On Language Spaces, Scales and Cross-Lingual Transfer of UD Parsers

Tanja Samardžić[1] **Ximena Gutierrez-Vasques**[1] **Rob van der Goot**[2]

Max Müller-Eberstein[2] **Olga Pelloni**[1] **Barbara Plank**[2,3]

[1] Text Group, URPP Language and Space, University of Zurich, Switzerland
[2] Department of Computer Science, IT University of Copenhagen, Denmark
[3] Center for Information and Language Processing, LMU Munich, Germany
{tanja.samardzic, ximena.gutierrezvasques, olga.pelloni}@uzh.ch
{robv, mamy}@itu.dk
b.plank@lmu.de

Abstract

Cross-lingual transfer of parsing models has been shown to work well for several closely-related languages, but predicting the success in other cases remains hard. Our study is a comprehensive analysis of the impact of linguistic distance on the transfer of Universal Dependencies (UD) parsers. As an alternative to syntactic typological distances extracted from URIEL, we propose three text-based feature spaces and show that they can be more precise predictors, especially on a more local scale, when only shorter distances are taken into account. Our analysis also reveals that the good coverage in typological databases is not among the factors that explain good transfer.[1]

1 Introduction

The goal of cross-lingual parsing is to process a target language as well as possible by exploiting training data available from (an)other language(s). While we know that parsing models can be transferred well across some well-known closely related languages (de Lhoneux et al., 2018), the success of cross-lingual transfer in all other cases remains hard to predict. Surprising cases of syntactic transfer between unrelated languages such as Irish and Indonesian (Lynn et al., 2014) illustrate well this unpredictability.

A possible explanation for such cases is that genealogically unrelated languages can still be similar enough to allow transfer. But what is the relevant measure of language similarity in such cases? One possible solution is to rely on language features stored in typological databases such as WALS (Dryer and Haspelmath, 2013; Comrie et al., 2013) or Glottolog (Hammarström et al., 2018). Taking these features as vector representations, languages can be embedded and compared regardless of their genealogical relations. A popular library URIEL (Littell et al., 2017) has facilitated the use of typological features to measure similarity between languages at different levels (phonology, syntax, geographical distribution). The problem with this solution is that the information in linguistic databases is often incomplete and unevenly distributed. Some languages are fully described, while only a few feature values are known for others (Ponti et al., 2019). Nevertheless, a study by Lauscher et al. (2020) on transferring models from English to several other languages suggests that the URIEL language similarity score is a good predictor of cross-lingual transfer for parsing Universal Dependencies (UD).

Our study brings a comprehensive analysis of the relationship between language similarity and the cross-lingual transfer in UD parsing. It extends previous work in two directions: first, we cover many more languages than any previous study (which are typically limited to a small set); second, we compare the URIEL representation with three text-based alternatives. These extensions allow us to ask new questions such as: What should we do for languages that do not have close relatives? Do measures of language similarity predict the transfer at any scale (for close and for distant languages)? Are there good alternatives to linguistic databases for measuring language similarity? We perform correlation tests between linguistic distances and parsing scores on various samples of UD treebanks designed to neutralize two kinds of biases. First, we balance the samples at the level of language, genus, and family,[2] reducing gradually the known bias of the UD towards Indo-European languages. Second,

[1] The analysis notebooks are available at https://github.com/MorphDiv/transfer-lang.

[2] Genus and family are two levels of language genealogy commonly used to group languages of the world. A list of families and genera can be found at https://wals.info/languoid/genealogy.

Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL), pages 266 - 281

we investigate the impact of the scale by comparing global correlations (considering a whole language space) with local correlations (considering smaller partitions of a language space).

We show that typological distances extracted from URIEL are reasonably good global predictors, while text-based distances are better local predictors. A surprising outcome of our analysis concerns the uneven coverage of languages in typological databases: most of the UD languages with many missing features are Indo-European. On the other hand, good database coverage does not guarantee good predictability of transfer for the languages outside of the Indo-European family.

2 Related Work

Thanks to evident structural alignments between languages the possibility of transferring syntactic parsing models across languages was investigated even before the wide-spread adoption of pre-trained language models in NLP (McDonald et al., 2006; Zeman and Resnik, 2008). However, this task proved non-trivial because such clear alignments tend to be found in similar languages, but are much rarer overall (Seeker and Kuhn, 2013; Goldberg and Elhadad, 2013).

The idea of using data from another language or a set of languages to improve syntactic parsing on any given language is tempting because annotated data is not available for the majority of the world's languages. Early work typically focused on several languages selected according to the availability of training data. In the meantime, the Universal Dependencies (UD) treebanks have become available for many different languages (Zeman et al., 2021)[3] opening the question of what language pairs are most suitable for model transfer. Most of the time, polyglot[4] models are trained on multiple languages, but preserving the identity of the languages (by adding the language ID to the text representation) turns out useful (Ammar et al., 2016). Smith et al. (2018) cluster languages according to similarity before training polyglot models. Cross-lingual parameter sharing is found to improve the performance overall, but especially for closely-related languages, which can share parameters in different layers of neural representation (de Lhoneux et al., 2018; van der Goot and de Lhoneux, 2021).

Cross-lingual transfer started being explored in other tasks too after the introduction of large pretrained models (Pires et al., 2019), making the question of linguistic similarity relevant to a more general scope of NLP research. Lin et al. (2019) propose a range of measures that can be used in order to choose the best transfer language, which they divide into data-dependent (data size, token overlap, TTR) and data independent (various distance measures extracted from the URIEL database). Lauscher et al. (2020) study how well different similarity scores predict the success of the transfer on different tasks (with mBERT and XLM-R as pretrained models) and find that syntactic features extracted from URIEL correlate strongly with the zero-shot cross-lingual UD parsing performance. Interestingly, these features are better predictors than genealogical relatedness, but data-dependent measures, such as the size of the training data, seem to predict better the cross-lingual zero-shot performance on other tasks such as XQuAD (Artetxe et al., 2020; Rajpurkar et al., 2016) or XNLI (Conneau et al., 2018; Bowman et al., 2015; Williams et al., 2018). While English turns out to be a good transfer language for many tasks due to the size of the training data, Turc et al. (2021) show that German is a better transfer language than English for quite a few, even less-related, languages. The fact that English is not the best transfer language on the task of part-of-speech (POS) tagging is confirmed by the most wide-scope study of cross-lingual transfer up to now (de Vries et al., 2022). Similarly to Lauscher et al. (2020), this study too finds that a surface string similarity measure (LDND distance, Wichmann et al. (2010)) is a better predictor of the transfer than genealogical relatedness. Somewhat contrary to this, Kudugunta et al. (2019) find an interesting genealogical clustering in the representations created by machine translation models.

Having counted mentions of successful cross-lingual transfer on many different tasks in the previous works (Ruder et al., 2021; Turc et al., 2021; Vázquez et al., 2021; Hu et al., 2020; Lauscher et al., 2020; Lin et al., 2019; Paul et al., 2013), we notice that English is most frequently mentioned as the best transfer language overall, but these mentions are almost entirely related to European target languages. For targets located outside of Europe, the best transfer languages are different and hard to predict. For instance, Greek is a good transfer language for Thai and Hindi, while Russian works

[3] http://universaldependencies.org

[4] We here used the term *polyglot* model (Mulcaire et al., 2019) most often also referred to as *multilingual* model.

well for these two languages and Arabic.

Our study shares the wide cross-lingual scope with de Vries et al. (2022). In contrast to their work, we focus on syntactic parsing models, rather than POS tagging. We follow some other previous studies in working with both typological and text-based language similarity measures, but our text-based measures can be regarded as generic rather than data-dependent and can be used as an alternative to URIEL in many cases.

3 Language Spaces and Similarity: Genealogy, Typology, Text

The most widely accepted method for comparing languages relies on *genealogical* classification: we consider languages located in the same region of a phylogenetic tree to be similar. This method currently prevails in NLP. Practitioners often discuss language similarity in terms of language family (Ponti et al., 2019; Tan et al., 2019; Shaffer, 2021). However, language families can be too broad for a meaningful comparison as they include typologically very different languages. For instance, English and Armenian belong to the same family (Indo-European), but are very different in terms of phoneme inventories, morphology, and word order. On the other hand, languages can be rather similar even if they are genealogically unrelated. For example, Bulgarian is closely related to other Slavic languages, but its morphology, word order and the use of the definite article makes it more similar to English than to other Slavic languages.

Typological features and geographical placement of languages can be regarded as potentially more objective and fine-grained alternatives to genealogical similarity. In other words, genealogically unrelated languages can turn out to be close in a typological vector space or in the geographical (physical) space. It is less common to have an intuitive perception of languages that are close in such spaces as similar, but typological proximity seems to be more useful as a predictor of cross-lingual transfer than genealogical relatedness (see Section 2).

The URIEL database and its associated Python library `lang2vec` (Littell et al., 2017) are very convenient resources for measuring the distance between languages in all of these spaces. URIEL combines features from several linguistic databases: Ethnologue (Lewis et al., 2015), Glottolog, PHOIBLE (Moran et al., 2014), SSWL[5] and WALS. It describes over 4,000 languages, but the available information strongly depends on the types of features. For example, geographic and genealogical feature values are known for all languages, while syntactic feature values, which are relevant to our study, are often missing.

When assessing linguistic similarity with `lang2vec`, one can use various subsets of features and the `knn` prediction option to fill in the missing features, which is what is typically used in previous research. With this option, all feature slots are filled with some predicted value. If a value is missing for some feature, the corresponding value from the most similar language (nearest neighbor) is returned. We work with the union of syntactic features (WALS + SSWL) completed with the `knn` prediction, but we also analyze the coverage of the UD languages in the URIEL sources by extracting the values before the `knn` prediction.

Text-based features can be regarded as a potential alternative to the features extracted from typological databases. Type-token ratio (TTR), for instance, is higher in morphologically rich than in morphologically poor languages and can be used for language comparison when the data size is controlled (Biber, 1988; Tweedie and Baayen, 1998; Bentz et al., 2017). Other text statistics, such as the *mean word length* (MWL) are also characteristic of languages (words are longer in morphologically rich languages), while being even less dependent on the data. In the work on cross-lingual transfer, it is common to consider all text-based measures to be *data-dependent* as opposed to typological measures, which are *data-independent* (Lin et al., 2019).[6] We assume that text-based features can reach various levels of data-independence, while providing a means for measuring language similarity at a more fine-grained level.

In the remainder of this section, we describe two text-based measures that we propose for comparing languages at two structural levels, morphology and syntax. Our morphological measure is more generic than the syntactic measure, which is more data-dependent.

[5] *Syntactic Structures of the World's Languages* by Chris Collins and Richard Kayne

[6] In NLP, data-dependent measures require access to text samples of the languages to estimate similarity statistics, which are viewed as specific to the samples (not easily generalized). In contrast to this, data-independent measures are often derived from data or linguistic observations yet the text sample is not required at estimation time.

3.1 The Language Space of BPE Subword Productivity

Capturing morphological phenomena, this measure departs from the observation that subword tokenization with BPE compresses the text vocabulary in a way that depends on typological properties of languages. Analyzing subword tokens formed in the first few hundred merges (Gutierrez-Vasques et al., 2021), we can distinguish between languages that have productive morphology (e.g. Hungarian), from languages that form words in a more idiosyncratic fashion (e.g. Chinese).

Following this intuition, we describe each language in terms of three features calculated over the tokens formed in the first 200 BPE merges. The first feature, *subword productivity* is the number of word types in which a subword appears. The second feature, *subword frequency* is the cumulative frequency of all word types in which a given subword appears. The third feature, *subword idiosyncrasy* is the ratio between the subword frequency and the subword productivity. A single vector representation for each language is constructed by averaging the values of all subword tokens. The resulting three-dimensional vectors are centered around zero and scaled with respect to the standard deviation. In this way, we construct a new space for comparing languages distinguishing between morphological types such as analytic, synthetic, and polysynthetic languages.

It is noteworthy that this approach does not depend on access to the information extracted from grammars and stored in typological databases. It also does not require any annotation: the scores are extracted directly from a relatively small sample of raw text (e.g. 50,000 words, fixed for our UD samples) in an unsupervised fashion. It thus provides a good alternative to hand-crafted descriptions which are hard to obtain. The drawback of this method is that it captures morphological features, which, despite the known universal trade-offs between syntax and morphology (Sinnemäki, 2010; Ehret and Szmrecsanyi, 2016; Futrell et al., 2015), might not be the most useful features for predicting the transfer of syntax.

3.2 The Language Space of Dependency Probes

To obtain text-based features capturing more precisely syntactic phenomena, we make use of *syntactic probes*, minimal models that can perform the dependency parsing task at hand. In constructing a language space with dependency probes, we build on the DepProbe approach of Müller-Eberstein et al. (2022) and the intuition that linear subspaces capture syntactic information while being much easier to interpret than the parameters of full parsers. Measuring the similarity of these linear subspaces using subspace angles (Knyazev and Argentati, 2002), we can further compare whether dependency structures and relations are represented similarly or dissimilarly across languages — even across unrelated languages not covered by manual typological annotations — which is crucial for cross-lingual transferability.

Conceptually, each probe contains the information on how pre-trained embeddings map to dependency structures. Therefore, similar mappings are expected to indicate similar languages. Comparing these subspaces for the purpose of transferability estimation has shown to be highly predictive (Müller-Eberstein et al., 2022). We rely on the same intuition, but use the probes for a different purpose: instead of predicting the performance of a full parser, which was the main goal for Müller-Eberstein et al. (2022), we see the probes as a sort of language embeddings for comparing different languages. This leads us to extend this initial study to the full set of languages in UD, and to analyze how these data-driven measures relate to linguistically motivated typological information.

There has been debate regarding what constitutes an appropriately parametrized probe (Hewitt and Liang, 2019; Voita and Titov, 2020). We follow the most common linear probing paradigm for dependency parsing by Hewitt and Manning (2019). It can be seen as learning a linear subspace within the existing, pre-trained latent space in which dependency information is particularly salient. For DepProbe specifically, these are the dependency structural subspace A and the dependency relational subspace L, which are respectively learned using the mean square error and cross-entropy loss to the target dependency tree. This approach is intermediary to training a full parser, which is computationally expensive, and manual features such as those from URIEL, which may lack coverage of the specific language variant used in any particular treebank. However, this measure requires at least some syntactically annotated data.

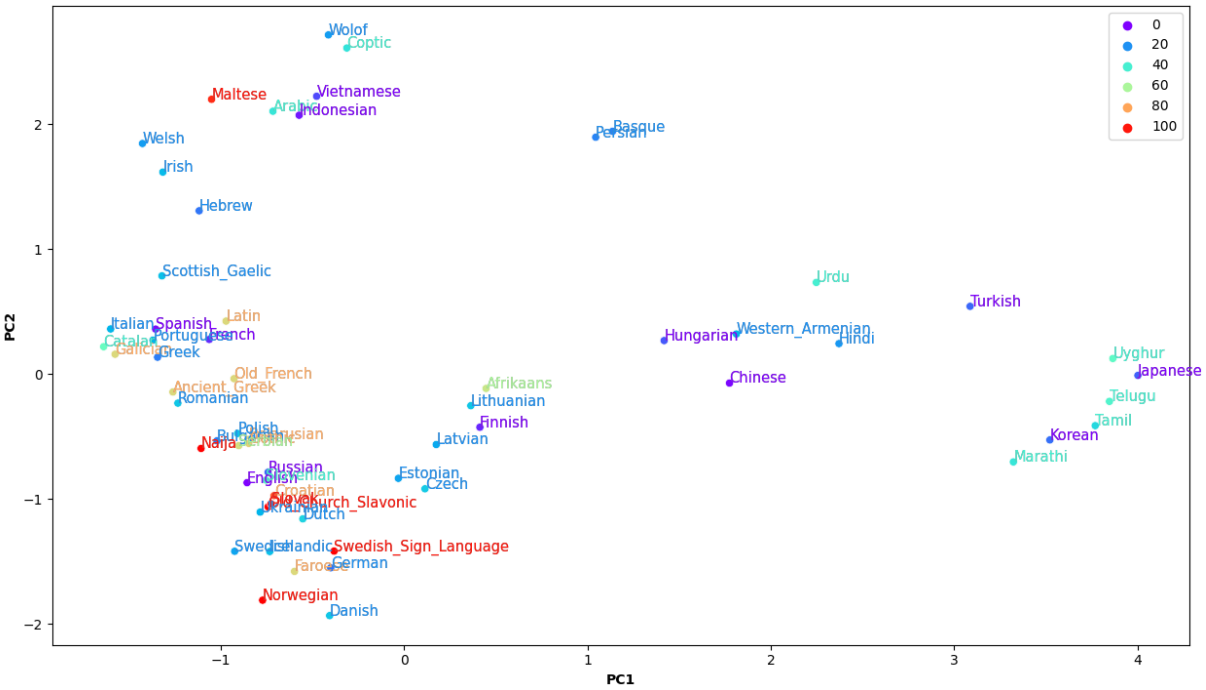

Figure 1: URIEL embeddings (reduced to 2 dimensions using PCA) for 62 UD languages that appear as target languages in our experiments. The color indicates the percentage of missing features in the URIEL sources. Languages with most missing features are located in the densely populated regions.

URIEL	Squared Euclidean distance in the URIEL space using WALS+SSWL syntax features and KNN prediction
probe-A	The distance between dependency probes trained on our UD samples to predict the dependency link (attachment)
probe-L	The distance between dependency probes trained on our UD samples to predict the dependency label
BPE	Squared Euclidean distance in the BPE productivity space constructed from the raw text extracted from our UD samples
MWL	The difference between mean word lengths, estimated on the raw text extracted from our UD samples as the average number of characters per word token in a treebank
MSL	The difference between mean sentence lengths, estimated on the raw text extracted from our UD samples as the average number of word tokens per sentence in a treebank

Table 1: Linguistic distances and baselines as experimental settings. Note: the MWL and MSL differences are, in fact, distances in a monodimensional space.

4 Data and Methods

From the linguistic spaces and measures described in Section 3, we create distance matrices. We then calculate multiple correlation scores between each of the linguistic distance matrices on one side and the scores obtained while testing parsers on a set of languages on the other. For each pair transfer-target language, we have one labeled attachment score (LAS), which we name xLAS in our experiments to underline the fact that these scores are obtained via cross-lingual transfer.[7] We expect higher xLAS scores when linguistic distances are smaller, thus a negative correlation.

In this section, we describe the details of the experimental design and the analyses.

4.1 Data

We carry out all our experiments on the Universal Dependencies V2.9 data (Zeman et al., 2021), and the additional unofficial set of treebanks used in van der Goot et al. (2021). In total our data has 116 languages in 223 treebanks. We removed all multi-word tokens with `ud-conversion-tools`.[8]

[7] We exclude all self-tranfer cases.

[8] Code-switched pairs are considered a new language as specified by the treebank-creators. Arabic-NYUAD and

Since data size has been identified as a factor that has an impact on cross-lingual transfer, controlling for the data size is necessary in order to isolate potential effects of linguistic distances, which are of interest for our study. We fix the training data size to 50,000 tokens for each transfer language. This size is determined as a good balance between the size of the data needed to achieve a reasonable parsing performance and the availability of the data for different languages. We thus use only treebanks with more than 50,000 tokens for training and cap them to the fixed size. This leaves us with 78 treebanks in 47 languages for training. Because we are not attempting to improve the state-of-the-art in this work and we do not tune the parser, we report our scores on the development data. To cover as many language varieties as possible in our analysis, we decided to use the test data set if there is no development set available for a treebank. On the target side, we have 116 treebanks in 62 languages.

4.2 Parser

To investigate how well linguistic distances defined by the three different language spaces (Section 3) predict the cross-lingual transfer of UD models, we perform zero-shot cross-lingual transfer from each of the 78 transfer treebanks to each of the 116 target treebanks (in a one-to-one setting). For this, we use MaChAmp, an NLP toolkit for training and testing models in a transfer-learning framework. This toolkit uses a transformer based language model as encoder, and can employ multiple decoder heads for multiple tasks. In our setup, we use the default UD model, but remove the morphological tagging and lemmatization task, as not all treebanks have annotation for these tasks. We use MaChAmp v0.3 beta (van der Goot et al., 2021) with default settings and mBERT embeddings (Devlin et al., 2019).

We train a single parser for each of the transfer treebanks, and evaluate on all of the target treebanks using the official CoNLL2018 evaluation script (Zeman et al., 2018). We disable early stopping in all experiments, and take the model after the whole training procedure (20 epochs) to avoid overfitting on the development data. Thus, for each target treebank, we test 78 parsers fine-tuned on transfer treebanks, one parser per transfer treebank. This results in a matrix of 78 $\times$ 116 labeled accuracy scores (xLAS). From these scores, we create various samples on which we then calculate correlation scores. For the 78 datasets, we checked the amount of unknown subwords assigned by the tokenizer of mBERT, which were on average only 0.4%. Outliers are Ancient Greek (~6%) and Old East Slavic (~14%). So, the scripts are mostly covered, and altough some languages might be underrepresented (Rust et al., 2021), at least almost all subwords are represented in the vocabulary.

4.3 Stratified sampling: language, genus, family

Recall that the UD data set is biased towards Indo-European languages in two ways. First, it contains many more treebanks in Indo-European languages than in language from any other family (Nivre et al., 2020). Second, for some languages (and those are usually Indo-European), there are multiple treebanks in the data set, while only single treebanks are available for other languages. To deal with the representation biases in the UD data set, we create stratified samples at three levels. Stratified sampling at the level of *language* means that we select one treebank per language; at the level of *genus* one treebank per genus; at the level of *family*, one treebank per family. The representatives of the three categories are selected randomly, but we repeat the tests 30 times to account for the variance in random sampling. We always report mean correlation scores of 30 random selections. The only level that neutralizes the bias towards Indo-European languages is the level of family, but we perform analyses at all the three levels to see how the scores change between them. Also, the analysis of the scales of the linguistic distances (Section 4.4) is performed only at the level of language.

4.4 Scales: global vs. local

When analyzing the effects of linguistic distances on the cross-lingual parsing scores, we distinguish between two scales. In the first case, which we call the *global scale*, we consider the whole spaces, that is all the data points sampled at the level of language regardless of where they are located in a linguistic space. The global scale thus includes both short and long distances. In the second case, called the *local scale*, we partition the linguistic spaces into smaller regions and consider the correlation scores within each region separately. To make the comparison between different spaces more straightforward, we consider only one partition created with the URIEL space and map all the other linguistic measures to this partition. The local scale

Japanese-BCCWJ are excluded as they are not freely available.

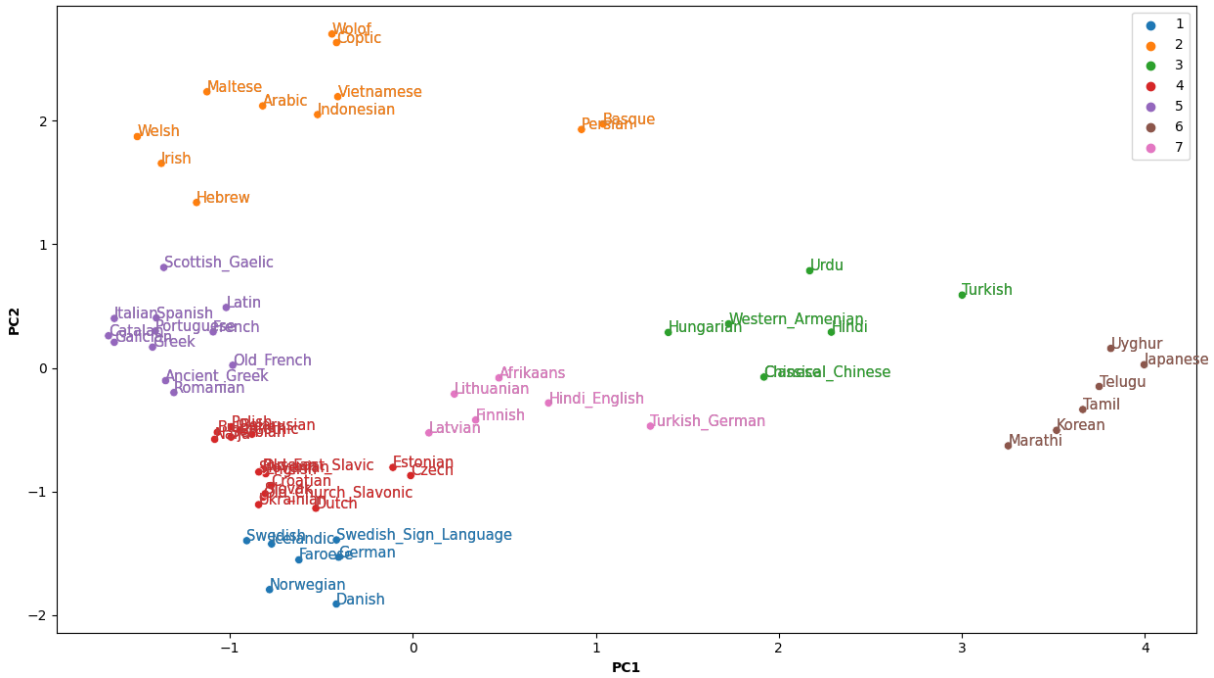

Figure 2: K-means clustering over URIEL embeddings (reduced to 2 dimensions using PCA) for 62 UD languages that appear as target languages in our experiments.

setting thus includes only short distances.

This analysis is motivated by some previous work on the interaction between linguistic variation and geographical phenomena, which has identified potential scale-related limitations. For instance, Jeszenszky et al. (2017) find that traveling times correlate with linguistic distances between Swiss dialects, but the correlation is stronger at shorter distances suggesting a non-linear relationship between the two measures. In other words, traveling times predict linguistic diversity well at short distances, but not so well at longer distances. On the other hand, if a correlation only holds on the global scale, then the observed effect might be driven by (or limited to) a subset of data points, while the rest of the data remains largely unexplained, as pointed out by Moran et al. (2012). Ideally, the correlation scores should not vary depending on the scale and this analysis is expected to show potential limitations of the observed effects.

4.5 Correlation settings

In all our correlation tests, the xLAS scores constitute the predicted variable and the linguistic distances are predictors. When calculating global correlations, we distinguish between three xLAS settings, depending on the sampling level: language, genus, family. Local correlations are only calculated in one setting, language, because other levels would give extremely sparse observations. However, we comment on the phenomena related to linguistic diversity in presenting the results.

Linguistic distance	Correlation with xLAS		
	Language level	Genus level	Family level
URIEL	-0.48	-0.39	-0.35
probe-A	**-0.66**	**-0.53**	**-0.50**
probe-L	-0.57	-0.38	-0.32
BPE	-0.39	-0.26	-0.10
MWL	-0.38	-0.36	-0.34
MSL	-0.12	-0.14	-0.16

Table 2: Global Spearman rank correlation between linguistic distance and xLAS scores. The reported values are the means of 30 random selections.

Table 1 summarizes the settings regarding the linguistic distances. Each of the spaces described in Section 3 is one predictor. In addition to these distances, we perform tests with two kinds of data statistics. We choose MWL as a good representative of text statistics that can be data-independent (see Section 3) and MSL as a representative of data-dependent text statistics.

5 Results

5.1 UD languages in URIEL

Having checked the coverage of the UD languages, we find that more than half of the feature values are missing. We note that missing features are not equally distributed across languages: some languages are well described with over 100 feature values, while for some no syntactic feature values are known. The full list of languages with the counts of missing features is in Appendix A.

To see how the UD languages are distributed in the URIEL space, we create a two-dimensional transformation of the original space with principle component analysis (PCA) and plot in Figure 1 all the languages tested as targets of UD transfer in our experiments (N=62). We color each data point according to the percentage of missing features.

The first thing that can be observed in the plot is a considerable asymmetry in the space density: the most populated area (in the left lower corner) hosts mostly European languages, showing the known bias of the UD data sets. We can also see a considerable covariance between typological, genealogical and geographical factors, which holds only at a very coarse level: Asian languages tend to occupy the right-hand side of the plot, African the upper-left corner. When we zoom in, we see quite a few mismatches between genealogical and syntactic (typological) proximity, especially in the areas outside of the the European corner. For instance, Hungarian and Chinese are rather close in URIEL but they are very far apart in the phylogenetic tree. Interestingly, one such case is the pair Irish-Indonesian mentioned before. Indonesian is an Austronesian language, but it is closer to Irish (which is an Indo-European language) than any other Indo-European language outside of the Celtic group.

Regarding the missing feature values, we notice that all the languages for which more than 50% of feature values are missing are European and their placement with the `knn` prediction is globally correct. At a more fine-grained level, we see some mismatches with what would be expected knowing the properties of languages. For example, Croatian and Serbian are placed rather far apart although they are syntactically identical, genealogically the same language and geographically adjacent. Also, the six languages in the rightmost cluster (Marathi, Korean, Tamil, Telugu, Japanese, Uyghur) come from five different languages families (genealogically distant).

We conclude that the URIEL space represents rather well the knowledge about language similarity globally, but it is rather imprecise at a more fine-grained level.

5.2 Global correlation

Table 2 shows the results of one-to-one correlation tests (one for each predictor). We report the Spearman rank correlation score, which is a non-parametric test best suited for our data. In this setting, we ask how well different linguistic distances predict xLAS scores generally, taking into account the whole spaces. First of all, we can see that the mean sentence (MSL) is the worst predictor despite the fact that its values vary considerably across treebanks. MWL, on the other hand, approaches some of the more elaborated linguistic distances. The values for these two statistics are listed in Appendix B.

The best predictor with solid scores turns out to be `probe-A`, the probe that encodes most of the structural information. This is not very surprising given the fact that the probes are trained to perform lightweight UD parsing. However, it is interesting to see that `probe-A` is a much better predictor than `probe-L` and more consistent across the samples. This means that the representations obtained for a structural task can be regarded as more relevant linguistic features than the representations obtained in a labeling task. The URIEL language space is a reasonably good predictor with moderate scores.[9] The BPE productivity space is close to MWL and sometimes even below it. A reason for this could be the fact that this space captures morphological properties which are not informative enough for predicting xLAS.

All the scores with linguistic distances and MWL decrease with higher sampling levels, which means that the scores at the level of language and genus might still be driven by representation biases in the data. While confirming the expected trends, our results provide a general sense of how big the change is.

5.3 Local correlations

To investigate the impact of the scale on the correlation between linguistic distances and xLAS,

[9] The scores that we observe are considerably lower than what was observed in previous work (Lauscher et al., 2020). This could be due to many reasons since our settings are very different, but it is most likely due to the different sampling approaches.

Linguistic distance	Correlation with xLAS						
	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7
URIEL	-0.35	-0.14	-0.11	-0.42	-0.54	0.03	0.11
probe-A	**-0.86**	**-0.82**	**-0.63**	**-0.83**	**-0.79**	**-0.11**	-0.34
probe-L	-0.71	-0.51	-0.55	0.80	-0.59	-0.08	-0.35
BPE	-0.55	0.11	-0.30	-0.38	-0.55	-0.01	-0.30
MWL	-0.55	-0.09	-0.45	-0.33	-0.18	**-0.11**	**-0.39**
MSL	-0.80	-0.21	-0.44	0.06	-0.30	0.14	0.31

Table 3: Local Spearman rank correlation between linguistic distance and xLAS scores. Cluster obtained from the URIEL space with k-means.

we measure local correlations within smaller areas. Figure 2 shows the partition of the URIEL space obtained by k-means clustering. The local correlation scores are given in Table 3. Dependency probes are still the best predictors within this scope, but the URIEL space is often below BPE and MWL. An important finding of this analysis is the difference in the correlations between the clusters: the correlations are stronger in clusters 1, 4, and 5, while they are very low in the other clusters (except for MWL in the cluster no. 7). An extreme case is the cluster no. 6, where no measure provides any explanation for the xLAS scores. We note that languages in this cluster come from many different families (6 languages from 5 families). The exceptional linguistic diversity is likely to be the reason for this result, but the exact explanation is still to be found. One possible explanation might be that these languages might be wrongly grouped together due to insufficient or inadequate linguistic descriptions in the linguistic databases. This might lead to overestimating their linguistic proximity, while cross-linguistic parser are struggling with real differences. Overall, predicting xLAS scores seems much more straightforward if the languages in a given sample come from the same language family.

6 Conclusion

In this paper, we have shown that various linguistic features can be good predictors of cross-linguistic transfer of UD parsing models. As an alternative to the typological syntactic features extracted from the URIEL database, we propose several text-based features and show that they are often better predictors. Those that encode syntactic structural information by design (dependency probes) are the strongest predictors, while those that capture morphology (BPE, MWL) are comparable to syntactic features extracted from URIEL, especially on a more local scale. In addition to the distance scales, all the scores are impacted by the genealogical composition of the language samples. Explanations for these findings remain an open question for future work.

Limitations

Focusing on the linguistic distances in this paper, we have not addressed the variation in xLAS scores, that is whether it is easier to predict higher than lower scores. Investigating different cases, we noticed that moderate scores seem to be associated with more noise in the correlation analysis, but this effect would need to be quantified and established in a separate study.

Another limitation of our work concerns potential interaction between the predictors that we studied. It might turn out that a combination of two or more of our predictors in a linear model would provide a better explanation for the xLAS scores than any individual predictor. Since we have introduced two novel measures, our principal goal in this paper was to test them in isolation. We leave the question of potential interactions for future work.

Acknowledgements

This research is supported by the Swiss National Science Foundation (SNSF) grant 176305 and by the Independent Research Fund Denmark (DFF) grant 9063-00077B.

References

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, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 4623–4637, Online. Association for Computational Linguistics.

Christian Bentz, Dimitrios Alikaniotis, Michael Cysouw, and Ramon Ferrer-i Cancho. 2017. The entropy of words—learnability and expressivity across more than 1000 languages. Entropy, 19(6):275.

Douglas Biber. 1988. Variation across Speech and Writing. Cambridge University Press.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Bernard Comrie, Matthew S. Dryer, David Gil, and Martin Haspelmath. 2013. Introduction. In Matthew S. Dryer and Martin Haspelmath, editors, The World Atlas of Language Structures Online. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Miryam de Lhoneux, Johannes Bjerva, Isabelle Augenstein, and Anders Søgaard. 2018. Parameter sharing between dependency parsers for related languages. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 4992–4997, Brussels, Belgium. Association for Computational Linguistics.

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.

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

Katharina Ehret and Benedikt Szmrecsanyi. 2016. An information-theoretic approach to assess linguistic complexity. In Complexity, isolation, and variation, pages 71–94. de Gruyter.

Richard Futrell, Kyle Mahowald, and Edward Gibson. 2015. Quantifying word order freedom in dependency corpora. In Proceedings of the third international conference on dependency linguistics (Depling 2015), pages 91–100.

Yoav Goldberg and Michael Elhadad. 2013. Word segmentation, unknown-word resolution, and morphological agreement in a Hebrew parsing system. Computational Linguistics, 39(1):121–160.

Ximena Gutierrez-Vasques, Christian Bentz, Olga Sozinova, and Tanja Samardzic. 2021. From characters to words: the turning point of BPE merges. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, pages 3454–3468, Online. Association for Computational Linguistics.

Harald Hammarström, Robert Forkel, and Martin Haspelmath. 2018. Glottolog 3.3. Leipzig.

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.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In International Conference on Machine Learning, pages 4411–4421. PMLR.

Péter Jeszenszky, Philipp Stoeckle, Elvira Glaser, and Robert Weibel. 2017. Exploring global and local patterns in the correlation of geographic distances and morphosyntactic variation in swiss german. Journal of Linguistic Geography, 5(2):86–108.

Andrew V. Knyazev and Merico E. Argentati. 2002. Principal angles between subspaces in an a-based scalar product: Algorithms and perturbation estimates. SIAM Journal on Scientific Computing, 23(6):2008–2040.

Sneha Kudugunta, Ankur Bapna, Isaac Caswell, and Orhan Firat. 2019. Investigating multilingual NMT representations at scale. 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 1565–1575, Hong Kong, China. 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), pages 4483–4499, Online. Association for Computational Linguistics.

M. Paul Lewis, Gary F. Simons, and Charles D. Fennig, editors. 2015. Ethnologue: Languages of the World, nineteenth edition. SIL International, Dallas, TX, USA.

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.

Teresa Lynn, Jennifer Foster, Mark Dras, and Lamia Tounsi. 2014. Cross-lingual transfer parsing for low-resourced languages: An Irish case study. In Proceedings of the First Celtic Language Technology Workshop, pages 41–49, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

Ryan McDonald, Kevin Lerman, and Fernando Pereira. 2006. Multilingual dependency analysis with a two-stage discriminative parser. In Proceedings of the Tenth Conference on Computational Natural Language Learning (CoNLL-X), pages 216–220, New York City. Association for Computational Linguistics.

Steven Moran, Daniel McCloy, and Richard Wright. 2012. Revisiting population size vs. phoneme inventory size. Language, 88(4):877–893.

Steven Moran, Daniel McCloy, and Richard Wright, editors. 2014. PHOIBLE Online. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Phoebe Mulcaire, Jungo Kasai, and Noah A. Smith. 2019. Polyglot contextual representations improve crosslingual transfer. 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 3912–3918, Minneapolis, Minnesota. Association for Computational Linguistics.

Max Müller-Eberstein, Rob Goot, and Barbara Plank. 2022. Probing for labeled dependency trees. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 7711–7726, Dublin, Ireland. Association for Computational Linguistics.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In Proceedings of the 12th Language Resources and Evaluation Conference, pages 4034–4043, Marseille, France. European Language Resources Association.

Michael Paul, Andrew Finch, and Eiichrio Sumita. 2013. How to choose the best pivot language for automatic translation of low-resource languages. ACM Transactions on Asian Language Information Processing (TALIP), 12(4):1–17.

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, Helen O'Horan, Yevgeni Berzak, Ivan Vulić, Roi Reichart, Thierry Poibeau, Ekaterina Shutova, and Anna Korhonen. 2019. Modeling language variation and universals: A survey on typological linguistics for natural language processing. Computational Linguistics, 45(3):559–601.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Sebastian Ruder, Noah Constant, Jan Botha, Aditya Siddhant, Orhan Firat, Jinlan Fu, Pengfei Liu, Junjie Hu, Dan Garrette, Graham Neubig, and Melvin Johnson. 2021. XTREME-R: Towards more challenging

and nuanced multilingual evaluation. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 10215–10245, Online and Punta Cana, Dominican Republic. 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.

Wolfgang Seeker and Jonas Kuhn. 2013. Morphological and syntactic case in statistical dependency parsing. Computational Linguistics, 39(1):23–55.

Kyle Shaffer. 2021. Language clustering for multilingual named entity recognition. In Findings of the Association for Computational Linguistics: EMNLP 2021, pages 40–45, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Kaius Sinnemäki. 2010. Word order in zero-marking languages. Studies in Language. International Journal sponsored by the Foundation "Foundations of Language", 34(4):869–912.

Aaron Smith, Bernd Bohnet, Miryam de Lhoneux, Joakim Nivre, Yan Shao, and Sara Stymne. 2018. 82 treebanks, 34 models: Universal Dependency parsing with multi-treebank models. In Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, pages 113–123, Brussels, Belgium. Association for Computational Linguistics.

Xu Tan, Jiale Chen, Di He, Yingce Xia, Tao Qin, and Tie-Yan Liu. 2019. Multilingual neural machine translation with language clustering. 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 963–973, Hong Kong, China. Association for Computational Linguistics.

Iulia Turc, Kenton Lee, Jacob Eisenstein, Ming-Wei Chang, and Kristina Toutanova. 2021. Revisiting the primacy of English in zero-shot cross-lingual transfer. arXiv preprint arXiv:2106.16171.

Fiona J Tweedie and R Harald Baayen. 1998. How variable may a constant be? measures of lexical richness in perspective. Computers and the Humanities, 32(5):323–352.

Rob van der Goot and Miryam de Lhoneux. 2021. Parsing with pretrained language models, multiple datasets, and dataset embeddings. In Proceedings of the 20th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2021), pages 96–104, Sofia, Bulgaria. Association for Computational Linguistics.

Rob van der Goot, Ahmet Üstün, Alan Ramponi, Ibrahim Sharaf, and Barbara Plank. 2021. Massive choice, ample tasks (MaChAmp): A toolkit for multi-task learning in NLP. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations, pages 176–197, Online. Association for Computational Linguistics.

Raúl Vázquez, Yves Scherrer, Sami Virpioja, and Jörg Tiedemann. 2021. The Helsinki submission to the AmericasNLP shared task. In Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas, pages 255–264, Online. Association for Computational Linguistics.

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, Eric W. Holman, Dik Bakker, and Cecil H. Brown. 2010. Evaluating linguistic distance measures. Physica A: Statistical Mechanics and its Applications, 389(17):3632–3639.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Daniel Zeman, Jan Hajič, Martin Popel, Martin Potthast, Milan Straka, Filip Ginter, Joakim Nivre, and Slav Petrov. 2018. CoNLL 2018 shared task: Multilingual parsing from raw text to Universal Dependencies. In Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, pages 1–21, Brussels, Belgium. Association for Computational Linguistics.

Daniel Zeman, Joakim Nivre, Mitchell Abrams, Elia Ackermann, Noëmi Aepli, Hamid Aghaei, Željko Agić, Amir Ahmadi, Lars Ahrenberg, Chika Kennedy Ajede, Gabrielė Aleksandravičiūtė, Ika Alfina, Lene Antonsen, Katya Aplonova, Angelina Aquino, Carolina Aragon, Maria Jesus Aranzabe, Bilge Nas Arıcan, Hórunn Arnardóttir, Gashaw Arutie, Jessica Naraiswari Arwidarasti, Masayuki Asahara, Deniz Baran Aslan, Luma Ateyah, Furkan Atmaca, Mohammed Attia, Aitziber Atutxa, Liesbeth Augustinus, Elena Badmaeva, Keerthana Balasubramani, Miguel Ballesteros, Esha Banerjee, Sebastian

Bank, Verginica Barbu Mititelu, Starkaður Barkarson, Rodolfo Basile, Victoria Basmov, Colin Batchelor, John Bauer, Seyyit Talha Bedir, Kepa Bengoetxea, Gözde Berk, Yevgeni Berzak, Irshad Ahmad Bhat, Riyaz Ahmad Bhat, Erica Biagetti, Eckhard Bick, Agnė Bielinskienė, Kristín Bjarnadóttir, Rogier Blokland, Victoria Bobicev, Loïc Boizou, Emanuel Borges Völker, Carl Börstell, Cristina Bosco, Gosse Bouma, Sam Bowman, Adriane Boyd, Anouck Braggaar, Kristina Brokaitė, Aljoscha Burchardt, Marie Candito, Bernard Caron, Gauthier Caron, Lauren Cassidy, Tatiana Cavalcanti, Gülşen Cebiroğlu Eryiğit, Flavio Massimiliano Cecchini, Giuseppe G. A. Celano, Slavomír Čéplö, Neslihan Cesur, Savas Cetin, Özlem Çetinoğlu, Fabricio Chalub, Shweta Chauhan, Ethan Chi, Taishi Chika, Yongseok Cho, Jinho Choi, Jayeol Chun, Juyeon Chung, Alessandra T. Cignarella, Silvie Cinková, Aurélie Collomb, Çağrı Çöltekin, Miriam Connor, Marine Courtin, Mihaela Cristescu, Philemon Daniel, Elizabeth Davidson, Marie-Catherine de Marneffe, Valeria de Paiva, Mehmet Oguz Derin, Elvis de Souza, Arantza Diaz de Ilarraza, Carly Dickerson, Arawinda Dinakaramani, Elisa Di Nuovo, Bamba Dione, Peter Dirix, Kaja Dobrovoljc, Timothy Dozat, Kira Droganova, Puneet Dwivedi, Hanne Eckhoff, Sandra Eiche, Marhaba Eli, Ali Elkahky, Binyam Ephrem, Olga Erina, Tomaž Erjavec, Aline Etienne, Wograine Evelyn, Sidney Facundes, Richárd Farkas, Jannatul Ferdaousi, Marília Fernanda, Hector Fernandez Alcalde, Jennifer Foster, Cláudia Freitas, Kazunori Fujita, Katarína Gajdošová, Daniel Galbraith, Marcos Garcia, Moa Gärdenfors, Sebastian Garza, Fabrício Ferraz Gerardi, Kim Gerdes, Filip Ginter, Gustavo Godoy, Iakes Goenaga, Koldo Gojenola, Memduh Gökırmak, Yoav Goldberg, Xavier Gómez Guinovart, Berta González Saavedra, Bernadeta Griciūtė, Matias Grioni, Loïc Grobol, Normunds Grūzītis, Bruno Guillaume, Céline Guillot-Barbance, Tunga Güngör, Nizar Habash, Hinrik Hafsteinsson, Jan Hajič, Jan Hajič jr., Mika Hämäläinen, Linh Hà Mỹ, Na-Rae Han, Muhammad Yudistira Hanifmuti, Sam Hardwick, Kim Harris, Dag Haug, Johannes Heinecke, Oliver Hellwig, Felix Hennig, Barbora Hladká, Jaroslava Hlaváčová, Florinel Hociung, Petter Hohle, Eva Huber, Jena Hwang, Takumi Ikeda, Anton Karl Ingason, Radu Ion, Elena Irimia, Ọlájídé Ishola, Kaoru Ito, Siratun Jannat, Tomáš Jelínek, Apoorva Jha, Anders Johannsen, Hildur Jónsdóttir, Fredrik Jørgensen, Markus Juutinen, Sarveswaran K, Hüner Kaşıkara, Andre Kaasen, Nadezhda Kabaeva, Sylvain Kahane, Hiroshi Kanayama, Jenna Kanerva, Neslihan Kara, Boris Katz, Tolga Kayadelen, Jessica Kenney, Václava Kettnerová, Jesse Kirchner, Elena Klementieva, Elena Klyachko, Arne Köhn, Abdullatif Köksal, Kamil Kopacewicz, Timo Korkiakangas, Mehmet Köse, Natalia Kotsyba, Jolanta Kovalevskaitė, Simon Krek, Parameswari Krishnamurthy, Sandra Kübler, Oğuzhan Kuyrukçu, Aslı Kuzgun, Sookyoung Kwak, Veronika Laippala, Lucia Lam, Lorenzo Lambertino, Tatiana Lando, Septina Dian Larasati, Alexei Lavrentiev, John Lee, Phuong Lê Hồng, Alessandro Lenci, Saran Lertpradit, Herman Leung, Maria Levina, Cheuk Ying Li, Josie Li, Keying Li, Yuan Li, KyungTae Lim, Bruna Lima Padovani, Krister Lindén, Nikola Ljubešić, Olga Loginova, Stefano Lusito, Andry Luthfi, Mikko Luukko, Olga Lyashevskaya, Teresa Lynn, Vivien Macketanz, Menel Mahamdi, Jean Maillard, Aibek Makazhanov, Michael Mandl, Christopher Manning, Ruli Manurung, Büşra Marşan, Cătălina Mărănduc, David Mareček, Katrin Marheinecke, Héctor Martínez Alonso, Lorena Martín-Rodríguez, André Martins, Jan Mašek, Hiroshi Matsuda, Yuji Matsumoto, Alessandro Mazzei, Ryan McDonald, Sarah McGuinness, Gustavo Mendonça, Tatiana Merzhevich, Niko Miekka, Karina Mischenkova, Margarita Misirpashayeva, Anna Missilä, Cătălin Mititelu, Maria Mitrofan, Yusuke Miyao, AmirHossein Mojiri Foroushani, Judit Molnár, Amirsaeid Moloodi, Simonetta Montemagni, Amir More, Laura Moreno Romero, Giovanni Moretti, Keiko Sophie Mori, Shinsuke Mori, Tomohiko Morioka, Shigeki Moro, Bjartur Mortensen, Bohdan Moskalevskyi, Kadri Muischnek, Robert Munro, Yugo Murawaki, Kaili Müürisep, Pinkey Nainwani, Mariam Nakhlé, Juan Ignacio Navarro Horñiacek, Anna Nedoluzhko, Gunta Nešpore-Bērzkalne, Manuela Nevaci, Luong Nguyễn Thị, Huyền Nguyễn Thị Minh, Yoshihiro Nikaido, Vitaly Nikolaev, Rattima Nitisaroj, Alireza Nourian, Hanna Nurmi, Stina Ojala, Atul Kr. Ojha, Adèdayọ̀ Olúòkun, Mai Omura, Emeka Onwuegbuzia, Petya Osenova, Robert Östling, Lilja Øvrelid, Şaziye Betül Özateş, Merve Özçelik, Arzucan Özgür, Balkız Öztürk Başaran, Hyunji Hayley Park, Niko Partanen, Elena Pascual, Marco Passarotti, Agnieszka Patejuk, Guilherme Paulino-Passos, Angelika Peljak-Łapińska, Siyao Peng, Cenel-Augusto Perez, Natalia Perkova, Guy Perrier, Slav Petrov, Daria Petrova, Jason Phelan, Jussi Piitulainen, Tommi A Pirinen, Emily Pitler, Barbara Plank, Thierry Poibeau, Larisa Ponomareva, Martin Popel, Lauma Pretkalniņa, Sophie Prévost, Prokopis Prokopidis, Adam Przepiórkowski, Tiina Puolakainen, Sampo Pyysalo, Peng Qi, Andriela Rääbis, Alexandre Rademaker, Mizanur Rahoman, Taraka Rama, Loganathan Ramasamy, Carlos Ramisch, Fam Rashel, Mohammad Sadegh Rasooli, Vinit Ravishankar, Livy Real, Petru Rebeja, Siva Reddy, Mathilde Regnault, Georg Rehm, Ivan Riabov, Michael Rießler, Erika Rimkutė, Larissa Rinaldi, Laura Rituma, Putri Rizqiyah, Luisa Rocha, Eiríkur Rögnvaldsson, Mykhailo Romanenko, Rudolf Rosa, Valentin Roșca, Davide Rovati, Olga Rudina, Jack Rueter, Kristján Rúnarsson, Shoval Sadde, Pegah Safari, Benoît Sagot, Aleksi Sahala, Shadi Saleh, Alessio Salomoni, Tanja Samardžić, Stephanie Samson, Manuela Sanguinetti, Ezgi Sanıyar, Dage Särg, Baiba Saulīte, Yanin Sawanakunanon, Shefali Saxena, Kevin Scannell, Salvatore Scarlata, Nathan Schneider, Sebastian Schuster, Lane Schwartz, Djamé Seddah, Wolfgang Seeker, Mojgan Seraji, Syeda Shahzadi, Mo Shen, Atsuko Shimada, Hiroyuki Shirasu, Yana Shishkina, Muh Shohibussirri, Dmitry Sichinava, Janine Siewert, Einar Freyr Sig-

urðsson, Aline Silveira, Natalia Silveira, Maria Simi, Radu Simionescu, Katalin Simkó, Mária Šimková, Kiril Simov, Maria Skachedubova, Aaron Smith, Isabela Soares-Bastos, Shafi Sourov, Carolyn Spadine, Rachele Sprugnoli, Steinþór Steingrímsson, Antonio Stella, Milan Straka, Emmett Strickland, Jana Strnadová, Alane Suhr, Yogi Lesmana Sulestio, Umut Sulubacak, Shingo Suzuki, Zsolt Szántó, Chihiro Taguchi, Dima Taji, Yuta Takahashi, Fabio Tamburini, Mary Ann C. Tan, Takaaki Tanaka, Dipta Tanaya, Samson Tella, Isabelle Tellier, Marinella Testori, Guillaume Thomas, Liisi Torga, Marsida Toska, Trond Trosterud, Anna Trukhina, Reut Tsarfaty, Utku Türk, Francis Tyers, Sumire Uematsu, Roman Untilov, Zdeňka Urešová, Larraitz Uria, Hans Uszkoreit, Andrius Utka, Sowmya Vajjala, Rob van der Goot, Martine Vanhove, Daniel van Niekerk, Gertjan van Noord, Viktor Varga, Eric Villemonte de la Clergerie, Veronika Vincze, Natalia Vlasova, Aya Wakasa, Joel C. Wallenberg, Lars Wallin, Abigail Walsh, Jing Xian Wang, Jonathan North Washington, Maximilan Wendt, Paul Widmer, Sri Hartati Wijono, Seyi Williams, Mats Wirén, Christian Wittern, Tsegay Woldemariam, Tak-sum Wong, Alina Wróblewska, Mary Yako, Kayo Yamashita, Naoki Yamazaki, Chunxiao Yan, Koichi Yasuoka, Marat M. Yavrumyan, Arife Betül Yenice, Olcay Taner Yıldız, Zhuoran Yu, Arlisa Yuliawati, Zdeněk Žabokrtský, Shorouq Zahra, Amir Zeldes, He Zhou, Hanzhi Zhu, Anna Zhuravleva, and Rayan Ziane. 2021. Universal dependencies 2.9. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Daniel Zeman and Philip Resnik. 2008. Cross-language parser adaptation between related languages. In Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages.

A Missing Syntax Features in URIEL

ISO-3 Lang code	Count feats no value	ISO-3 Lang code	Count feats no value
afr	66	lat	70
grc	70	lav	24
arb	37	lit	29
eus	18	mlt	97
bel	70	mar	37
bul	17	pcm	103
cat	47	nno	103
zho	1	chu	103
cop	36	fro	70
hrv	91	pes	18
ces	30	pol	29
dan	29	por	26
nld	32	ron	29
eng	0	rus	13
est	23	gla	28
fao	70	srp	68
fin	5	slk	103
fra	6	slv	43
glg	70	spa	2
deu	16	swe	23
got	71	swl	103
ell	17	tam	33
heb	16	tel	42
hin	21	tur	14
hun	12	ukr	26
isl	28	urd	40
ind	3	uig	45
gle	27	vie	10
ita	26	cym	22
jpn	13	hye	26
kor	15	wol	22

Table 4: The counts of missing syntactic features in URIEL for languages included in UD. The table contains some languages that were not included in our experiments (due to sampling), but are listed as available in UD.

B Data Statistics

Treebank	MSL	MWL
UD_Afrikaans-AfriBooms	25.76	4.98
UD_Ancient_Greek-PROIEL	12.46	5.06
UD_Ancient_Greek-Perseus	13.93	4.59
UD_Arabic-PADT	31.58	4.52
UD_Armenian-ArmTDP	21.18	5.0
UD_Basque-BDT	13.52	5.6
UD_Belarusian-HSE	11.95	5.31
UD_Bulgarian-BTB	13.96	4.63
UD_Catalan-AnCora	31.75	4.29
UD_Chinese-GSD	24.67	1.58
UD_Chinese-GSDSimp	24.67	1.58
UD_Classical_Chinese-Kyoto	4.84	1.04
UD_Coptic-Scriptorium	11.89	5.4
UD_Croatian-SET	22.11	5.0
UD_Czech-CAC	20.09	5.06
UD_Czech-CLTT	32.27	5.45
UD_Czech-FicTree	13.1	4.01
UD_Czech-PDT	17.1	4.84
UD_Danish-DDT	18.34	4.41
UD_Dutch-Alpino	15.14	4.7
UD_Dutch-LassySmall	12.98	4.83
UD_English-Atis	11.38	4.71
UD_English-ESL	19.04	3.87
UD_English-EWT	16.1	4.11
UD_English-GUM	18.05	4.18
UD_English-GUMReddit	18.45	3.94
UD_English-LinES	18.06	3.98
UD_English-ParTUT	24.41	4.53
UD_English-Tweebank2	15.1	4.08
UD_Estonian-EDT	13.99	5.55
UD_Estonian-EWT	12.05	4.7
UD_Faroese-FarPaHC	22.64	3.58
UD_Finnish-FTB	8.5	5.95
UD_Finnish-TDT	13.31	6.49
UD_French-FTB	29.96	4.33
UD_French-GSD	23.86	4.41
UD_French-ParTUT	29.04	4.64
UD_French-Rhapsodie	14.67	3.5
UD_French-Sequoia	22.03	4.57
UD_Galician-CTG	31.66	4.86
UD_German-GSD	18.76	5.27
UD_German-HDT	17.99	5.67
UD_German-tweeDE	9.25	4.74
UD_Gothic-PROIEL	10.34	5.21
UD_Greek-GDT	24.8	5.11
UD_Hebrew-HTB	18.76	4.03
UD_Hindi-HDTB	21.13	3.83
UD_Hindi_English-HIENCS	13.95	3.75
UD_Hungarian-Szeged	22.16	5.46
UD_Icelandic-IcePaHC	20.72	4.04
UD_Icelandic-Modern	23.04	4.43
UD_Indonesian-GSD	21.39	5.25
UD_Irish-IDT	23.94	4.52
UD_Italian-ISDT	19.63	4.65
UD_Italian-ParTUT	25.53	4.93
UD_Italian-PoSTWITA	17.77	4.72
UD_Italian-TWITTIRO	19.91	4.56
UD_Italian-VIT	25.22	4.75
UD_Japanese-GSD	23.88	1.65
UD_Japanese-GSDLUW	18.48	2.13
UD_Korean-GSD	12.88	2.84
UD_Korean-Kaist	12.88	2.84
UD_Latin-ITTB	17.16	5.06
UD_Latin-LLCT	26.64	4.91
UD_Latin-PROIEL	10.81	5.38
UD_Latin-UDante	32.76	4.9
UD_Latvian-LVTB	16.92	5.1
UD_Lithuanian-ALKSNIS	20.35	5.58
UD_Lithuanian-HSE	20.98	5.1
UD_Maltese-MUDT	20.37	4.56
UD_Marathi-UFAL	7.32	4.03
UD_Naija-NSC	15.37	2.97
UD_Norwegian-Bokmaal	15.54	4.47
UD_Norwegian-Nynorsk	17.31	4.51
UD_Norwegian-NynorskLIA	10.32	3.15
UD_Old_Church_Slavonic-PROIEL	9.08	4.5
UD_Old_East_Slavic-TOROT	8.9	4.5
UD_Old_French-SRCMF	11.21	3.5
UD_Persian-PerDT	17.01	3.82
UD_Persian-Seraji	25.0	3.78
UD_Polish-LFG	7.6	4.64
UD_Polish-PDB	15.78	5.07
UD_Portuguese-Bosque	22.65	4.42
UD_Portuguese-GSD	24.75	4.34
UD_Romanian-Nonstandard	22.09	3.77
UD_Romanian-RRT	23.02	4.69
UD_Romanian-SiMoNERo	31.19	5.19
UD_Russian-GSD	19.46	5.28
UD_Russian-SynTagRus	17.3	5.04
UD_Russian-Taiga	11.01	4.58
UD_Scottish_Gaelic-ARCOSG	19.02	4.2
UD_Serbian-SET	22.31	4.93
UD_Slovak-SNK	9.5	4.41
UD_Slovenian-SSJ	17.37	4.63
UD_Spanish-AnCora	30.98	4.43
UD_Spanish-GSD	26.44	4.41
UD_Swedish-LinES	17.46	4.46
UD_Swedish-Talbanken	15.49	4.98
UD_Swedish_Sign_Language-SSLC	7.4	8.91
UD_Tamil-TTB	14.34	7.21
UD_Telugu-MTG	4.84	4.66
UD_Turkish-Atis	8.47	6.65
UD_Turkish-BOUN	12.46	5.51
UD_Turkish-FrameNet	7.14	5.36
UD_Turkish-IMST	10.05	5.41
UD_Turkish-Kenet	9.31	5.41
UD_Turkish-Penn	11.21	5.61
UD_Turkish-Tourism	4.64	5.03
UD_Turkish_German-SAGT	17.31	4.53
UD_Ukrainian-IU	16.8	4.64
UD_Urdu-UDTB	26.88	3.57
UD_Uyghur-UDT	11.63	5.48
UD_Vietnamese-VTB	14.49	3.99
UD_Welsh-CCG	19.73	4.06
UD_Western_Armenian-ArmTDP	18.13	5.06
UD_Wolof-WTB	19.21	3.46

Table 5: Mean Sentence Length (MSL) and Mean Word Length (MWL) values per treebank.