Cross-Linguistic Syntactic Difference in Multilingual BERT: How Good is It and How Does It Affect Transfer?

Ningyu Xu[1,2], **Tao Gui**[2*], **Ruotian Ma**[1], **Qi Zhang**[1], **Jingting Ye**[3,4], **Menghan Zhang**[2], **Xuanjing Huang**[1,2]
[1]School of Computer Science, Fudan University
[2]Institute of Modern Languages and Linguistics, Fudan University
[3]Department of Chinese Language and Literature, Fudan University
[4]Department of Linguistic and Cultural Evolution, Max Planck Institute for Evolutionary Anthropology
nyxu22@m.fudan.edu.cn {tgui,rtma19,qz,yejingting,mhzhang,xjhuang}@fudan.edu.cn

Abstract

Multilingual BERT (mBERT) has demonstrated considerable cross-lingual syntactic ability, whereby it enables effective zero-shot cross-lingual transfer of syntactic knowledge. The transfer is more successful between some languages, but it is not well understood what leads to this variation and whether it fairly reflects difference between languages. In this work, we investigate the distributions of grammatical relations induced from mBERT in the context of 24 typologically different languages. We demonstrate that the distance between the distributions of different languages is highly consistent with the syntactic difference in terms of linguistic formalisms. Such difference learnt via self-supervision plays a crucial role in the zero-shot transfer performance and can be predicted by variation in morphosyntactic properties between languages. These results suggest that mBERT properly encodes languages in a way consistent with linguistic diversity and provide insights into the mechanism of cross-lingual transfer.

1 Introduction

Cross-lingual transfer aims to address the huge linguistic disparity in NLP by transferring the knowledge acquired in high-resource languages to low-resource ones, where pretrained multilingual encoders, such as Multilingual BERT (mBERT) (Devlin et al., 2019), have proven a powerful facilitator. Compared to other approaches learning certain cross-lingual alignment in a supervised (Gouws et al., 2015; Mikolov et al., 2013; Faruqui and Dyer, 2014) or unsupervised (Artetxe et al., 2017; Zhang et al., 2017; Lample et al., 2018) manner, mBERT directly learns to encode different languages in a shared representation space through self-supervised joint training, dispensing with explicit alignment. It has exhibited notable cross-lingual ability and can perform effective zero-shot cross-lingual transfer across a variety of downstream tasks, albeit the performance varies (Wu and Dredze, 2019; Pires et al., 2019).

The simplicity and efficacy of mBERT are crucial for cross-lingual transfer and have sparked interest in investigating the reason for its success. Previous work has looked into its representation space and found that mBERT automatically performs certain alignment across languages (Cao et al., 2019; Gonen et al., 2020; Conneau et al., 2020; Chi et al., 2020). The extent of alignment is shown correlated with the transfer performance (Muller et al., 2021). Despite these insights into the source of the transfer, it is also intriguing why different languages are aligned to varying degrees and what implication such variation bears. Another line of work has demonstrated that the zero-shot transfer performance is affected by certain linguistic features such as word order (Pires et al., 2019; Karthikeyan et al., 2019), whereas the underlying mechanism is left unexplored. Taken together, it remains unclear how different aspects of cross-linguistic differences impact the representations and further affect the cross-lingual transfer of different tasks.

In this paper, we focus on the syntactic level and investigate the cross-lingual transfer of mBERT based on 24 typologically distinct languages, with the purpose of figuring out the following questions:

Q1: Does mBERT properly induce cross-linguistic syntactic difference via self-supervision? The distance between distributions over mBERT representations of grammatical relations in different languages can be used to evaluate the syntactic difference between languages encoded in mBERT (Section 2). We compare it with the cross-linguistic syntactic difference in terms of linguistic formalisms for validation and rely on it to investigate the cross-lingual ability of mBERT.

Q2: How does the syntactic difference learnt

*Corresponding author.

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 8073–8092
December 7-11, 2022

by mBERT impact its cross-lingual transfer? The zero-shot cross-lingual transfer is typically realized through fine-tuning the pretrained multilingual model on a certain source language. We analyze the change pretraining and fine-tuning brought to the distance between distributions (i.e., the syntactic difference between languages in mBERT) to understand the mechanism behind the transfer (Section 3).

Q3: **If and to what extent do various morphosyntactic properties impact the transfer performance?** We then investigate the reason for the variation in the transfer performance based on syntactic-related linguistic properties. We exploit all the morphosyntactic properties available in the World Atlas of Language Structures (WALS) (Dryer and Haspelmath, 2013) and examine the extent to which variation in them impacts the distance between distributions and further affects the transfer performance through regression analysis (Section 4).

Our quantitative results and qualitative analysis demonstrate that:

1) The distance between distributions of grammatical relations in mBERT is highly consistent with the cross-linguistic syntactic difference in the context of linguistic formalisms. 2) The syntactic difference learnt during pretraining plays a crucial role in the zero-shot cross-lingual transfer of dependency parsing. While fine-tuning on a specific language augments the transfer with task-specific knowledge, it can distort the established cross-linguistic knowledge. 3) Variation in morphosyntactic properties is predictive of the syntactic difference in mBERT, which further impacts the transfer performance. Encouragingly, these linguistic features can be exploited to optimize the cross-lingual transfer, whereby we can efficiently select the best language for fine-tuning without the need for any dataset.[1]

2 A Measure of Cross-Linguistic Syntactic Difference in mBERT

mBERT learns to encode different languages in a shared representation space, which provides a basis for cross-linguistic comparison. However, the syntactic properties of a language are not explicitly realized at a word or sentence level. To bridge the gap between the linguistic knowledge at a language level and the word-level contextual representations, we look into the distributions over mBERT representations of different languages. We first derive representations of syntactic information (i.e., grammatical relations) from mBERT and then use the divergence between distributions over the representations to measure the language-wide difference encoded in it. Finally, we compare the measured difference with the cross-linguistic syntactic difference in terms of formal syntax to examine whether mBERT properly induces syntactic difference via self-supervision.

2.1 Method

Multilingual BERT mBERT is a Transformer-based (Vaswani et al., 2017) neural language model, which has the same architecture as BERT-Base but is pretrained on a concatenation of monolingual Wikipedia corpora from 104 languages. For each input sentence tokenized into a sequence of n tokens $w_{1:n}$, mBERT runs them through an embedding layer and 12 layers of transformer encoders, producing a sequence of contextual representations $\mathbf{h}^{\ell}_{1:n}$ for each token at each layer ℓ, where $1 \leq \ell \leq 12$. As there is no explicit cross-lingual alignment provided during the entire training procedure, it is intriguing **how common linguistic properties vary across languages in mBERT representation space**.

Representations of grammatical relations in mBERT We adopt the framework of Universal Dependencies (UD) (de Marneffe et al., 2021) in describing abstract syntactic structure across typologically diverse languages, where the dependency grammatical relations are universal and allow for cross-linguistic comparison. In the light of work of the structural probe (Hewitt and Manning, 2019; Chi et al., 2020), we use the difference between mBERT representations of a head-dependent pair of words $(w_{\text{head}}, w_{\text{dep}})$ to represent the grammatical relation between them:

$$\mathbf{d}^{\ell}_{(\text{head,dep})} = \mathbf{h}^{\ell}_{\text{head}} - \mathbf{h}^{\ell}_{\text{dep}}, \qquad (1)$$

and verify its effectiveness through a linear classifier decoding the grammatical relation from it. We then visualize the representations $\mathbf{d}^{\ell}_{(\text{head,dep})}$ to get a qualitative understanding of the grammatical information encoded in them[2].

[1] Our code is available at https://github.com/ningyuxu/cl-syntactic-difference-mbert.

[2] See Appendix A.1 for details.

Evaluation of cross-linguistic syntactic difference in mBERT To evaluate the language-wide difference in terms of grammatical relations, we abstract away from single sentences and look into the distributions of representations. Formally, we regard the dataset in language L as a set of N feature-label pairs, i.e., $\mathcal{D}_L = \left\{\left(\mathbf{x}^{(i)}, \mathbf{y}^{(i)}\right)\right\}_{i=1}^{N} \sim P_L(\mathbf{x}, \mathbf{y})$, where feature $\mathbf{x}$ is our representation $\mathbf{d}_{(\text{head,dep})}$ and the label $\mathbf{y}$ is the gold grammatical relation between the word pair $(w_{\text{head}}, w_{\text{dep}})$. $P_L(\mathbf{x}, \mathbf{y})$ denotes the joint distribution over the feature-label space. We define the syntactic difference between L_A and L_B ($d_{\mathcal{S}}(L_A, L_B)$) as the distance between their joint distributions:

$$d_{\mathcal{S}}(L_A, L_B) \triangleq d\left(P_{L_A}(\mathbf{x}, \mathbf{y}), P_{L_B}(\mathbf{x}, \mathbf{y})\right). \quad (2)$$

The optimal transport dataset distance (OTDD) (Alvarez-Melis and Fusi, 2020) is employed for the estimation of the distance[3], as it has a solid theoretical footing, discards extra parameters, and yields distance both between datasets and between labels, benefiting fine-grained analysis of the representation space.

Validation of cross-linguistic syntactic difference in mBERT We validate the effectiveness of our measure through comparison with the cross-linguistic syntactic difference in the context of linguistic formalisms. We adopt the formal syntactic distance provided in Ceolin et al. (2020), which is measured based on the theory of Principles-and-Parameters developed since Chomsky (2010). It compares the syntactic structure of different languages through a finite set of universal abstract grammatical parameters characterizing possible cross-linguistic differences, which in principle enables a systematic comparison between syntax of different languages (Longobardi and Guardiano, 2009). In detail, each parameter is coded as a binary value, and a language L is represented by the list of parameters S_L it takes. The formal syntactic distance between language L_A and L_B is measured by Jaccard distance (Jaccard, 1901) between them:

$$d_{\mathcal{F}}(L_A, L_B) \triangleq d_{\text{Jaccard}}(S_{L_A}, S_{L_B}). \quad (3)$$

2.2 Experimental Setup

Data The data for all our experiments is from UD treebanks. We adopt all the grammatical relations defined in it except for *root* as it does not denote relations between words. We select 24 typologically different languages covering a reasonable variety of language families, which are Arabic, Bulgarian, German, Greek, English, Spanish, Estonian, Persian, Finnish, French, Hebrew, Hindi, Italian, Japanese, Korean, Latvian, Dutch, Polish, Portuguese, Romanian, Russian, Turkish, Vietnamese and Chinese[4].

[3]See Appendix A.2 for details.

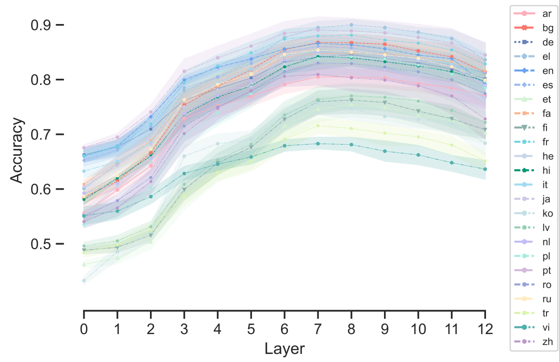

Figure 1: Accuracy in recovering grammatical relations of different languages across the layers of mBERT. The colored bands denote 95% confidence intervals.

Baseline We compare mBERT with the following two baselines:

- **MBERT0** The layer 0 of mBERT, which does not involve any contextual information.
- **MBERTRAND** A model same as mBERT but without pretrained weights. The subword embeddings remain unchanged.

2.3 Results

Evaluating cross-linguistic syntactic difference in mBERT The probing result (Figure 1) demonstrates that **grammatical relation can be successfully extracted from the representations** computed based on our method in contrast to baselines[5]. The 7th and 8th layer are most effective in encoding grammatical relations across the languages.

The representations we derive from the 7th layer of mBERT generally form clusters reflecting their grammatical relations (Figure 2). Moreover, we can find that the distributions of different languages differ and such difference reflects certain difference

[4]Constrained by the availability of UD datasets and mBERT's pretraining, many languages belong to the Indo-European family. See Appendix E.1 for the datasets we use.

[5]See Appendix A.1 Table 3 for comparison with baselines.

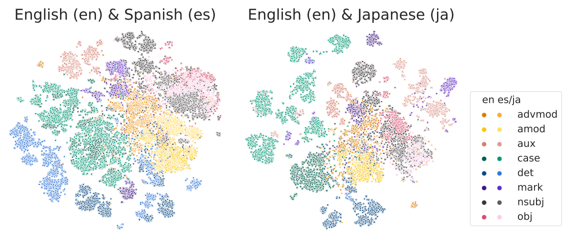

Figure 2: Visualization of the representations of different grammatical relations derived from the 7th layer of mBERT. English is shown more similar to Spanish than to Japanese as to the distributions of grammatical relations such as *case*, *obj* and *aux*.

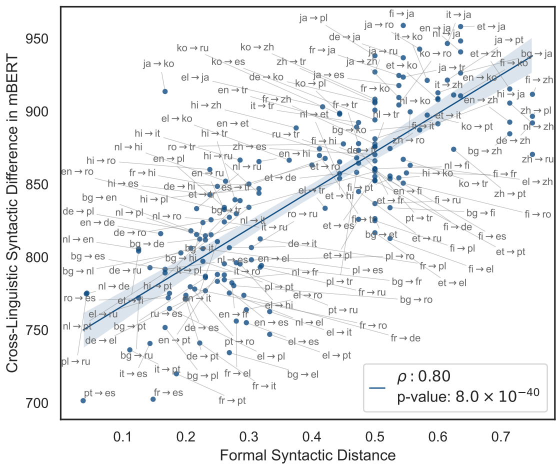

Figure 3: Comparison of the formal syntactic distance and the cross-linguistic syntactic difference induced from mBERT, evaluated through Spearman's correlation.

between languages. The representations of the same grammatical relations better clustered together between English and Spanish than between English and Japanese, in line with the fact that English is more similar to Spanish than to Japanese at the syntactic level.

Validating cross-linguistic syntactic difference in mBERT The cross-linguistic syntactic difference measured based on mBERT shows significantly high correlation with the formal syntactic distance (Figure 3). And the correlation is higher in the 7th layer ($\rho = 0.80$) than baselines ($\rho = 0.72$ for MBERT0 and 0.68 for MBERTRAND), which indicates that **mBERT properly induces difference in syntactic structure via self-supervision**.

2.4 Discussion

Grammatical relations can be largely derived from the representations computed based on our method, but to different degrees. As shown in

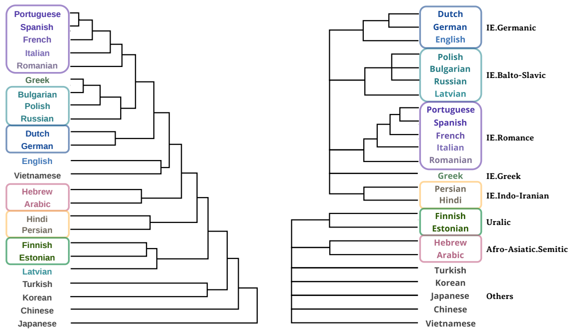

Figure 4: Left: Hierarchical clustering based on cross-linguistic syntactic difference derived from mBERT. Right: The gold phylogenetic tree from Glottolog (Hammarström et al., 2021). IE stands for the Indo-European family.

the probing result (Figure 1), the representations are less effective in encoding syntactic knowledge for languages such as Turkish, Hebrew, Estonian, Finnish, Korean and Chinese, where the former four have rich morphology and the latter are tokenized with CJK characters[6]. Previous work has demonstrated similar disparity in mBERT (Chi et al., 2020; Mueller et al., 2020) and suggests that the inadequacy in tokenization can be a possible reason (Rust et al., 2021).

While the syntactic difference induced from mBERT is highly consistent with the distance in formal grammar, certain deviation can be observed, especially for languages poorly represented where the probe classifier achieves a relatively lower performance.

We further perform a hierarchical clustering based on our measure to understand the relationship between languages it reveals. **Languages in the same family are generally clustered together, analogous to conventional understanding in linguistic taxonomy, while there exist discrepancies** regarding languages such as Vietnamese and Latvian (Figure 4). Besides the deficiency in representations, they might stem from i) the sampling bias in the UD treebanks, especially for low-resource languages such as Vietnamese, and ii) the difference between languages in terms of dependency grammar better reflecting grammatical diversity. For instance, though belonging to the Indo-European family, Latvian bears structural similarities to Finno-Ugric languages (Kalnaca, 2014). Such result is in line with previous work showing certain correlation between grammatical

[6]Additionally, the deficiency in Vietnamese may result from lack of training data as its treebank is relatively small.

typology and historical relatedness (Dunn et al., 2005; Wichmann and Saunders, 2007; Longobardi and Guardiano, 2009; Abramov and Mehler, 2011) and suggests that the relationship between languages in terms of syntax should be properly reflected in mBERT representation space.

3 Mechanism behind Cross-Lingual Transfer

The training procedure of zero-shot transfer typically involves two steps: pretraining on a multilingual corpus and fine-tuning on a specific source language. To understand the mechanism behind the zero-shot transfer and why the transfer performance varies across languages, we look into the change they bring to the syntactic difference in mBERT. Specifically, we first compare the syntactic difference learnt during pretraining with the transfer performance to evaluate the impact of pretraining on the transfer and then examine how fine-tuning on a specific language changes the syntactic difference.

3.1 Method

Analyzing the effect of pretraining We investigate what effect the syntactic difference learnt during pretraining has on the transfer performance through a correlation analysis. The performance of dependency parsing is measured by labeled attachment score (LAS). Let

$$drop(L_S, L_T) \triangleq \text{LAS}_{L_S} - \text{LAS}_{L_T} \tag{4}$$

denote the drop in LAS when transferring the model fine-tuned on a source language L_S to a target language L_T, we compare it with the syntactic difference $d_S^{(\text{pre})}(L_S, L_T)$ measured based on (2) in pretrained mBERT.

Analyzing the effect of fine-tuning To understand how fine-tuning on a source language impacts the zero-shot transfer of the dependency parsing task, we investigate the change it brings to the syntactic difference between the source and target languages in mBERT. We first visualize mBERT representations of grammatical relations[7] before and after fine-tuning to get a qualitative understanding, and then quantitatively compare the syntactic difference in pretrained mBERT and mBERT fine-tuned on the source language.

[7] We use the same method of visualization as in Section 2.1. See Appendix A.1 for details.

To further explore whether the change in the syntactic difference impacts the variation in transfer performance among different target languages, we perform a correlation analysis of their syntactic difference with the source language before and after fine-tuning. We also compare the change that fine-tuning on different source languages brings to the syntactic difference to understand how fine-tuning on a particular language may affect the overall cross-linguistic syntactic knowledge in mBERT.

3.2 Experimental Setup

Following the setup of Wu and Dredze (2019), we use the parser with deep biaffine attention mechanism (Dozat and Manning, 2017) as the task-specific layer on top of mBERT for dependency parsing, which has been shown to perform the best on average across typologically different languages (Ahmad et al., 2019). Instead of providing gold Part-of-Speech (POS) tags, we train a linear model to predict POS tags on the source side, and apply it to the target language. We employ this strategy to avoid introducing additional cross-lingual information, as we focus on the cross-lingual ability of mBERT itself.

We take the 7th layer of pretrained mBERT as information about grammatical relations is best manifested here. For fine-tuned models, we focus on the 12th layer as the representations here are directly fed to the parser and impact the transfer performance.

3.3 Results

Effect of pretraining The syntactic difference acquired during pretraining strongly correlates with the drop in LAS across typologically diverse languages (Figure 5), in contrast to the baselines ($\rho = 0.51$ for MBERT0 and 0.45 for MBERTRAND). The result suggests that, with a given source language, **the syntactic difference learnt during pretraining plays a crucial role in the cross-lingual transfer performance**.

Effect of fine-tuning After fine-tuning on English, representations of the same grammatical relation better cluster together (Figure 6), indicating a task-specific improvement in both the source and target languages. Our quantitative analysis reveals that the syntactic difference with the source language in mBERT generally decreases after fine-tuning (Figure 7), where the distance between the

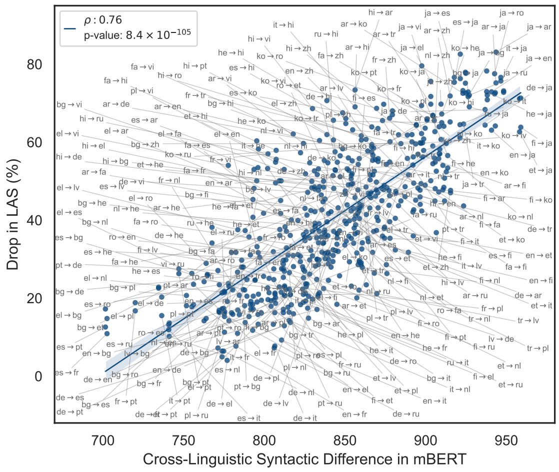

Figure 5: Comparison of the cross-linguistic syntactic difference in pretrained mBERT and drop in zero-shot cross-lingual transfer performance (LAS).

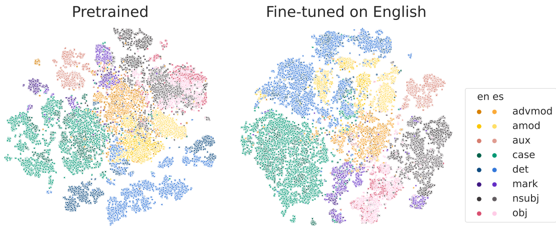

Figure 6: Visualization of the representations of grammatical relations in English (en) and Spanish (es) derived from pretrained mBERT and mBERT fine-tuned on English. Representations of the same grammatical relation in different languages better cluster together after fine-tuning.

same grammatical relations decrease much more drastically than the others (Figure 8). These results, together, suggest that fine-tuning facilitates the zero-shot cross-lingual transfer with task-specific knowledge.

Through a correlation analysis, we find an approximately linear relationship between the syntactic difference with the source language before and after fine-tuning. Namely, **fine-tuning on a specific language benefits other languages according to the similarity between them learnt during pretraining**. However, **it can distort the overall cross-linguistic syntactic knowledge**, especially for languages with a bigger difference. Figure 9 shows that the syntactic difference with English is worse correlated with $d_S^{(\text{pre})}(\text{en}, \cdot)$ when fine-tuning on typologically distant languages such as Polish than on English, indicating that **the relationship among languages can be deformed when augmenting the pretrained model with syntactic knowledge in a single language**.

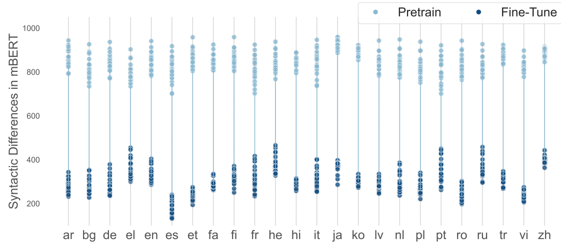

Figure 7: The syntactic difference in mBERT before and after fine-tuning on the language on the x-axis. Each point represents the syntactic difference between the source language on the x-axis and another language.

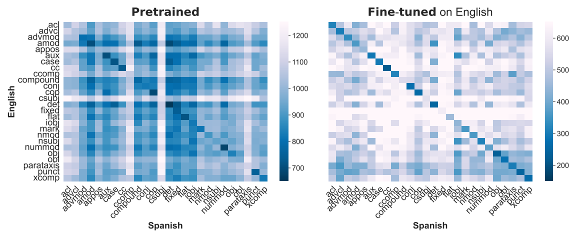

Figure 8: Distance between distributions of grammatical relations in English and Spanish before and after fine-tuning on English. The distance between the same grammatical relation becomes much smaller after fine-tuning, indicating a task-specific improvement in transfer.

3.4 Discussion

Our experiment results are complementary to previous work in the monolingual setting, which has shown that fine-tuning benefits downstream tasks with clearer distinction between samples belonging to different labels but also largely preserves the original spatial structure of the pretrained model (Merchant et al., 2020; Zhou and Srikumar, 2022). In mBERT, we can further explore how fine-tuning on a specific language impacts the representations of other languages, i.e., samples in different domains. While for that language, fine-tuning augments the effect of pretraining and benefits the transfer, it distorts the established cross-linguistic knowledge especially for languages with a larger divergence in distributions.

4 Impact of Linguistic Diversity

To better understand the cross-linguistic syntactic difference learnt by mBERT, we employ the structural and functional features in linguistic typology which allows for description of linguistic diversity and analyze to what extent variation in these features affects the syntactic difference

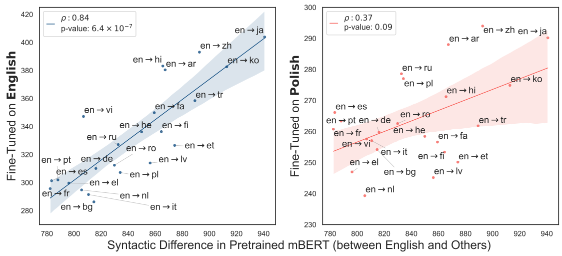

Figure 9: Left: Comparison of the syntactic difference between English and other languages derived from pretrained mBERT (x-axis) and mBERT fine-tuned on English (y-axis). Right: Comparison of the syntactic difference between English and others derived from pretrained mBERT (x-axis) and mBERT fine-tuned on Polish (y-axis). The syntactic difference in mBERT fine-tuned on Polish is not significantly correlated with that in the pretrained model ($p > 0.05$).

in mBERT. We further examine whether these features can be exploited to select better source languages and thus benefit cross-lingual transfer.

4.1 Method

Typological features We exploit all the morphosyntactic features available in WALS (Dryer and Haspelmath, 2013), covering areas including Morphology, Nominal Categories, Verbal Categories, Nominal Syntax, Word Order, Simple Clauses, and Complex Sentences.[8]

Evaluation of difference in typological features For each feature f, there are between 2 to 28 different values in WALS and they may not be mutually exclusive. We regard each feature as a vector $\mathbf{v}_f^L = [v_1^L, \cdots, v_m^L]$ where m is the number of possible values for a feature f and each entry $v_i^L (i = 1, \cdots, m)$ typically represents a binary value that a language L may take (see Table 1 for an example). We use cosine distance to measure the difference between language L_A and L_B in this feature:

$$d_f(L_A, L_B) \triangleq 1 - \cos\left(\mathbf{v}_f^{L_A}, \mathbf{v}_f^{L_B}\right). \quad (5)$$

The overall difference between L_A and L_B is represented by

$$\mathbf{d}_F(L_A, L_B) = [d_{f_1}, \cdots, d_{f_n}], \quad (6)$$

where $n = 116$ is the total number of features.

[8]We filter out the features which have missing values for all the languages we study, which results in a total of 116 features. See Appendix C.1 for all the features we use.

Language	NRel	RelN	Correlative
English	1	0	0
Hindi	0	0	1
Hungarian	1	1	0
Japanese	0	1	0

Table 1: A truncated example of WALS feature 90A: Order of Relative Clause (Rel) and Noun (N). Each entry typically takes a binary value for a particular language. For Hungarian, there is not a dominant type of the order of Rel and N, and instead, both NRel and RelN exist.

Regression analysis Given the observable correlation and potential interdependence between these features[9], we use a gradient boosting regressor[10] combined with impurity-based and permutation importance to analyze the impact of different features, as it is robust to multicollinearity, generally achieves high empirical performance, and is relatively interpretable. The regressor $\mathcal{G}$ takes as input $\mathbf{d}_F(L_A, L_B)$ and the target is to predict the syntactic difference between them, i.e., $d_S^{(\text{pre})}(L_A, L_B)$.

Selection of source languages To further examine our findings and also improve the cross-lingual transfer, we extend the regressor to predict the syntactic difference between the J languages we study $\{L_1, L_2, \cdots, L_J\}$ and another language L_K and then test whether $L_S = \text{argmin}_j\, \mathcal{G}\left(\mathbf{d}_F(L_j, L_K)\right)$ is among the best source languages for zero-shot cross-lingual transfer.

4.2 Experimental Setup

Model and evaluation in regression analysis We train the gradient boosting regressor with 100 estimators where each has a maximum depth of three. Its performance is evaluated through the average of R^2 in 10-fold cross-validation. For feature importance, we report both permutation importance with 30 repeats and impurity-based importance.

Evaluation of source language selection We test the effectiveness of our regressor in source language selection on five other languages including Czech, Catalan, Hungarian, Tamil and Urdu. Specifically, each of them is taken as a target

[9]For instance, implicational universals of word order (Greenberg, 1990; Dryer, 1992)) may be driven by some universal constraints (Levshina, 2019; Hahn et al., 2020).

[10]https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingRegressor.html

language and our goal is to choose the best source language among the languages we select. For each target language, we use the regressor to predict the syntactic differences between it and the source languages, and rank them from low to high to get the predicted ranking of source languages. To get the gold ranking for evaluation, we fine-tune an mBERT on each of the source languages, test it on the target language to obtain the LAS, and rank the scores from high to low. Similar to Lin et al. (2019), we use the Normalized Discounted Cumulative Gain (Järvelin and Kekäläinen, 2002) at position 3 (NDCG@3)[11] as the evaluation metric. It measures the quality of ranking and yields a score between 0 and 1, where the gold ranking gets a score of 1.

Baseline We compare the trained regressor with these baselines:

- **AVE** The average distance of all morphosyntactic features.
- **URIEL** The different kinds of distance provided in Littell et al. (2017), including syntactic d_{syn}[12], genetic d_{gen}, featural d_{fea}, geographic d_{geo}, phonological d_{pho} and inventory distance d_{inv}.

4.3 Results

Regression Analysis The R^2 score of the regressor reaches 85%, showing that **the differences in morphosyntactic features are predictive of the syntactic difference between languages learnt by mBERT**. Additionally, that the correlation score ($\rho = 0.89$) between the predicted and the computed syntactic difference is higher than baselines ($\rho = 0.58$ for AVERAGE and 0.68 for d_{syn} in URIEL) suggests that these features should be treated with different importance.

Feature importance Figure 10 shows the five most important features[13]. The dominant role of features belonging to the area of word order supports previous work emphasizing the importance of word order typology in characterizing the difference between languages (Ahmad et al., 2019; Pires et al., 2019; Karthikeyan et al., 2019; Dufter and Schütze, 2020).

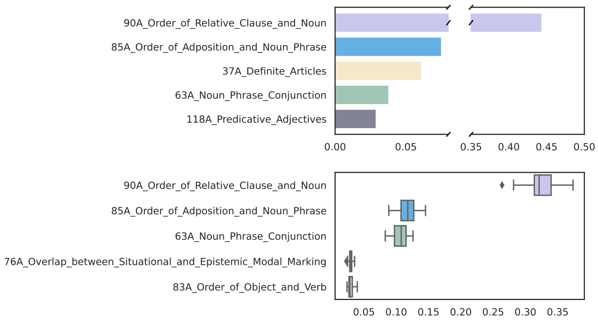

Figure 10: The five typological features having the biggest impact on the syntactic difference in mBERT. Above: Impurity-based importance. Below: Permutation importance.

Method	NDCG(%)	Method	NDCG(%)
AVE	66.1	d_{geo}	52.6
d_{syn}	71.7	d_{pho}	23.7
d_{gen}	61.6	d_{inv}	59.4
d_{fea}	57.7	**REG**	**77.0**

Table 2: Results of source language selection strategy evaluated by NDCG@3 (%). REG is our method.

Source language selection Our regressor effectively selects better source languages for zero-shot cross-lingual transfer of dependency parsing than baselines (Table 2), which further verifies our findings and indicates that **morphosyntactic features are good indicators of transfer performance**.

4.4 Discussion

Previous work has tried to predict the cross-lingual transfer performance based on typological features (Lin et al., 2019; Pires et al., 2019), whereas a general metric of typological similarity may not be informative enough. Dolicki and Spanakis (2021) conducted a finer-grained analysis, but the aim is to choose the best source language for a certain downstream task, not focusing on specific language pairs.

We here show that the morphosyntactic features are predictive of the cross-linguistic syntactic difference learnt during pretraining and have a great potential to benefit the cross-lingual transfer. As our method is based on distributions and is not constrained at a language level, it can be extended to cross-domain and multi-source transfer scenarios, where data from different languages or domains can be treated as one dataset and the effects of linguistic properties may be reevaluated. Combined with finer-grained linguistic features, it

[11] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.ndcg_score.html

[12] The syntactic distance here is the cosine distance between feature vectors derived from typological databases including WALS.

[13] For importance of all features, see Appendix C.2.

is promising to provide more insight into the cross-lingual transfer.

5 Related Work

Probing for linguistic knowledge Contextualized word embeddings have been found to be especially effective at the syntactic level (Linzen and Baroni, 2021; Baroni, 2021). Through probing methods, prior work has shown that syntactic knowledge including syntactic tree depth, subject-verb agreement (Conneau et al., 2018; Jawahar et al., 2019), constituent labels, grammatical relations (Tenney et al., 2019; Liu et al., 2019) and dependency parse trees (Hewitt and Manning, 2019) can be largely derived from these embeddings. In the multilingual setting, mBERT has been found to encode morphosyntactic properties such as syntactic structure (Chi et al., 2020) and morphosyntactic alignment (Papadimitriou et al., 2021) in a similar way across languages. There has been work noting problems related to the probing method (Hewitt and Liang, 2019; Pimentel et al., 2020; Voita and Titov, 2020), suggesting that the extra classifier can interfere with the analysis of the embedding space. We here derive representations of syntactic knowledge through a simple subtraction between embeddings and discard task-specific parameters through a measure of distance between their representations.

Linguistic diversity Difference in linguistic properties across languages has been associated with the hardness of transfer (Ponti et al., 2018; Lin et al., 2019) and typological resources have been exploited to guide parameter and information sharing among languages (Naseem et al., 2012; Täckström et al., 2013; Ammar et al., 2016) and data selection (Ponti et al., 2018; Lin et al., 2019). Previous work has demonstrated that the transfer performance is greatly affected by typological features such as word order both in a delexicalized setting before the emergence of large pretrained language models (Aufrant et al., 2016) and in the context of multilingual language models (Pires et al., 2019; Karthikeyan et al., 2019; Dufter and Schütze, 2020). Moreover, much typological information is found encoded in mBERT representations (Choenni and Shutova, 2020) and blinding mBERT to it impedes successful cross-lingual transfer (Bjerva and Augenstein, 2021). On the other hand, Singh et al. (2019) shows that the representation space of mBERT is partitioned in a way similar to genealogical relatedness. While most previous work investigates sentence-level or word-level representations and mixes various aspects of linguistic knowledge, we here focus on the cross-lingual syntactic transfer and extract representations in a targeted manner.

6 Conclusion

Languages vary profoundly at almost every level including lexicon, grammar and meaning. Pre-trained multilingual encoders learn to encode them in a shared representation space simply via self-supervision, but it is unclear how they address the linguistic variation at different levels. This paper investigates the cross-lingual syntactic ability of mBERT. Through a measure of distance between distributions over its representations, we demonstrate that mBERT encodes universal grammatical relations in a way highly consistent with the cross-linguistic syntactic difference in terms of formal syntax. Such cross-linguistic syntactic knowledge plays a decisive role in the zero-shot cross-lingual transfer performance of dependency parsing. This evidence suggests that linguistic knowledge such as typological resources can be incorporated in improvement of cross-lingual transfer and thus help to better accommodate the rich linguistic diversity.

Limitations

At the core of our method is a measure of divergence between distributions, which highly correlates with the zero-shot cross-lingual transfer performance. As it is challenging to choose an appropriate measure of divergence between joint distributions, we empirically compared several measures, and they yield similar results. We here employ the optimal transport distance between datasets (Alvarez-Melis and Fusi, 2020) as it provides interpretable correspondence and characterize the geometry of the representation space. A detailed analysis of the best measure of divergence in the multilingual setting is left for future work.

Combined with representations of grammatical relations derived from mBERT, our method provides a quantitative evaluation of the cross-linguistic difference learnt by mBERT in terms of dependency grammar. It can be related with typological diversity and help to analyze the effects of various morphosyntactic properties. Future work can extend to finer-grained description of linguistic variation and other downstream tasks involving different aspects of language. By clarifying the source of cross-lingual transfer and understanding how linguistic diversity affects the model, significant improvements on efficient cross-lingual transfer can be expected.

Acknowledgements

The authors would like to thank the anonymous reviewers for their helpful comments. This work was partially funded by National Natural Science Foundation of China (No. 62206057, 62076069, 61976056).

References

Olga Abramov and Alexander Mehler. 2011. Automatic language classification by means of syntactic dependency networks. *Journal of Quantitative Linguistics*, 18(4):291–336.

Wasi Ahmad, Zhisong Zhang, Xuezhe Ma, Eduard Hovy, Kai-Wei Chang, and Nanyun Peng. 2019. On Difficulties of Cross-Lingual Transfer with Order Differences: A Case Study on Dependency Parsing. 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 2440–2452, Minneapolis, Minnesota. Association for Computational Linguistics.

David Alvarez-Melis and Nicolo Fusi. 2020. Geometric Dataset Distances via Optimal Transport. In *Advances in Neural Information Processing Systems*, volume 33, pages 21428–21439. Curran Associates, Inc.

Waleed Ammar, George Mulcaire, Miguel Ballesteros, Chris Dyer, and Noah A. Smith. 2016. Many languages, one parser. *Transactions of the Association for Computational Linguistics*, 4:431–444.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2017. Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 451–462, Vancouver, Canada. Association for Computational Linguistics.

Lauriane Aufrant, Guillaume Wisniewski, and François Yvon. 2016. Zero-resource dependency parsing: Boosting delexicalized cross-lingual transfer with linguistic knowledge. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 119–130, Osaka, Japan. The COLING 2016 Organizing Committee.

Marco Baroni. 2021. On the proper role of linguistically-oriented deep net analysis in linguistic theorizing. *arXiv:2106.08694 [cs]*. ArXiv: 2106.08694.

Johannes Bjerva and Isabelle Augenstein. 2021. Does Typological Blinding Impede Cross-Lingual Sharing? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 480–486, Online. Association for Computational Linguistics.

Steven Cao, Nikita Kitaev, and Dan Klein. 2019. Multilingual alignment of contextual word representations. In *International Conference on Learning Representations*.

Andrea Ceolin, Cristina Guardiano, Monica Alexandrina Irimia, and Giuseppe Longobardi. 2020. Formal Syntax and Deep History. *Frontiers in Psychology*, 11.

Ethan A. Chi, John Hewitt, and Christopher D. Manning. 2020. Finding Universal Grammatical Relations in Multilingual BERT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5564–5577, Online. Association for Computational Linguistics.

Rochelle Choenni and Ekaterina Shutova. 2020. What does it mean to be language-agnostic? Probing multilingual sentence encoders for typological properties. *arXiv:2009.12862 [cs]*. ArXiv: 2009.12862.

Noam Chomsky. 2010. *Lectures on Government and Binding: The Pisa Lectures*. De Gruyter Mouton.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136, Melbourne, Australia. Association for Computational Linguistics.

Alexis Conneau, Shijie Wu, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Emerging Cross-lingual Structure in Pretrained Language Models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6022–6034, Online. Association for Computational Linguistics.

Marco Cuturi. 2013. Sinkhorn Distances: Lightspeed Computation of Optimal Transport. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

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

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.

Błażej Dolicki and Gerasimos Spanakis. 2021. Analysing The Impact Of Linguistic Features On Cross-Lingual Transfer. *arXiv:2105.05975 [cs]*. ArXiv: 2105.05975.

Timothy Dozat and Christopher D. Manning. 2017. Deep Biaffine Attention for Neural Dependency Parsing. *arXiv:1611.01734 [cs]*. ArXiv: 1611.01734.

Matthew Dryer. 1992. The greenbergian word order correlations. *Language*, 68:138 – 81.

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

Philipp Dufter and Hinrich Schütze. 2020. Identifying Elements Essential for BERT's Multilinguality. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4423–4437, Online. Association for Computational Linguistics.

Michael Dunn, Angela Terrill, Ger Reesink, Robert A. Foley, and Stephen C. Levinson. 2005. Structural phylogenetics and the reconstruction of ancient language history. *Science*, 309(5743):2072–2075.

Manaal Faruqui and Chris Dyer. 2014. Improving vector space word representations using multilingual correlation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 462–471, Gothenburg, Sweden. Association for Computational Linguistics.

Hila Gonen, Shauli Ravfogel, Yanai Elazar, and Yoav Goldberg. 2020. It's not Greek to mBERT: Inducing Word-Level Translations from Multilingual BERT. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 45–56, Online. Association for Computational Linguistics.

Stephan Gouws, Yoshua Bengio, and Greg Corrado. 2015. Bilbowa: Fast bilingual distributed representations without word alignments. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, page 748–756. JMLR.org.

Joseph Harold Greenberg. 1990. Some universals of grammar with particular reference to the order of meaningful elements. *On Language*.

Michael Hahn, Dan Jurafsky, and Richard Futrell. 2020. Universals of word order reflect optimization of grammars for efficient communication. *Proceedings of the National Academy of Sciences*, 117(5):2347–2353. ISBN: 9781910923115 Publisher: National Academy of Sciences Section: Social Sciences.

Harald Hammarström, Robert Forkel, Martin Haspelmath, and Sebastian Bank. 2021. glottolog/glottolog: Glottolog database 4.5. Zenodo.

John Hewitt and Percy Liang. 2019. Designing and Interpreting Probes with Control Tasks. 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 2733–2743, Hong Kong, China. Association for Computational Linguistics.

John Hewitt and Christopher D. Manning. 2019. A Structural Probe for Finding Syntax in Word Representations. 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 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.

Paul Jaccard. 1901. Etude de la distribution florale dans une portion des alpes et du jura. *Bulletin de la Societe Vaudoise des Sciences Naturelles*, 37:547–579.

Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–446.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What Does BERT Learn about the Structure of Language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy. Association for Computational Linguistics.

Andra Kalnaca. 2014. A typological perspective on latvian grammar. In *A Typological Perspective on Latvian Grammar*. De Gruyter Open Poland.

K Karthikeyan, Zihan Wang, Stephen Mayhew, and Dan Roth. 2019. Cross-lingual ability of multilingual bert: An empirical study. In *International Conference on Learning Representations*.

Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *International Conference on Learning Representations*.

Natalia Levshina. 2019. Token-based typology and word order entropy: A study based on Universal Dependencies. *Linguistic Typology*, 23(3):533–572. Publisher: De Gruyter Mouton.

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.

Tal Linzen and Marco Baroni. 2021. Syntactic Structure from Deep Learning. *Annual Review of Linguistics*, 7(1):195–212. ArXiv: 2004.10827.

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.

Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. 2019. Linguistic Knowledge and Transferability of Contextual Representations. *arXiv:1903.08855 [cs]*. ArXiv: 1903.08855.

Giuseppe Longobardi and Cristina Guardiano. 2009. Evidence for syntax as a signal of historical relatedness. *Lingua*, 119(11):1679–1706.

Amil Merchant, Elahe Rahimtoroghi, Ellie Pavlick, and Ian Tenney. 2020. What happens to BERT embeddings during fine-tuning? In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 33–44, Online. Association for Computational Linguistics.

Tomas Mikolov, Quoc V Le, and Ilya Sutskever. 2013. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*.

Aaron Mueller, Garrett Nicolai, Panayiota Petrou-Zeniou, Natalia Talmina, and Tal Linzen. 2020. Cross-Linguistic Syntactic Evaluation of Word Prediction Models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5523–5539, Online. Association for Computational Linguistics.

Benjamin Muller, Yanai Elazar, Benoît Sagot, and Djamé Seddah. 2021. First Align, then Predict: Understanding the Cross-Lingual Ability of Multilingual BERT. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2214–2231, Online. Association for Computational Linguistics.

Tahira Naseem, Regina Barzilay, and Amir Globerson. 2012. Selective sharing for multilingual dependency parsing. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 629–637, Jeju Island, Korea. Association for Computational Linguistics.

Isabel Papadimitriou, Ethan A. Chi, Richard Futrell, and Kyle Mahowald. 2021. Deep Subjecthood: Higher-Order Grammatical Features in Multilingual BERT. *arXiv:2101.11043 [cs]*. ArXiv: 2101.11043.

Tiago Pimentel, Josef Valvoda, Rowan Hall Maudslay, Ran Zmigrod, Adina Williams, and Ryan Cotterell. 2020. Information-Theoretic Probing for Linguistic Structure. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4609–4622, Online. 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.

Edoardo Maria Ponti, Roi Reichart, Anna Korhonen, and Ivan Vulić. 2018. Isomorphic Transfer of Syntactic Structures in Cross-Lingual NLP. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1531–1542, Melbourne, Australia. Association for Computational Linguistics.

Phillip Rust, Jonas Pfeiffer, Ivan Vulić, Sebastian Ruder, and Iryna Gurevych. 2021. How Good is Your Tokenizer? On the Monolingual Performance of Multilingual Language Models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3118–3135, Online. Association for Computational Linguistics.

Jasdeep Singh, Bryan McCann, Richard Socher, and Caiming Xiong. 2019. BERT is Not an Interlingua and the Bias of Tokenization. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, pages 47–55, Hong Kong, China. Association for Computational Linguistics.

Oscar Täckström, Ryan McDonald, and Joakim Nivre. 2013. Target language adaptation of discriminative transfer parsers. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1061–1071, Atlanta, Georgia. Association for Computational Linguistics.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R. Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R. Bowman, Dipanjan Das, and Ellie Pavlick. 2019. What do you learn from context? probing for sentence structure in contextualized word representations. In *International Conference on Learning Representations*.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Elena Voita and Ivan Titov. 2020. Information-Theoretic Probing with Minimum Description Length. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 183–196, Online. Association for Computational Linguistics.

Søren Wichmann and Arpiar Saunders. 2007. How to use typological databases in historical linguistic research. *Diachronica*, 24:373–404.

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.

Meng Zhang, Yang Liu, Huanbo Luan, and Maosong Sun. 2017. Earth mover's distance minimization for unsupervised bilingual lexicon induction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1934–1945, Copenhagen, Denmark. Association for Computational Linguistics.

Yichu Zhou and Vivek Srikumar. 2022. A closer look at how fine-tuning changes BERT. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1046–1061, Dublin, Ireland. Association for Computational Linguistics.

A Additional Materials for Measure of Syntactic Difference in mBERT

A.1 Representations for Grammatical Relations

Grammatical relation probe For each language, we train a linear classifier via stochastic gradient descent[14] to identify the grammatical relation between a word pair $(w_{\text{head}}, w_{\text{dep}})$ given the input representation $\mathbf{d}^{\ell}_{(\text{head,dep})}$.

Table 3 shows that the layer 7 of mBERT significantly outperforms the baselines in representing grammatical relations ($W = 0.0$ and $p = 1.19 \times 10^{-7}$)[15].

[14]https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.SGDClassifier.html

[15]As MBERTRAND performs similar across different layers, we take the 7th layer of it for comparison in the following experiments.

Visualization of the representation space We combine t-SNE[16] (van der Maaten and Hinton, 2008) with PCA to visualize the representations in two dimensions[17]. As to t-SNE, the perplexity is set to 30 and the maximum number of iteration is set to 1000.

A.2 Evaluation of Syntactic Difference in mBERT

Distance between Distributions The method of optimal transport dataset distance (OTDD) (Alvarez-Melis and Fusi, 2020) relies on optimal transport and defines the metric space as $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, where $\mathcal{X}$ is the feature space and $\mathcal{Y}$ is the label set. The metric on $\mathcal{Z}$ is defined as

$$d_{\mathcal{Z}}\left(\mathbf{z}, \mathbf{z}'\right) = \left(d_{\mathcal{X}}\left(\mathbf{x}, \mathbf{x}'\right)^p + d_{\mathcal{Y}}\left(\mathbf{y}, \mathbf{y}'\right)^p\right)^{1/p}$$

where $\mathbf{z} = (\mathbf{x}, \mathbf{y})$ is a feature-label pair and $p \geq 1$. Euclidean distance is employed for metric $d_{\mathcal{X}}$ on the feature space $\mathcal{X}$. For $d_{\mathcal{Y}}$, labels are regarded as distributions over $\mathcal{X}$ where samples with label $\mathbf{y}$ are drawn. $d_{\mathcal{Y}}$ is measured through p-Wasserstein distance between distributions of labels. The distance between dataset $\mathcal{D}$ and $\mathcal{D}'$ is calculated as

$$d_{\mathrm{OT}}\left(\mathcal{D}, \mathcal{D}'\right) = \min_{\pi \in \Pi(\mathcal{D}, \mathcal{D}')} \int_{\mathcal{Z} \times \mathcal{Z}} d_{\mathcal{Z}}\left(\mathbf{z}, \mathbf{z}'\right) \pi(\mathbf{z}, \mathbf{z}')$$

where π is the coupling matrix. For more details, see Alvarez-Melis and Fusi (2020).

A.3 Validation of Syntactic Difference in mBERT

Figure 11 and Figure 12 show the comparison of syntactic difference derived from two baselines and the formal syntactic distance. The lower Spearman's ρ indicates that the similarities and differences between languages are not well captured by these baselines.

B Additional Materials for Mechanism behind Cross-Lingual Transfer

B.1 Comparison with Baselines

Figure 13 and Figure 14 are comparisons of the syntactic difference induced by two baselines

[16]https://scikit-learn.org/stable/modules/generated/sklearn.manifold.TSNE.html

[17]As t-SNE can be slow for high-dimensional data, the representations are first projected to 37 dimensions via PCA and then visualized through t-SNE.

Language	ar	bg	de	el	en	es	et	fa	fi	fr	he	hi
Layer 7	83.3	89.3	87.0	92.0	88.0	88.9	77.3	88.0	79.3	90.1	85.8	86.0
MBERT0	60.1	60.7	68.2	61.1	69.2	66.4	47.8	62.7	49.9	67.8	58.2	60.2
MBERTRAND	58.8	61.6	69.0	62.5	70.2	67.4	47.7	61.9	50.5	68.3	59.1	60.8
Language	**it**	**ja**	**ko**	**lv**	**nl**	**pl**	**pt**	**ro**	**ru**	**tr**	**vi**	**zh**
Layer 7	88.7	86.5	77.8	79.8	87.1	86.4	92.5	86.4	89.2	73.7	70.8	84.3
MBERT0	65.6	61.2	45.4	50.6	62.2	54.8	68.4	56.3	60.9	50.1	58.6	56.3
MBERTRAND	66.0	61.6	46.2	51.1	63.8	57.0	69.1	58.2	61.9	51.2	59.6	57.9

Table 3: Comparison of the 7th layer of mBERT and the two baselines. We take the best layer of MBERTRAND for comparison.

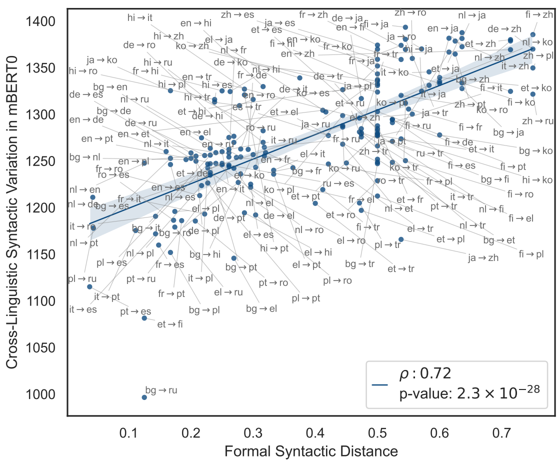

Figure 11: Comparison of formal syntactic distance and cross-linguistic syntactic difference derived from MBERT0.

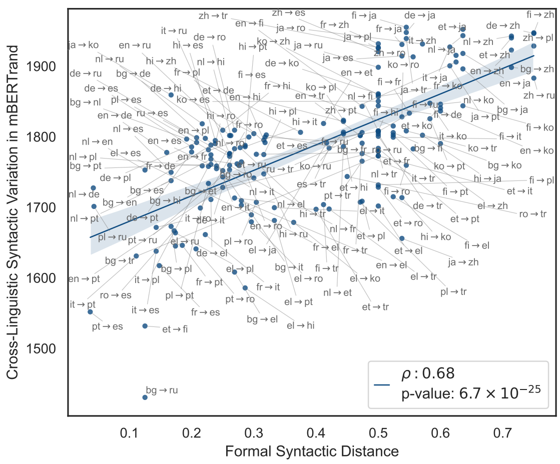

Figure 12: Comparison of formal syntactic distance and cross-linguistic syntactic difference derived from MBERTRAND.

and the performance drop in zero-shot cross-lingual transfer performance of dependency parsing (Section 3.3).

C Additional Materials for Impact of Linguistic Diversity

C.1 Typological Features

Table 5 shows the morphosyntactic features we employ in Section 4. We delete Feature 95A: *Relationship between the Order of Object and Verb and the Order of Adposition and Noun Phrase*, 96A: *Relationship between the Order of Object and Verb and the Order of Relative Clause and Noun* and 97A: *Relationship between the Order of Object and Verb and the Order of Adjective and Noun* as they can be inferred from other features in the area of word order.

C.2 Importance of Typological Features

The feature importance of all the morphosyntactic features we use is shown in Figure 15.

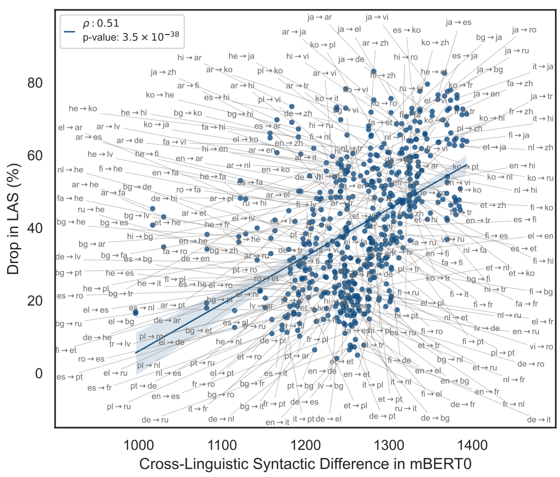

Figure 13: Comparison of the cross-linguistic syntactic difference in MBERT0 and drop in zero-shot cross-lingual transfer performance (LAS).

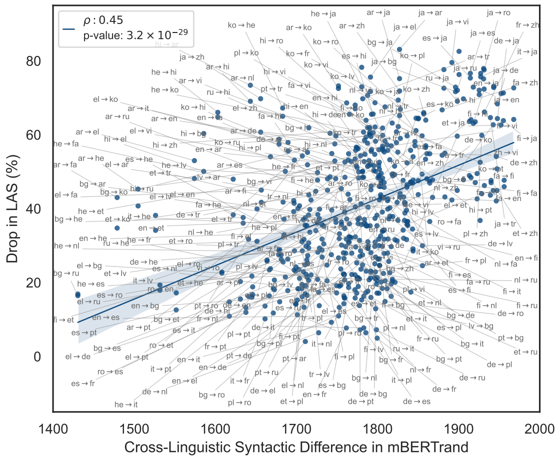

Figure 14: Comparison of the cross-linguistic syntactic difference in MBERTRAND and drop in zero-shot cross-lingual transfer performance (LAS).

D Implementation Details

Multilingual BERT We use the pretrained *bert-base-multilingual-cased* model[18] for all our experiments.

Grammatical relation probe We train the linear classifier via stochastic gradient descent[19] to classify the grammatical relations between a head-dependent word pair. We use logistic regression, set the max number of iteration to 10000 and allow for early stopping. We report the 95% confidence interval computed based on different regularization strengths (1.e-09, 1.e-08, 1.e-07, 1.e-06, 1.e-05, 1.e-04, 1.e-03, and 1.e-02) in Figure 1.

Measure of the syntactic difference in mBERT We use the public source code of Alvarez-Melis and Fusi (2020)[20] to compute the syntactic difference in mBERT. The p-Wasserstein distance ($p = 2$) is computed based on Sinkhorn algorithm (Cuturi, 2013) and the entropy regularization strength is set to 1e-1.

Dependency parsing We follow the setup of Wu and Dredze (2019), which replaces the LSTM encoder in Dozat and Manning (2017) with mBERT. For each language, we train the model with ten epochs and validate it at the end of each epoch. We choose the model performing the best (i.e., achieving the highest LAS) on the development set. We use the Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.99$, eps $= 1 \times 10^{-8}$, and a learning rate of 5e-5. The batch size is 16 and the max sequence length is 128.

Gradient boosting regressor We use a gradient boosting regressor[21] with 100 estimators and each has a maximum depth of 3. We use the squared error for regression with the default learning rate of 1e-1.

E Data for Experiments

E.1 Universal Dependencies Treebanks

Table 4 shows the languages and UD treebanks (version 2.8)[22] we use. We follow the split of training, development and test set in UD.

[18] https://huggingface.co/bert-base-multilingual-cased

[19] https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.SGDClassifier.html

[20] https://github.com/microsoft/otdd

[21] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingRegressor.html

[22] https://lindat.mff.cuni.cz/repository/xmlui/bitstream/handle/11234/1-3687/ud-treebanks-v2.8.tgz?sequence=1&isAllowed=y

Language	Abbr.	Language Family	UD Treebanks	#Sentences	#Tokens
Arabic†	ar†	Afro-Asiatic.Semitic	Arabic-PADT†	7,664	282,384
Bulgarian	bg	IE.Balto-Slavik	Bulgarian-BTB	11,138	156,149
Catalan*	ca*	IE.Romance	Catalan-AnCora*	166,678	530,767
Czech*	cs*	IE.Balto-Slavic	Czech-PDT*	87,913	1,503,732
German	de	IE.Germanic	German-GSD	15,590	287,740
Greek	el	IE.Greek	Greek-GDT	2,521	61,773
English	en	IE.Germanic	English-EWT	16,621	251,494
Spanish	es	IE.Romance	Spanish-GSD	16,013	423,346
Estonian	et	Uralic	Estonian-EDT	30,972	437,767
Persian (Farsi)†	fa†	IE.Indo-Iranian	Persian-PerDT†	29,107	494,163
Finnish	fi	Uralic	Finnish-TDT	151,136	201,950
French	fr	IE.Romance	French-GSD	16,341	389,224
Hebrew†	he†	Afro-Asiatic.Semitic	Hebrew-HTB†	6,216	115,529
Hindi	hi	IE.Indo-Iranian	Hindi-HDTB	16,647	351,704
Hungarian*	hu*	Uralic	Hungarian-Szeged*	1,800	42,032
Italian	it	IE.Romance	Italian-VIT	10,087	259,479
Japanese	ja	Japonic	Japanese-GSD	8,100	193,654
Korean	ko	Koreanic	Korean-Kaist	27,363	350,090
Latvian†	lv†	IE.Balto-Slavic	Latvian-LVTB†	15,351	252,334
Dutch	nl	IE.Germanic	Dutch-Alpino	13,603	208,613
Polish	pl	IE.Balto-Slavic	Polish-PDB	22,152	347,377
Portuguese	pt	IE.Romance	Portuguese-GSD	12,078	297,938
Romanian	ro	IE.Romance	Romanian-RRT	9,524	218,511
Russian	ru	IE.Balto-Slavic	Russian-GSD	5,030	98,000
Tamil*	ta*	Dravidian	Tamil-TTB*	600	8,635
Turkish	tr	Turkic	Turkish-BOUN	9,761	121,214
Urdu*	ur*	IE.Indo-Iranian	Urdu-UDTB*	5,130	138,077
Vietnamese†	vi†	Austroasiatic.Vietic	Vietnamese-VTB†	3,000	43,754
Chinese (Mandarin)	zh	Sino-Tibetan.Sinitic	Chinese-GSDSimp	4,997	123,291

Table 4: Languages and UD Treebanks we use. Languages marked with a dagger (†) aren't involved in the comparison with formal syntactic distance due to lack of corresponding data in Ceolin et al. (2020). Languages used for test of the strategy for source language selection in Section 4 is marked with an asterisk ($*$). The phylogenetic information is obtained from Glottolog (Hammarström et al., 2021). IE stands for the Indo-European family.

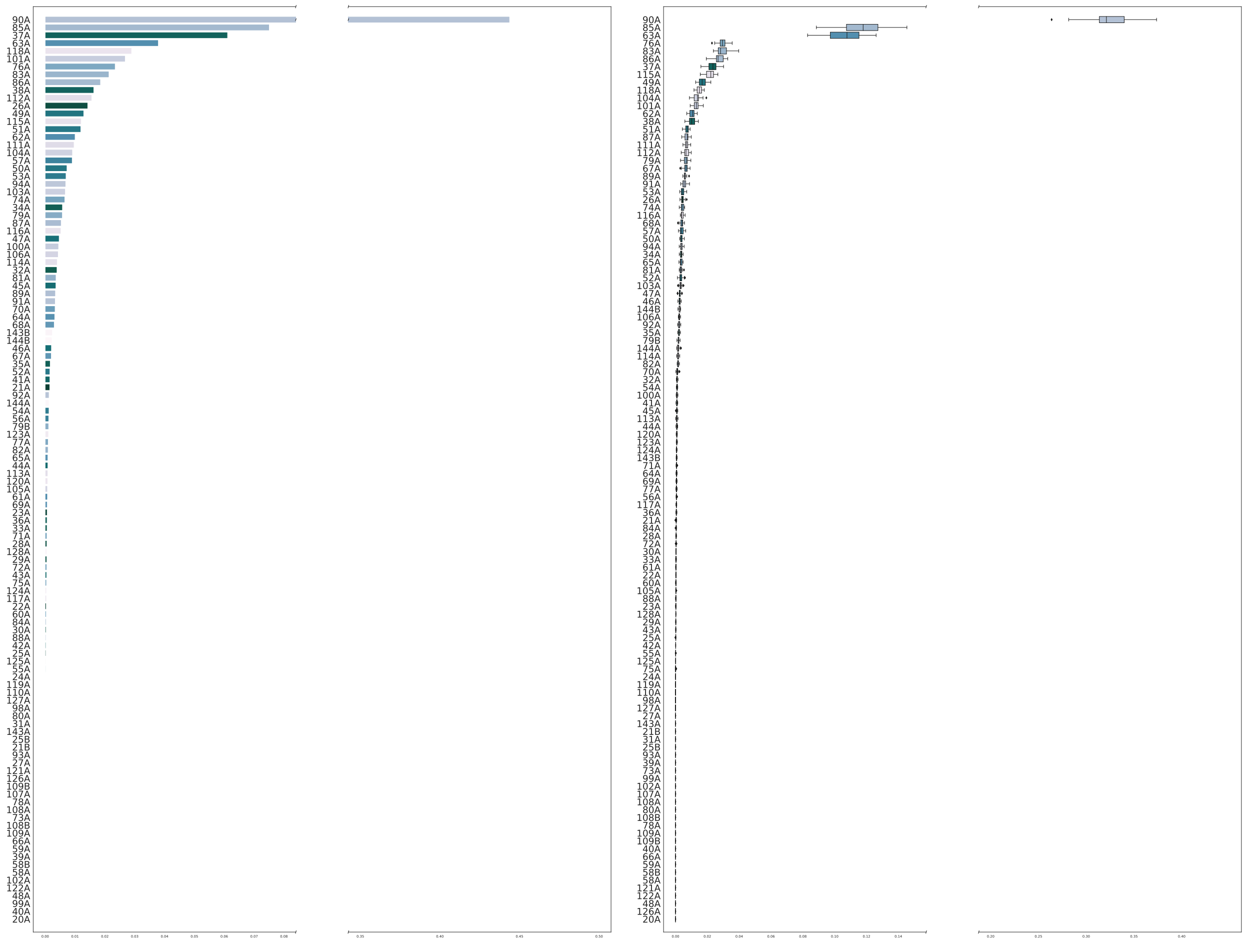

Figure 15: Rank of importance of all the morphosyntactic features we use.

Table 5: Features in WALS used in our work. As WALS entries can be sparse, we provide in the column **#Languages** information about how many languages involved in the experiment (Section 4) have a valid entry for the feature. The **left** side of "/" indicates the number of languages for which the feature is not missing among the languages involved in the training procedure of the gradient boosting regressor, including Arabic, Bulgarian, German, Greek, English, Spanish, Estonian, Persian, Finnish, French, Hebrew, Hindi, Italian, Japanese, Korean, Latvian, Dutch, Polish, Portuguese, Romanian, Russian, Turkish and Chinese. For the **right** side, the five languages used to test the strategy for source language selection is involved, including Czech, Catalan, Hungarian, Tamil and Urdu.

ID	Name	#Languages
20A	Fusion of Selected Inflectional Formatives	15 / 16
21A	Exponence of Selected Inflectional Formatives	15 / 16
21B	Exponence of Tense-Aspect-Mood Inflection	15 / 16
22A	Inflectional Synthesis of the Verb	15 / 16
23A	Locus of Marking in the Clause	15 / 16
24A	Locus of Marking in Possessive Noun Phrases	15 / 16
25A	Locus of Marking: Whole-language Typology	15 / 16
25B	Zero Marking of A and P Arguments	15 / 16
26A	Prefixing vs. Suffixing in Inflectional Morphology	23 / 27
27A	Reduplication	17 / 20
28A	Case Syncretism	16 / 17
29A	Syncretism in Verbal Person/Number Marking	16 / 17
30A	Number of Genders	14 / 16
31A	Sex-based and Non-sex-based Gender Systems	14 / 16
32A	Systems of Gender Assignment	14 / 16
33A	Coding of Nominal Plurality	23 / 27
34A	Occurrence of Nominal Plurality	18 / 20
35A	Plurality in Independent Personal Pronouns	16 / 18
36A	The Associative Plural	21 / 22
37A	Definite Articles	22 / 25
38A	Indefinite Articles	20 / 24
39A	Inclusive/Exclusive Distinction in Independent Pronouns	16 / 17
40A	Inclusive/Exclusive Distinction in Verbal Inflection	16 / 17
41A	Distance Contrasts in Demonstratives	16 / 20
42A	Pronominal and Adnominal Demonstratives	16 / 20
43A	Third Person Pronouns and Demonstratives	14 / 15
44A	Gender Distinctions in Independent Personal Pronouns	19 / 20
45A	Politeness Distinctions in Pronouns	21 / 24
46A	Indefinite Pronouns	23 / 25
47A	Intensifiers and Reflexive Pronouns	22 / 25
48A	Person Marking on Adpositions	19 / 20
49A	Number of Cases	21 / 24
50A	Asymmetrical Case-Marking	21 / 24
51A	Position of Case Affixes	23 / 27
52A	Comitatives and Instrumentals	20 / 24
53A	Ordinal Numerals	23 / 27
54A	Distributive Numerals	20 / 23
55A	Numeral Classifiers	15 / 16
56A	Conjunctions and Universal Quantifiers	12 / 14
57A	Position of Pronominal Possessive Affixes	18 / 20
58A	Obligatory Possessive Inflection	15 / 16
58B	Number of Possessive Nouns	15 / 16
59A	Possessive Classification	15 / 16

ID	Name	#Languages
60A	Genitives, Adjectives and Relative Clauses	11 / 12
61A	Adjectives without Nouns	13 / 14
62A	Action Nominal Constructions	19 / 21
63A	Noun Phrase Conjunction	20 / 22
64A	Nominal and Verbal Conjunction	17 / 19
65A	Perfective/Imperfective Aspect	19 / 21
66A	The Past Tense	19 / 21
67A	The Future Tense	19 / 21
68A	The Perfect	19 / 21
69A	Position of Tense-Aspect Affixes	23 / 27
70A	The Morphological Imperative	23 / 27
71A	The Prohibitive	23 / 27
72A	Imperative-Hortative Systems	23 / 26
73A	The Optative	18 / 21
74A	Situational Possibility	21 / 24
75A	Epistemic Possibility	21 / 24
76A	Overlap between Situational and Epistemic Modal Marking	21 / 24
77A	Semantic Distinctions of Evidentiality	20 / 22
78A	Coding of Evidentiality	20 / 22
79A	Suppletion According to Tense and Aspect	19 / 21
79B	Suppletion in Imperatives and Hortatives	19 / 21
80A	Verbal Number and Suppletion	19 / 21
81A	Order of Subject, Object and Verb	23 / 28
82A	Order of Subject and Verb	23 / 28
83A	Order of Object and Verb	23 / 28
84A	Order of Object, Oblique, and Verb	12 / 13
85A	Order of Adposition and Noun Phrase	23 / 28
86A	Order of Genitive and Noun	23 / 28
87A	Order of Adjective and Noun	23 / 28
88A	Order of Demonstrative and Noun	23 / 28
89A	Order of Numeral and Noun	22 / 27
90A	Order of Relative Clause and Noun	23 / 28
91A	Order of Degree Word and Adjective	22 / 25
92A	Position of Polar Question Particles	23 / 27
93A	Position of Interrogative Phrases in Content Questions	20 / 24
94A	Order of Adverbial Subordinator and Clause	21 / 25
98A	Alignment of Case Marking of Full Noun Phrases	16 / 17
99A	Alignment of Case Marking of Pronouns	16 / 17
100A	Alignment of Verbal Person Marking	19 / 20
101A	Expression of Pronominal Subjects	21 / 24
102A	Verbal Person Marking	19 / 20
103A	Third Person Zero of Verbal Person Marking	19 / 20
104A	Order of Person Markers on the Verb	19 / 20
105A	Ditransitive Constructions: The Verb 'Give'	17 / 19
106A	Reciprocal Constructions	17 / 18
107A	Passive Constructions	19 / 20
108A	Antipassive Constructions	16 / 18
108B	Productivity of the Antipassive Construction	16 / 18
109A	Applicative Constructions	16 / 18
109B	Other Roles of Applied Objects	16 / 18

ID	Name	#Languages
110A	Periphrastic Causative Constructions	13 / 15
111A	Nonperiphrastic Causative Constructions	16 / 18
112A	Negative Morphemes	23 / 27
113A	Symmetric and Asymmetric Standard Negation	17 / 18
114A	Subtypes of Asymmetric Standard Negation	17 / 18
115A	Negative Indefinite Pronouns and Predicate Negation	22 / 25
116A	Polar Questions	23 / 28
117A	Predicative Possession	17 / 20
118A	Predicative Adjectives	20 / 23
119A	Nominal and Locational Predication	20 / 23
120A	Zero Copula for Predicate Nominals	20 / 23
121A	Comparative Constructions	15 / 17
122A	Relativization on Subjects	18 / 19
123A	Relativization on Obliques	18 / 19
124A	’Want’ Complement Subjects	18 / 19
125A	Purpose Clauses	15 / 17
126A	’When’ Clauses	16 / 18
127A	Reason Clauses	16 / 18
128A	Utterance Complement Clauses	15 / 17
143A	Order of Negative Morpheme and Verb	23 / 28
143B	Obligatory Double Negation	23 / 28
144A	Position of Negative Word With Respect to Subject, Object, and Verb	23 / 28
144B	Position of negative words relative to beginning and end of clause and with respect to adjacency to verb	23 / 28