transcend specific model designs, offering broader insights into cross-lingual transfer that could enable us to better leverage MLMs for low-resource applications.

2 Related Works

2.1 Ranking Transfer Languages

Lin et al. (2019) rank transfer languages using both dataset-dependent and linguistic features from the URIEL knowledge base (Littell et al., 2017). We build on their work with key adaptations: 1) Instead of varying dataset size, which obscures the role of fine-grained features, we hold corpus size constant across all language pairs. 2) In addition to bilingual biLSTMs, we examine zero-shot transfer with finetuned MLMs. 3) We replace typological distance measures with element-wise comparisons of typological feature vectors, following Dolicki and Spanakis (2021).

Khan et al. (2025) build on the work in Littell et al. (2017) to enhance the coverage of URIEL and lang2vec with novel linguistic databases and customizable distance calculations. We follow suit by comparing the impact of incorporating URIEL syntactic vectors versus Grambank syntactic vectors on the transfer language ranking task

2.2 Transfer Language Choice for Zero-shot Cross-lingual Transfer with MLMs

Lauscher et al. (2020) show a correlation between linguistic proximity and successful zero-shot transfer, but only test English as the source language. We experiment with 18 source languages. de Vries et al. (2022) find that XLM-R finetuned on a suitable transfer language performs almost three times better than when using a suboptimal transfer language. They highlight the influence of linguistic similarity but do not consider dataset features.

3 Experiments

3.1 Languages

We experiment with a total of 20 target and 18 source languages across seven language families. We determine our set of target and source languages based on the availability of sufficient data in Universal Dependencies 2.0 (UD) (de Marneffe et al., 2021). We consider target languages that have a training corpus with at least 500 lines and source languages with at least 2000. Justification for this threshold is described in 3.2.1. We also eliminate languages that are not present in URIEL and/or Grambank. Our full set of target languages is given in Table 1. Languages that also serve as source languages are italicized. While many of the languages covered by our experiments are high-resource, several others fall into a middle range and are underserved by the NLP research community at large.

3.2 Testbed Tasks

We generate gold ranking-data by training a suite of biLSTMs and finetuned XLM-R and M-BERT models for POS tagging across all possible source/target language pairs. To remove the influence of dataset size, we cap each source language training set at 2000 lines. Then, for each target language, we create a ranking of all potential source languages based on the relative performance of each model on a held out test set. Model details are outlined in following sections.

3.2.1 biLSTMs

We train a suite of 378 biLSTMs using Stanza (Qi et al., 2020)– one for each target/source pair. We train each model on 500 instances of UD data in the target language and 2000 instances in the source

language. We choose this split to simulate a setting where limited training data is available in the target language but comparatively greater data is available in the source language. We set the data thresholds to ensure that sufficient training data is present for model convergence, but training data in the target language is still limited enough to make the task non-trivial. All models are trained on default Stanza hyperparameters *without* pre-trained word embeddings for a maximum of 6000 steps. We evaluate each model on a held out test set drawn from the same corpus as the target training data.

3.2.2 Fine-tuned XLM-R and M-BERT

We finetune XLM-R and M-BERT equivalently on each of our 18 source languages with a modified implementation[1] from de Vries et al. (2022). Each model is trained on the same 2000 instance UD dataset that we use to train our biLSTM models. All models are trained for 1,000 batches of 10 samples with a linearly decreasing learning rate starting at 5e-5. We use 10% dropout between transformer layers and 10% self-attention dropout.

Language	**Treebank**
Basque	UD_Basque-BDT
Czech	UD_Czech-PDT
Danish	UD_Danish-DDT
Dutch	UD_Dutch-LassySmall
Finnish	UD_Finnish-FTB
Hindi	UD_Hindi-HDTB
Hungarian	UD_Hungarian-Szeged
Indonesian	UD_Indonesian-GSD
Galician	UD_Galician-CTG
Italian	UD_Italian-PoSTWITA
Korean	UD_Korean-GSD
Latin	UD_Latin-ITTB
Latvian	UD_Latvian-LVTB
Turkish	UD_Turkish-IMST
Polish	UD_Polish-LFG
Portuguese	UD_Portuguese-Bosque
Russian	UD_Russian-SynTagRus
Catalan	UD_Catalan-AnCora
French	UD_French-Sequoia
English	UD_English-LinES
Ukrainian	UD_Ukrainian-IU

Table 1: Full list of target languages and their corresponding treebanks. Languages that also serve as source languages are italicized.

[1]https://github.com/wietsedv/xpos

3.3 Our Ranking System

Given a target language t and a list of n potential source languages $S = [s_1, s_2...s_n]$, our goal is to rank all source languages in S based on the expected performance of POS-tagging models trained on each source/target pair (s_i,t). Building on Lin et al. (2019), we train a series of gradient boosted decision trees using the LightGBM implementation (MIT License) (Ke et al., 2017) of the LambdaRank algorithm. Models are trained on gold ranking-data described in Section 3.2.

Input to our ranking system consists of vector representations of each source/target pair. Vectors are defined as a set of features, categorized into two types. We calculate **dataset-dependent** features by comparing source and target corpora using four metrics: word overlap, type-token ratio in the source language corpus, type-token ratio in the target language corpus, and the difference between the source and target language type-token ratios. **Dataset-independent** features capture linguistic similarity between the source and target languages using five measures: *genetic*, *syntactic*, *phonological*, (phonetic) *inventory*, and *geographic*. *Syntactic*, *phonological* and *inventory* features are defined using binary feature vectors sourced from typological databases. We call these our *Typology-Vector* features. By default, Typology-Vector features are represented by distance measures computed as the cosine difference between URIEL (Littell et al., 2017) vectors representing source and target, but we experiment with different representations (described in Sections 3.3.1 and 3.3.2). All features are briefly summarized in Table 2 and feature vector lengths are given in Table 3. For more detailed descriptions, refer to Lin et al. (2019).

3.3.1 Distance-Measure vs. Fully Featured

By default, we express the linguistic similarity between *syntactic*, *phonological*, and *inventory* features as a series of distance measures. We call these **distance** Typology-Vector representations. At predict time, the ranker receives a feature vector a representing the target and a feature vector b representing the source and computes the cosine distance: $1 - cos(a, b) = d$. We concatenate d to the final ranking model input vector.

To analyze the impact of fine-grained features on transfer language suitability, we experiment with an expanded representation, using an element-wise *and* operation to compare a and b: $a \wedge b = v$. We refer to v as the **full** Typology-Vector representation.

Feature Type	Description
Genetic Distance	Genealogical distance derived from language descent trees described in Glottolog.
Geographic Distance	Defined as the orthodromic distance divided by the antipodal distance between rough locations of source and target languages on the surface of the Earth.
Syntactic, *Phonological*, and *Inventory* Distances (**distance** Typology-Vector)	Computed as the cosine difference between corresponding URIEL (Littell et al., 2017) or Grambank (Skirgård et al., 2023a) feature vectors representing source and target languages.
Syntactic, *Phonological*, and *Inventory* Vectors (**full** Typology-Vector)	Computed as element-wise AND operation between corresponding URIEL (Littell et al., 2017) or Grambank (Skirgård et al., 2023a) feature vectors representing source and target languages.
Dataset-Dependent Features	Word overlap, transfer type-token ration, source type-token ration, type-token ratio distance

Table 2: All possible ranker features

Vector Type	Description
URIEL Syntactic	104
Grambank Syntactic	113
Phonological	28
Inventory	158

Table 3: Typological feature vector lengths

We concatenate v to ranker input.

3.3.2 URIEL vs. Grambank

Many typological analyses of crosslingual transfer rely on URIEL (CC BY-SA 4.0) feature vectors, which are heavily based on the World Atlas of Language Structures (CC BY 4.0) (Dryer and Haspelmath, 2013). WALS has incomplete genealogical coverage and over 80% missing data (Skirgård et al., 2023a). As such, we experiment with switching to Grambank (CC BY 4.0) (Skirgård et al., 2023a), which addresses some of WALS' shortcomings. We impute all undefined features in either database as follows.

URIEL. We use URIEL vectors that have been pre-imputed by Littell et al. (2017) using k-nearest-neighbors.[2]

Grambank. 24% of total feature values in Grambank 1.0.3 (across all languages in the database) are undefined. In order to produce fully defined feature vectors for our experiments, we first eliminate any features that are undefined for greater than 25% of languages and any languages that have greater than 25% missing data. After cropping, only 4.03% of values are missing. We impute the remaining values with the MissForest algorithm for nonparametric missing value imputation (Stekhoven and Bühlmann, 2012). We adapt our imputation procedure from Skirgård et al. (2023b).

[2]vectors available at https://github.com/antonisa/lang2vec

3.3.3 Dataset Features

We experiment with the inclusion and exclusion of dataset dependent features to assess the impact the training corpus might have on successful crosslingual transfer. We control for training corpus size in our gold rankings, but we do not control for any other corpus features across source languages. Therefore, it is necessary to evaluate the relevance of features like type-token ratio and word overlap.

3.3.4 Evaluation

As in Lin et al. (2019), we evaluate our ranking models with leave-one-out cross-validation. For each cross-validation fold, we exclude one target language from our test set of n languages, and train our ranking model using gold transfer language rankings for each $n-1$ remaining languages. We then evaluate the model's performance on the held-out language. We evaluate our ranking models using Normalized Distributed Cumulative Gain (NCDG)(Järvelin and Kekäläinen, 2002).

Specifically, we use NCDG@p, a metric that considers the top-p elements, which is defined by:

$$NDCG@p = \frac{DCG@p}{IDCG@p},$$

where the Discounted Cumulative Gain (DCG) at position p is defined as

$$DCG@p = \sum_{i=1}^{p} \frac{2^{\gamma_i} - 1}{log_2(i+1)}.$$

γ_i is a relevance score corresponding to the language at position i of the predicted ranking that we are evaluating. For all $i \leq p$, $\gamma_i = p - i$, where p represents the number of ranked items we wish to assign relevance. We set $p = 5$, meaning that the true best transfer language has a relevance score of $\gamma = 5$. All languages below the top-5 are assigned

Syntactic Feature-Src		Dataset Features	Typology-Vector Representation	NDCG@5 *biLSTMs*	*XLM-R*	M-BERT
URIEL	**a**	✓	distance	**0.799**	0.755	0.654
	b	-	distance	0.385	0.643	0.625
	c	✓	full	0.776	**0.782**	0.680
	d	-	full	0.721	0.670	**0.689**
	Avg			0.670	0.713	0.662
Grambank	**a**	✓	distance	0.768	0.826	0.653
	b	-	distance	0.447	0.574	0.638
	c	✓	full	**0.788**	**0.827**	0.665
	d	-	full	0.721	0.707	**0.692**
	Avg			0.681	0.734	0.662
Avg ***(std)***				0.676 *(0.153)*	0.723 *(0.085)*	0.662 *(0.023)*

Table 4: Average NDCG@5 for all model configurations trained on gold rankings. Every model configuration includes *genetic* and *geographic* features.

$\gamma = 0$. The Ideal Discounted Cumulative Gain (IDCG) is calculated the same as DCG except it is calculated over the gold-standard ranking. An NCDG@p of 1 indicates that the top-p predicted elements match the top-p gold elements exactly. We report the average NDCG@5 across all N leave-one-out models.

3.4 Analyzing Feature Importance

To compare the most relevant features for transfer in POS tagging across architectures, we use our most full featured ranking model, incorporating dataset-dependent features, *syntactic* features from Grambank, and **full** Typology-Vectors. We train three rankers, one for each architecture. During training, each feature is assigned an importance score based on the gain resulting from splits made on that feature. For a given split, we calculate gain as the reduction in squared error from the parent node to the child nodes, summed across all trees in the ranking model. We report average gain over all cross-validation folds and identify the top-5 most important features for each model.

4 Results

4.1 Dataset vs. Typological Features

In Table 4, we observe that regardless of *syntactic* vector source, models trained with **distance** Typology-Vector representations and *without* dataset features (setting **b**) perform relatively poorly. This suggests that coarse grained information from **distance** Typology-Vector representations may not be sufficient for choosing a transfer language. However, when we replace **distance** Typology-Vector representations with **full**, performance increases substantially. On average, NDCG@5 jumps by 0.148 between settings **b** and **d** over all 6 architecture/feature-source pairings. The performance gains from including dataset features are even more significant. On average, NDCG@5 jumps by 0.19 between settings **b** and **a**.

These findings suggest that both fine-grained typological features *and* dataset-dependent features support more accurate transfer language ranking. Both feature sources provide meaningful signals to the ranker, but setting **c** results in the best average ranker performance, suggesting that an integrated view of transfer language choice is most effective.

M-BERT stands out as a notable outlier, as setting **d** produces the highest-performing M-BERT rankers. It is unclear why excluding dataset features benefits transfer language ranking for M-BERT. However, it is noteworthy that M-BERT exhibits by far the lowest standard deviation in performance, suggesting its rankers are less sensitive to variations in feature configuration. We leave further analysis of this phenomenon to future work.

4.2 Grambank vs. URIEL

Rankers leveraging Grambank *syntactic* features outperform those trained with URIEL *syntactic* features in ranking biLSTMs and XLM-R on average, suggesting that the typological information captured by Grambank may be more informative for selecting a transfer language. However, M-BERT is yet again an outlier– on average, M-BERT rankers perform equivalently regardless of *syntactic* feature-source.

XLM-R		M-BERT		BiLISTM	
Feature	**Gain**	**Feature**	**Gain**	**Feature**	**Gain**
genetic	272.95	*genetic*	283.41	*word_overlap*	264.24
word_overlap	102.82	*word_overlap*	130.90	*transfer_ttr*	118.17
transfer_ttr	67.60	*transfer_ttr*	42.49	*genetic*	100.78
distance_ttr	25.74	*distance_ttr*	24.67	*distance_ttr*	12.66
GB093	11.96	*task_ttr*	10.06	*INV_VOW_10_MORE*	7.90
Standard Deviation	17.08		17.96		17.42

Table 5: Feature importance for top-5 features by model for ranker trained *with* dataset features and full Grambank vectors.

4.3 Feature Importance

We investigate feature importance within our most fully-featured ranking model, which incorporates dataset-dependent features, *syntactic* features from Grambank, and full Typology-Vectors. Though this is not always the highest performing setting, it enables us to elucidate the interplay between the dataset-dependent and typological features most clearly. We identify the top-5 most important features for each of our models in Table 5. Four out of five features are shared across architectures: genetic, word_overlap, transfer_ttr, and distance_ttr. Notably, these are primarily dataset-dependent features. This consistency in relative feature importance across models suggests that the features that determine a suitable transfer language choice may not be architecture-dependent. On the other hand, it is interesting that *genetic* is most important for XLM-R and M-BERT but not for biLSTMs. It is possible that the shared representation space built during multilingual pretraining already contains features like word-overlap making them less relevant for selecting a finetuning dataset.

Feature	**Gain**
XLM-R	
genetic	362.93
GB020	11.62
GB080	8.90
GB093	7.68
INV_OPEN_FRONT_UNROUNDED_VOWEL	7.48
Standard Deviation	20.93
M-BERT	
genetic	407.08
GB022	8.44
GB093	7.07
INV_PALATAL_LATERAL_APPROXIMANT	6.42
GB020	6.39
GB114	5.32
Standard Deviation	23.46
biLSTM	
genetic	342.61
INV_OPEN_MID_CENTRAL_UNROUNDED_VOWEL	21.75
GB172	19.12
INV_MID_CENTRAL_UNROUNDED_VOWEL	17.66
INV_LABIODENTAL_NASAL	12.22
Standard Deviation	19.83

Table 6: Feature importance for rankers Trained with full Grambank vectors and *without* dataset features

5 Supplementary Analyses

5.1 Excluding Dataset Features

For the sake of comparison, we also analyze the top-5 features for a ranking model trained with *syntactic* features from Grambank and **full** Typology-Vectors *without* dataset-dependent features. These rankers do not consistently underperform their dataset-dependent counterparts, raising the question of which dataset-independent features carry the most weight.

Looking at Table 6, we find that the *genetic* feature yields substantially more gain than any other feature. It is possible that *genetic* scores so highly because it serves as a proxy for many of the other features. This intuition is supported by Skirgård et al. (2023a), who show that phylogenetic relationships explain a majority of the variance in all but a few Grambank features.

Other than *genetic*, M-BERT and XLM-R seem to share more top features with each other than with biLSTMs– GB093 and GB020 both ranking highly. However, this does not necessarily indicate a meaningful difference between the architectures. Excluding *genetic*, gain is relatively low and consistent across features. This finding suggests that it may not be possible to identify especially salient fine-grained features, because relevance is distributed over the full feature set. In a sense, the

Src/Tgt	XLM-R Rank	BiLSTM Rank	Diff.	Src/Tgt	XLM-R Rank	BiLSTM Rank	Diff.
eus/cat	354	22	332	ukr/pol	10	339	329
kor/cat	360	29	331	ces/pol	8	302	294
kor/glg	339	13	326	rus/pol	32	324	292
kor/fra	359	54	305	dan/fin	66	345	279
pol/cat	323	24	299	rus/lav	26	304	278
eus/glg	301	12	289	lav/pol	86	337	251
eus/fra	334	46	288	eng/fin	96	347	251
pol/fra	331	49	282	ces/rus	15	258	243
tur/cat	305	27	278	ukr/lav	56	297	241
pol/glg	282	7	275	fra/fin	108	348	240

Table 7: Greatest difference in relative performance differences between XLM-R and biLSTM. Better biLSTM performance (left) vs. better XLM-R performance (right).

XLM-R		biLSTM	
Language Family Pair	**Count**	**Language Family Pair**	**Count**
Indo-European/Indo-European	125	Basque/Indo-European	13
Indo-European/Uralic	14	Koreanic/Indo-European	14
Austronesian/Indo-European	5	Indo-European/Indo-European	71
Basque/Uralic	1	Turkic/Indo-European	12
Turkic/Uralic	1	Koreanic/Uralic	1
Austronesian/Uralic	1	Koreanic/Austronesian	1
Indo-European/Turkic	14	Indo-European/Uralic	14
Indo-European/Basque	6	Basque/Uralic	1
Turkic/Indo-European	3	Indo-European/Koreanic	14
Basque/Indo-European	2	Indo-European/Austronesian	14
Koreanic/Uralic	1	Turkic/Austronesian	1
Austronesian/Turkic	1	Turkic/Koreanic	1
Basque/Turkic	1	Basque/Austronesian	1
Koreanic/Indo-European	1	Austronesian/Uralic	1
Koreanic/Turkic	1	Austronesian/Indo-European	10
		Austronesian/Koreanic	1
		Basque/Koreanic	1
		Koreanic/Basque	1
		Turkic/Basque	1
		Indo-European/Basque	8
		Austronesian/Basque	1
		Turkic/Uralic	1

Table 8: Distribution of language family pairs that ranked relatively higher in XLM-R performance rankings (left) vs. those that ranked relatively higher in biLSTM performance rankings (right)

whole may be greater than the sum of its parts.

5.2 Ranking Analysis: BiLSTMs vs. XLM-R

To contextualize our findings, we conducted a comparative analysis of gold transfer language rankings for biLSTMs and XLM-R. For each architecture, we generated an ordered list of source-target pairs based on performance. We then compared rank differences across architectures for each pair. Table 7 highlights the top-10 language pairs with the most divergent rankings.

XLM-R performs best on language pairs within the same family or subfamily, such as Slavic pairs, likely due to better typological alignment. Meanwhile, biLSTMs excel on pairs with weaker genetic ties. To further explore these trends, we counted occurrences of language family pairs where either XLM-R or biLSTM had a relative ranking advantage in Table 8.

We see that XLM-R comparatively excels on

Indo-European/Indo-European pairs, while biLSTMs perform relatively better on unrelated or weakly related pairs. These results align with expectations: XLM-R's zero-shot approach benefits from well-matched transfer pairs, whereas biLSTMs can make effective use of small amounts of target language training data.

6 Conclusion

We find that features such as word overlap, type-token ratio, and genealogical distance are consistently influential in transfer language selection regardless of model architecture; their importance may be somewhat model-agnostic.

Our findings also highlight the crucial role of dataset-dependent features in ranking transfer languages for cross-lingual transfer. Rankers trained with these features outperform those relying solely on coarse-grained typological features.

At the same time, while coarse-grained typological features alone are insufficient, rankers trained with *fine-grained* typological features achieve impressive results even without dataset-dependent features. The most successful ranking performance comes from combining both dataset-dependent and fine-grained typological features, underscoring the value of a comprehensive approach to transfer language selection.

Crucially, these insights enable us to better support languages that are not well-represented in MLM pretraining. By identifying effective transfer languages with interpretable features, we can improve cross-lingual transfer for lower-resource languages, expanding the reach of NLP beyond those languages that benefit from large-scale pretraining.

Limitations

Since the scope of this paper is limited to crosslingual transfer for POS tagging, it would be interesting to explore whether our results are extensible to other tasks. We are also limited in that we consider a set of just 20 target languages, 13 of which are Indo-European. This paper represents a step forward in explaining the dynamics at play in successful crosslingual transfer, but more work is necessary to determine whether our findings generalize across diverse linguistic contexts.

References

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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 (v2020.3)*. Zenodo.

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

Juuso Eronen, Michal Ptaszynski, Fumito Masui, Masaki Arata, Gniewosz Leliwa, and Michal Wroczynski. 2022. Transfer language selection for zero-shot cross-lingual abusive language detection. *Information Processing & Management*, 59(4):102981.

Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of ir techniques. *ACM Trans. Inf. Syst.*, 20(4):422–446.

Karthikeyan K, Zihan Wang, Stephen Mayhew, and Dan Roth. 2020. Cross-lingual ability of multilingual bert: An empirical study. In *International Conference on Learning Representations*.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu.

2017. Lightgbm: A highly efficient gradient boosting decision tree. In *Neural Information Processing Systems*.

Aditya Armaan Khan, Mason Stephen Shipton, David Anugraha, Kaiyao Duan, Phuong H. Hoang, Eric Khiu, A. Seza Doğruöz, and Annie Lee. 2025. URIEL+: Enhancing linguistic inclusion and usability in a typological and multilingual knowledge base. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6937–6952, Abu Dhabi, UAE. Association for Computational Linguistics.

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)*, page 4483–4499, Online. Association for Computational Linguistics.

Jimson Layacan, Isaiah Edri W. Flores, Katrina Tan, Ma. Regina E. Estuar, Jann Montalan, and Marlene M. De Leon. 2024. Zero-shot cross-lingual POS tagging for Filipino. In *Proceedings of the 3rd Workshop on NLP Applications to Field Linguistics (Field Matters 2024)*, pages 69–77, Bangkok, Thailand. Association for Computational Linguistics.

Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, Antonios Anastasopoulos, Patrick Littell, and Graham Neubig. 2019. Choosing transfer languages for cross-lingual learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3125–3135, Florence, Italy. Association for Computational Linguistics.

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*, pages 8–14, Valencia, Spain. Association for Computational Linguistics.

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.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Hedvig Skirgård, Hannah J. Haynie, Damián E. Blasi, Harald Hammarström, Jeremy Collins, Jay J. Latarche, Jakob Lesage, Tobias Weber, Alena Witzlack-Makarevich, Sam Passmore, Angela Chira, Luke Maurits, Russell Dinnage, Michael Dunn, Ger Reesink, Ruth Singer, Claire Bowern, Patience Epps, Jane Hill, Outi Vesakoski, Martine Robbeets, Noor Karolin Abbas, Daniel Auer, Nancy A. Bakker, Giulia Barbos, Robert D. Borges, Swintha Danielsen, Luise Dorenbusch, Ella Dorn, John Elliott, Giada Falcone, Jana Fischer, Yustinus Ghanggo Ate, Hannah Gibson, Hans-Philipp Göbel, Jemima A. Goodall, Victoria Gruner, Andrew Harvey, Rebekah Hayes, Leonard Heer, Roberto E. Herrera Miranda, Nataliia Hübler, Biu Huntington-Rainey, Jessica K. Ivani, Marilen Johns, Erika Just, Eri Kashima, Carolina Kipf, Janina V. Klingenberg, Nikita König, Aikaterina Koti, Richard G. A. Kowalik, Olga Krasnoukhova, Nora L. M. Lindvall, Mandy Lorenzen, Hannah Lutzenberger, Tânia R. A. Martins, Celia Mata German, Suzanne van der Meer, Jaime Montoya Samamé, Michael Müller, Saliha Muradoglu, Kelsey Neely, Johanna Nickel, Miina Norvik, Cheryl Akinyi Oluoch, Jesse Peacock, India O. C. Pearey, Naomi Peck, Stephanie Petit, Sören Pieper, Mariana Poblete, Daniel Prestipino, Linda Raabe, Amna Raja, Janis Reimringer, Sydney C. Rey, Julia Rizaew, Eloisa Ruppert, Kim K. Salmon, Jill Sammet, Rhiannon Schembri, Lars Schlabbach, Frederick W. P. Schmidt, Amalia Skilton, Wikaliler Daniel Smith, Hilário de Sousa, Kristin Sverredal, Daniel Valle, Javier Vera, Judith Voß, Tim Witte, Henry Wu, Stephanie Yam, Jingting Ye, Maisie Yong, Tessa Yuditha, Roberto Zariquiey, Robert Forkel, Nicholas Evans, Stephen C. Levinson, Martin Haspelmath, Simon J. Greenhill, Quentin D. Atkinson, and Russell D. Gray. 2023a. Grambank reveals the importance of genealogical constraints on linguistic diversity and highlights the impact of language loss. *Science Advances*, 9(16):eadg6175.

Hedvig Skirgård, Hannah J. Haynie, Harald Hammarström, Damián E. Blasi, Jeremy Collins, Jay Latarche, Jakob Lesage, Tobias Weber, Alena Witzlack-Makarevich, Michael Dunn, Ger Reesink, Ruth Singer, Claire Bowern, Patience Epps, Jane Hill, Outi Vesakoski, Noor Karolin Abbas, Sunny Ananth, Daniel Auer, Nancy A. Bakker, Giulia Barbos, Anina Bolls, Robert D. Borges, Mitchell Browen, Lennart Chevallier, Swintha Danielsen, Sinoël Dohlen, Luise Dorenbusch, Ella Dorn, Marie Duhamel, Farah El Haj Ali, John Elliott, Giada Falcone, Anna-Maria Fehn, Jana Fischer, Yustinus Ghanggo Ate, Hannah Gibson, Hans-Philipp Göbel, Jemima A. Goodall, Victoria Gruner, Andrew Harvey, Rebekah Hayes, Leonard Heer, Roberto E. Herrera Miranda, Nataliia Hübler, Biu H. Huntington-Rainey, Guglielmo Inglese, Jessica K. Ivani, Marilen Johns, Erika Just, Ivan Kapitonov, Eri Kashima, Carolina Kipf, Janina V. Klingenberg, Nikita König, Aikaterina Koti, Richard G. A. Kowalik, Olga Krasnoukhova, Kate Lynn Lindsey, Nora L. M. Lindvall, Mandy Lorenzen, Hannah Lutzenberger, Alexandra Marley, Tânia R. A. Martins, Celia Mata German, Suzanne van der Meer, Jaime Montoya, Michael Müller, Saliha Muradoğlu, HunterGatherer, David Nash, Kelsey

Neely, Johanna Nickel, Miina Norvik, Bruno Olsson, Cheryl Akinyi Oluoch, David Osgarby, Jesse Peacock, India O.C. Pearey, Naomi Peck, Jana Peter, Stephanie Petit, Sören Pieper, Mariana Poblete, Daniel Prestipino, Linda Raabe, Amna Raja, Janis Reimringer, Sydney C. Rey, Julia Rizaew, Eloisa Ruppert, Kim K. Salmon, Jill Sammet, Rhiannon Schembri, Lars Schlabbach, Frederick W. P. Schmidt, Dineke Schokkin, Jeff Siegel, Amalia Skilton, Hilário de Sousa, Kristin Sverredal, Daniel Valle, Javier Vera, Judith Voß, Daniel Wikalier Smith, Tim Witte, Henry Wu, Stephanie Yam, Jingting Ye, Maisie Yong, Tessa Yuditha, Roberto Zariquiey, Robert Forkel, Nicholas Evans, Stephen C. Levinson, Martin Haspelmath, Simon J. Greenhill, Quentin D. Atkinson, and Russell D. Gray. 2023b. Grambank v1.0. Dataset.

Daniel J. Stekhoven and Peter Bühlmann. 2012. Missforest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 833–844, Hong Kong, China. Association for Computational Linguistics.