Language Embeddings Sometimes Contain Typological Generalizations

Robert Östling
Stockholm University
Department of Linguistics
robert@ling.su.se

Murathan Kurfalı*
Stockholm University
Department of Psychology
murathan.kurfali@su.se

To what extent can neural network models learn generalizations about language structure, and how do we find out what they have learned? We explore these questions by training neural models for a range of natural language processing tasks on a massively multilingual dataset of Bible translations in 1,295 languages. The learned language representations are then compared to existing typological databases as well as to a novel set of quantitative syntactic and morphological features obtained through annotation projection. We conclude that some generalizations are surprisingly close to traditional features from linguistic typology, but that most of our models, as well as those of previous work, do not appear to have made linguistically meaningful generalizations. Careful attention to details in the evaluation turns out to be essential to avoid false positives. Furthermore, to encourage continued work in this field, we release several resources covering most or all of the languages in our data: (1) multiple sets of language representations, (2) multilingual word embeddings, (3) projected and predicted syntactic and morphological features, (4) software to provide linguistically sound evaluations of language representations.

1. Introduction and Related Work

In highly multilingual natural language processing (NLP) systems covering hundreds or even thousands of languages, one must deal with a considerable portion of the total diversity present in the world's approximately 7,000 languages. This goes far beyond the standard training resources for most tasks, even after the advent of highly multilingual resources such as the Universal Dependencies Treebanks (McDonald et al.

* Work carried out while at the Department of Linguistics.

Action Editor: Rico Sennrich. Submission received: 10 January 2023; revised version received: 12 May 2023; accepted for publication: 2 June 2023.

https://doi.org/10.1162/coli_a_00491

2013, 114 languages) and UniMorph (Sylak-Glassman et al. 2015, 110 languages) and unsupervised representation learning models such as multilingual BERT (Devlin et al. 2019, 104 languages), mT5 (Xue et al. 2021, 101 languages), and XLM-R (Conneau et al. 2020, 100 languages). The discrepancy between the total set of languages and available NLP resources is even greater if one considers the *diversity* of languages. To take one representative example, the mT5 model of Xue et al. (2021) contains 101 languages from 16 families, of which only 3 are mainly spoken outside Eurasia. This is to be compared to the 424 spoken natural language families in the Glottolog database of languages (Hammarström et al. 2022), of which the vast majority are predominantly spoken outside Eurasia. Reducing this discrepancy is one of the most challenging and important projects of our field.

1.1 Highly Multilingual Natural Language Processing

From the beginning of the statistical NLP era, researchers have attempted to bridge the gap between the few languages with digital resources, and those without. Early attempts were based on annotation projection (e.g., Yarowsky and Ngai 2001; Hwa et al. 2005), followed by different types of transfer learning using, for instance, multilingual word embeddings (a comprehensive overview can be found in Søgaard et al. 2019), multilingual neural models (e.g., Ammar et al. 2016), and more recently, multilingual pre-trained language models that achieved a breakthrough with the multilingual BERT model (Devlin et al. 2019).

Even after large language models such as GPT-3 (Brown et al. 2020) and PaLM (Chowdhery et al. 2022) have demonstrated human-like performance in an increasing number of natural language problems for some languages, similar models still have severe problems with handling data in low-resource, predominantly non-Eurasian, languages (Blasi, Anastasopoulos, and Neubig 2022). As an illustrative example, we consider the evaluation of Ebrahimi et al. (2022), who developed and applied an XLM-R (Conneau et al. 2020) based Natural Language Inference (NLI) model on a newly created dataset for 10 languages from the Americas, ranging from large languages such as Quechua and Guaraní with millions of speakers, to languages such as Shipibo-Konibo (Panoan, 35,000 speakers) that are closer to the median of the distribution of speaker counts among the world's languages. Their best system manages to achieve a mean accuracy of 49% (range over languages: 41%–62%), as compared to a 33% random baseline and 85% in English.

Our focus in this work is not to directly improve the state of the art in any specific NLP application, but rather to explore the properties of massively multilingual (1,000+ languages) neural models.

1.2 Linguistic Typology and Natural Language Processing

As the number of languages one wishes to study grows, it becomes increasingly important to consider the systematic differences and similarities between languages. This is the object of study in the field of linguistic typology (e.g., Croft 2002; Shopen 2007), and a growing body of research aims to transfer insights back and forth between linguistic typology and natural language processing. Below, we discuss three main directions: predicting typological features, applying typological information to improve NLP, and applying NLP to obtain typological information.

1.2.1 Extraction of Typological Features from Diverse Sources of Data. Traditionally, typological databases have been constructed manually,[1] where collaborations of linguistics researchers classify languages according to a predetermined set of typological parameters. This is a slow and costly process, and often leaves large gaps where language documentation is missing or incomplete, or where there has been insufficient researcher time or interest to perform the analysis for some languages.

Researchers have long attempted to collect or infer the values of typological features from existing data sources. In particular, parallel text has been an important resource, and has been applied to multiple domains of linguistic typology: tense markers (Dahl 2007; Asgari and Schütze 2017), semantic categories of motion verbs (Wälchli and Cysouw 2012), word order (Östling 2015), colexification patterns (Östling 2016), and affixation (Hammarström 2021). All these methods rely on some type of word or morpheme alignment (for an overview, see, e.g., Tiedemann 2011) combined with manually specified, feature-specific rules. For instance, given a parsed text and word alignments to other languages, one can construct rules for estimating word order properties in those languages. Apart from being a time-consuming process, this also means one has to know beforehand which features to look for before hand-crafting rules to estimate their values. In contrast, our interest in this work is to investigate models that are not "aware" of the existence of specific typological parameters, but rather has to discover them from data.

A different but somewhat similar method is text mining of grammatical descriptions. Hammarström (2021) uses such an approach to investigate affixing tendencies in 4,287 languages with digital or digitized grammars, which makes this one of very few studies using a majority of the world's languages as its sample. He searches for terms describing suffixing or prefixing in the language of the description, and reports an overall binary agreement of 75% over the 917 language subset that is also present in the manually collected database of Dryer (2013f).

1.2.2 Typological Feature Prediction. This line of work is looking at finding methods for predicting which features we would expect a given language to have, given its known relatives, location in the world, typological properties of genealogically and/or geographically close languages, as well as other (known) features of the same language.

Murawaki (2019) designed a Bayesian model to infer latent representations that explain observable typological features. Based on universal tendencies, for instance, that noun–adjective order tends to imply noun–relative clause order, combined with spatial and phylogenetic information, this model is able to accurately predict typological feature values outside the training data.

Bjerva et al. (2019) designed a probabilistic model that learns a set of latent language representations along with a mapping to binary typological feature vectors. They apply this to fill gaps in a language–feature matrix given information from related languages and/or other features in the same language, and find that this can be done with near-perfect accuracy as long as a sufficient amount of information is provided.

1 Although some databases contain automatically computed feature values that can be logically derived from manually specified parameters, AUTOTYP (Bickel et al. 2017) being a prime example of this, we count these as manually constructed since all analysis of language data is performed by humans.

The shared task on predicting typological features (Bjerva et al. 2020) brought together a number of different approaches of the same general direction. The best-performing system (Vastl, Zeman, and Rosa 2020) combines a simple method based on directly exploiting correlations between features, with a neural network predicting missing feature values from known values as well as spatial information.

Our interest is in a sense opposite to this line of work, since we are interested in how typological generalizations can be made from language data alone. Rather than exploiting genealogical and geographical information along with correlations between typological features, we treat them as confounding variables to be controlled for. After all, the most interesting aspects of languages are not the ones that are identical to their neighboring relatives, but rather those that are in some aspect unique or divergent. We do, however, note that if one, for whatever reason, is obliged to provide a "best guess" for the value of a certain feature without actually studying any language data, the methods described above are useful. It would also be possible to combine this with our line of research, in order to compare the expected to actual feature values, and hopefully identify those interesting cases.

1.2.3 Application of Typological Information in NLP. As multilingual NLP became more established, several authors attempted to use existing typological information to inform models for the purpose of improving their accuracy (overviews can be found in O'Horan et al. 2016; Ponti et al. 2019).

Such research predates the neural era and modern representation learning. For instance, Naseem, Barzilay, and Globerson (2012) used a feature-rich model with sparse indicator features depending on a number of typological parameters from the World Atlas of Language Structures (WALS) database (Dryer and Haspelmath 2013), which results in parameter sharing between structurally similar languages. They observed substantial gains in parsing accuracy using this approach.

In their multilingual neural parser covering seven related European languages, Ammar et al. (2016) attempted to insert typological features from the WALS database. However, they saw less benefit from this than learning language embeddings by only providing a language identifier for each example. This work was an early demonstration of the usefulness of learning language representations. While Ammar et al. (2016) showed that language embeddings improve parsing accuracy, they did not investigate whether the language embeddings learned to *generalize* over languages rather than simply using the embeddings to *identify* each language. A generalization relevant to a parser would be, for instance, whether adjectives tend to precede or to follow the noun that they modify. To say that the model has made this generalization, we would need evidence that the information on adjective/noun order is somehow consistently encoded in the embedding of each language. In contrast, even random language embeddings could serve to identify each language, so that for instance the (random) German embedding is simply used to activate some opaquely coded German syntactic model in the parser's neural network.

Bjerva and Augenstein (2021) hypothesize that one reason for the modest performance improvement when adding typological features to neural multilingual NLP systems could be that the systems are already capable of making the necessary generalizations. They show that blinding such models to typological information, by adding a gradient reversal layer that ensures the representations predict typological features *poorly*, yields somewhat reduced performance across a range of NLP tasks.

A different direction was taken by Wang, Ruder, and Neubig (2022), who use artifical data generated from lexical resources to adapt pretrained multilingual language

models to languages with only a lexicon available. This allows some coverage even for the thousands of languages with no other digital resources than a lexicon. While potentially useful for some NLP applications, we do not consider this line of research further since it is not useful for studying structural features of languages.

1.2.4 Extracting Typological Information from Neural NLP Models. Another line of research focuses on identifying how multilingual language models encode structural properties across languages, and in particular to what extent those representations are common across typologically similar languages.

Östling and Tiedemann (2017) trained a character-level LSTM language model on translated Bible text from 990 different languages, also conditioned on language embeddings, and showed that the resulting embeddings can be used to reconstruct language genealogies. Similar results have later been obtained using a variety of multilingual neural machine translation (NMT) models that learn either language embeddings (Tiedemann 2018; Platanios et al. 2018), or language representations derived from encoder activations (Kudugunta et al. 2019), or both (Malaviya, Neubig, and Littell 2017). In the following we will use the term **language representations** to refer to vector representations of languages, regardless of how they were estimated.

The property that similar languages (i.e., languages that share many properties) have similar vector representations is a basic requirement of any useful language representation whose purpose is to generalize across languages. It is generally sufficient to improve practical NLP models, especially when evaluated on datasets with many similar languages. However, an aggregate measure of language similarity contains relatively little information. With only similarity information it is impossible to capture the *differences* between otherwise similar languages, for instance, when one language differs from its close relatives in some aspect. Even worse, we have no way to learn properties of languages that are not similar to other languages in our data, for instance, isolates such as Basque or Burushaski.

Starting with Malaviya, Neubig, and Littell (2017), several authors have attempted to probe directly whether the language representations learned by neural models encode the same types of generalizations across languages that have been studied in the field of linguistic typology. Malaviya, Neubig, and Littell (2017) used logistic regression classifiers to probe whether typological features can be predicted from language representations derived from a multilingual NMT system trained on Bible translations in 1,017 languages. They used features from the URIEL database (Littell et al. 2017), which contains typological data sourced from Dryer and Haspelmath (2013), Moran and McCloy (2019), and Eberhard, Simons, and Fennig (2019). Based on their classification experiments, they conclude that their language representations have generalized in several domains of language, from phonological to syntactic features. This finding was later supported by Oncevay, Haddow, and Birch (2020), who compared the original representations of Malaviya, Neubig, and Littell (2017) with a novel set of representations that combined Malaviya's with URIEL features using canonical correlation analysis (CCA).

Similar results have been reported by Bjerva and Augenstein (2018a), who use the language embeddings from Östling and Tiedemann (2017) and fine-tune them using specific NLP tasks of several types: grapheme-to-phoneme conversion (representing phonology), word inflection (morphology), and part-of-speech tagging (syntax). Using a k-nearest-neighbors (kNN) classifier for probing, they conclude that typological features from all three domains of language that were investigated (phonology, morphology, syntax) are present in the language representations.

Another, smaller-scale, study on the same topic is that of He and Sagae (2019). They use a denoising autoencoder to reconstruct sentences in 27 languages, using a multilingual dictionary so that the model is presented only with English vocabulary. Based on a syntactic feature classification task, they report that properties of verbal categories, word order, nominal categories, morphology, and lexicon were encoded in the language embeddings learned by their autoencoder. They did not see any difference from baseline classification accuracy for features relating to phonology and nominal syntax, a fact that they ascribe to the small number of languages available for their evaluation.

In their study of multilingual semantic drift, Beinborn and Choenni (2020) demonstrate that similarities in the multilingual sentence representations of Artetxe and Schwenk (2019) as well as word representations from Conneau et al. (2018) follow language phylogenies when applied to multilingual concept lists. Similarly, Rama, Beinborn, and Eger (2020) extract the internal representations of multilingual BERT (Devlin et al. 2019) for each of 100 languages, obtained by encoding a single word at a time from multilingual concept lists. As expected, given the lack of input beyond the word level, they find a low correlation between the language representations obtained and structural properties of the languages studied, although the language representations correlate well with language phylogeny. Such methods could potentially be useful within lexical typology, which studies the systematic variation between languages in the structure of their semantic spaces. A more direct approach in this direction was taken by Östling (2016), who projected multilingual concept lists through parallel texts for direct study of lexical **colexification**, semantically distinct concepts referred to by the same word form.

Chi, Hewitt, and Manning (2020) use the structural probing technique of Hewitt and Manning (2019) to find a syntactic subspace in multilingual BERT encodings of different languages, which allows a direct look at how the model encodes syntactic relations rather similarly across languages. Similarly, Stanczak et al. (2022) investigate morphosyntactic properties encoded by the multilingual BERT and XLM-R models, by analyzing their encodings of data from the Universal Dependency treebanks (Nivre et al. 2018). They find that the same morphosyntectic categories are encoded by sets of neurons that overlap to a significantly above-chance degree. While these studies suggest that pretrained multilingual language models to some degree make typological abstractions, they are limited to a relatively small and biased sample of languages.

Recently, Choenni and Shutova (2022) performed an in-depth investigation of how typological information is stored in the LASER Artetxe and Schwenk (2019), multilingual BERT (Devlin et al. 2019), XLM (Conneau and Lample 2019), and XLM-R models Conneau et al. (2020). They use seven pairs of (pairwisely) closely related languages from four different branches of the Indo-European family, and train simple two-layer perceptron classifiers from encoded sentences to typological feature values. These classifiers obtain high F_1 classification scores on a sentence level, although at a language level many features are consistently predicted incorrectly (Choenni and Shutova 2022, Figure 1). Given the small and highly biased set of languages, it is difficult to draw solid conclusions about how well structural properties are encoded in general across the languages of the world.

In summary, the results of previous work (Malaviya, Neubig, and Littell 2017; Bjerva and Augenstein 2018a; Oncevay, Haddow, and Birch 2020; He and Sagae 2019; Stanczak et al. 2022; Choenni and Shutova 2022) indicate that a range of neural models can learn language representations, which in most cases capture a range of generalizations in multiple domains of language.

A potential problem with the studies listed above is that the method for probing whether a certain feature is captured by some language representations varies, and in several cases is vulnerable to false positives due to the high correlation between features in similar languages. For instance, suppose that a classifier correctly predicts from the Dutch language representation that it tends to use adjective–noun order. Is this because the order of adjective and noun is coded in the language representation space, or because the language representations indicate that Dutch (in the test set) is lexically similar to German (in the training set), which also uses adjective–noun order? Some authors do not control for this correlation between typological features and genealogical or areal connections between languages at all (He and Sagae 2019) or only partially (Oncevay, Haddow, and Birch 2020); others provide the baseline classifiers with genealogical and geographic information (Malaviya, Neubig, and Littell 2017). Bjerva and Augenstein (2018a) hold out the largest single language family for each feature as a test set.[2] No attempt was made to control for correlations due to language contact. We also note that only a limited set of neural models have been explored in previous work, with most studies relying on the language representations from Östling and Tiedemann (2017) or Malaviya, Neubig, and Littell (2017).

2. Research Questions and Contributions

Given the inconclusive nature of previous work in the area, we here set out to systematically explore our overarching research question of which typological features can be discovered using neural models trained on a massively parallel corpus. Specifically, we set out to investigate the following research questions:

(RQ1) When given a low-dimensional language embedding space and no prior information on typological features, which types of features are discovered, and by which types of models?

(RQ2) What level of accuracy can be achieved for typological feature prediction *without* using typological information about geographically or genealogically close languages?

We note here that *discover* refers to the model identifying typological generalizations, by using a part of its allotted language embedding space to store the values of typological features. In order to evaluate how well this works, we limit ourselves to previously known typological features with data available for evaluation. For future work, it would be very interesting to more fully map the part of the embedding space that is *not* used to encode the subset of features we are exploring.

The main contributions of this work are listed below.[3]

1. A thorough investigation of a number of language representations from previous work as well as newly designed models, including a novel

2 Some details of the evaluation are unclear in the original paper; our summary in this work is also based on personal communication with the authors.

3 The code required to reproduce the results in this article is available at `https://github.com/robertostling/parallel-text-typology` and data is available at Zenodo (Östling and Kurfalı 2023).

word-level language model that can be trained efficiently on the full vocabularies of thousands of languages.

2. Publicly available resources derived from parallel texts, for 1,295 languages: language representations, multilingual word embeddings, partial inflectional paradigms, and projected token-level typological features relating to word order and affixation type. This data is also applicable to research beyond computational linguistics, such as in linguistic typology and language evolution.

3. A method and publicly available software for detecting typological features encoded into language representations.

The rest of this article is structured as follows.

First, we describe our evaluation framework in Section 3. In brief, we follow previous work in training classifiers to predict typological features from language representations. To avoid general language similarity from affecting the results, we use a cross-validation scheme that ensures languages in the test fold are not related to, geographically close to, or in a potential contact situation with any of the languages in the training fold. We also provide baselines from lexically derived language representations that are guaranteed *not* to directly code generalizations about language structure.

Second, we describe a diverse set of multilingual neural NLP models that we have implemented (Section 6), based on data derived in various ways from a massively parallel corpus of Bible translations (Section 5). All models use language embeddings. Because different tasks require analysis at different levels of language, and given the results of Bjerva and Augenstein (2018b), we expect that the language embedding spaces will mainly capture properties relevant to the task at hand.

Finally, we apply our evaluation framework to both our language embeddings, and to several sets of embeddings from previous work (Section 7). Surprisingly, we failed to detect any signal of linguistic generalizations in the representations from several previous studies, as well as in most of our own models (except the WORDLM and REINFLECT models). We demonstrate multiple ways in which spurious results can be obtained. For some of our models we show that typological features can be predicted with high accuracy, indicating that while neural models *can* discover typological generalizations, they do so less readily than suggested by previous research.

3. Evaluation Framework

3.1 Languages and Doculects

In this work, we generally use two levels of granularity, with the following terminology: **languages**, for our purposes identified by a unique ISO 639-3 code, and **doculects**, which is a particular language variety documented in a grammar, dictionary, or text (Cysouw and Good 2013). A typical situation encountered is that for a single language, say Kirghiz (ISO 639-3 code: kir), there are multiple Bible translations and multiple reference grammars. We count this as one language with multiple doculects, which may differ with respect to some features. Here we use the term *doculect* to emphasize that there may be multiple items (e.g., Bible translations, word embeddings, language representations) sorting under the same (ISO 639-3) language.

3.2 Linguistically Sound Cross-validation

The basis of our evaluation framework consists of typological feature classification, using constrained leave-one-out cross-validation. Thus, for each feature that we evaluate, the predicted label of a given doculect is obtained from a model that was trained on data from only **independent doculects**. We consider a potential training fold doculect to be independent of the test fold doculect if none of the following criteria apply:

1. Same family: The training doculect shares top-level family in Glottolog (Hammarström et al. 2022), including as a special case when they belong to the same language.
2. Same macro-area: The test and training doculects belong to the same linguistic macro-area. Although several definitions of macro-areas exist with some differences between them (Hammarström and Donohue 2014), we rely on the division found in Glottolog (Hammarström et al. 2022).
3. Potential long-distance contact: The training and test doculects are listed as potential contact languages in the phoneme inventory borrowings dataset of Grossman et al. (2020).

The first two criteria cover genealogical and areal correlations, respectively. The third criterion covers some cases that are not directly captured by the previous heuristics, in particular, languages such as English and Arabic that are influential globally.

For classification, we follow Malaviya, Neubig, and Littell (2017) in using L2-regularized logistic regression,[4] as implemented in Pedregosa et al. (2011). The language representations are used directly as features. Our classification models thus contain $k + 1$ parameters, for k-dimensional language representations and a bias term.

Naively applying this cross-validation scheme could still lead to problems due to correlations between the representations of related languages, for reasons like lexical similarity. If two large language families (A and B) that share a certain typological parameter P by chance have representations that are similar in some way, and a classifier for a language in family A is trained using (among others) the many languages in B, it will likely predict the parameter P with high accuracy. The effect of this will be demonstrated empirically in Section 8.3. Because the relationships between languages would be complex to model explicitly, we use family-wise Monte Carlo sampling to estimate classification accuracy and its uncertainty. To compute one sample of the classification accuracy of a given parameter, we uniformly sample one language from each family as the test set. For each language in the test set, we then uniformly sample one language from each family, but only among the *independent* languages (as defined above) to form the corresponding training fold. This procedure is repeated 201 times, yielding 201 samples of the classification accuracy for the given parameter and language representations. For each classification accuracy sample a_c, we also collect a paired

4 We use a fixed regularization strength $C = 10^{-3}$, and all features (i.e., individual language representation dimensions) are scaled to zero mean and unit variance. The use of strong regularization encodes the prior belief that most language representation dimensions are not predictors of a given typological feature. Insufficient regularization in preliminary experiments resulted in strong chance effects from minority class data points.

baseline accuracy sample a_b by randomly shuffling the training labels. This allows us to verify that the baseline behaves like a binomial distribution with $p = 0.5$.

3.3 Overview of the Method

We now provide a high-level description of how our language representations are created. This procedure will be described in more detailed over the following sections.

1. We start from the highly multilingual Bible corpus of Mayer and Cysouw (2014), and process it in several ways:
    (a) Remove verses present in few translations, and translations with few verses. The goal of this step is to approximate an ideal multi-parallel text where each verse is present in every translation.
    (b) Remove translations without word-level segmentation.
    (c) For languages where high-quality parsers are available, pick one translation (typically the most modern) and lemmatize, PoS-tag, and parse it.
2. Perform word alignment of each non-parsed Bible translation with each of the parsed translations.
3. Project the following information onto the non-parsed Bible translations:
    (a) Multilingual word embeddings.
    (b) Language-independent semantic concept labels.
    (c) Parts of speech and dependency relations.
4. Create inflectional paradigms for nouns and verbs in each language by combining projected parts of speech and concept labels.
5. Run a suite of multilingual neural models using the data produced so far, thereby creating the language representations used for this study.
6. Create word order statistics for each language and a number of word order parameters, using the projected dependency relations. These will be used for evaluating the language representations, in addition to being informative for typological research.

4. External Resources

In this section we describe the data we use from external sources, leaving data sets produced by us as part of this work to Section 5, and the typological databases used for evaluation to Section 7.1.

Table 1
The diversity of language families represented in the corpus of Bible translations, and in the mT5 language model. The division into linguistic macro-areas follows Hammarström et al. (2022), and each family is placed in the macro-area where most of its members are located. Pidgins, artificial languages, and unclassified languages are not counted.

Macro-area	Bible	mT5
North America	17	0
South America	39	0
Eurasia	19	13
Africa	15	2
Papunesia	36	1
Australia	6	0
Total	132	16

As the main multilingual resource, we use a corpus of Bible translations crawled from online sources (Mayer and Cysouw 2014). In the version used by us, it contains 1,846 translations in 1,401 languages, representing a total of 132 language families or isolates according to the classification of Hammarström et al. (2022). This is a unique resource, both in terms of the number of languages, and in their diversity. Table 1 compares the number and distribution of language families of this corpus, with the 101-language mT5 model (Xue et al. 2021) used for comparison as a representative of typical highly multilingual language models.[5]

The discrepancy between the number of translations and languages is due to some languages having multiple translations. Here we define **language** as corresponding to a unique ISO 639-3 code, while **doculect** refers to the language documented in a single translation. We exclude partial translations with fewer than 80% of the New Testament verses translated. The count of verses varies somewhat between different traditions, but we compute a canonical set of verses, defined as all verses that occur in at least 80% of all translations, in total 7,912 verses. We also exclude a few translations without suitable word segmentation. A total of 1,707 translations in 1,299 languages satisfy these criteria. For languages that we intend to use as source languages for annotation projection, we manually choose a single **preferred translation** per language. We apply the following criteria for the doculect in the preferred translation, in order of priority:

- The doculect should be as close as possible to the modern written variety of the language, in order to match the external resources. This typically means excluding old translations, based on metadata on publication year in the corpus.
- The translation should be as literal as possible, without extensive added elaborations or divergences.

5 Other models, such as mBERT and XLM-R, use data from very similar sets of languages.

- The translation should cover the Old Testament part of the Bible, in order to maximize the amount of parallel data to those target texts that contain both the New Testament and the Old Testament.

A total of 43 such translations are chosen. These are only used as sources for annotation projection, which brings the number of available *target* translations down to 1,664, in 1,295 languages. Note that most (39) of the 43 languages used as sources have multiple translations, which means that the non-preferred translations are used as targets during annotation projection (discussed further in Section 5).

For the word embedding projection (Section 5.2) we use the multilingual word embeddings of Smith et al. (2017), trained on monolingual Wikipedia data and aligned into a multilingual space using the English embeddings as a pivot. We chose the 32 languages with the highest word translation accuracy in the evaluation of Smith et al. (2017), and refer to these embeddings as **high-resource embeddings** below.

For dependency relation and part of speech tag projection, we lemmatize, PoS-tag, and parse Bible translations using the multilingual Turku NLP Pipeline (Kanerva et al. 2018). A total of 35 languages in the Bible corpus are supported by this model, and the preferred translation in each of these languages is annotated with lemmas, PoS tags, and dependency structure following the Universal Dependencies framework (McDonald et al. 2013).

For concept labels (Section 5.3) we rely on the Intercontinental Dictionary Series (IDS) (Key and Comrie 2015) and its connection to the Concepticon list of semantic concepts (List et al. 2022). This is a collection of digital lexicons for 329 languages or varieties, of which 25 are supported by the TurkuNLP lemmatizer. Since the IDS contains only citation forms, we only use the lexicons for these 25 languages.

5. Multi-source Projection of Information

We now turn to several types of resources that we have produced, for use as training data and evaluation. These resources rely on aligning words between a large amount of pairwise translations in the Bible corpus, and Section 5.1 below describes an efficient method for performing this task.

5.1 Subword-based Word Alignment

Word alignment is performed using subword-level co-occurrence statistics.[6] Since the typical translation pair is unrelated and the languages have very different morphological properties, we prefer this method over word-based alignments. The alignment score of two items, w (from language L_1) and u (from language L_2), compares two models for explaining co-occurences between w and u:

- M_1: Whether w and u occur in a given Bible verse is decided by draws from two independent Bernoulli distributions.

6 We initially experimented with using a more complex two-step procedure, where subword-level co-occurrence alignment was used to compute Dirichlet priors for a Gibbs sampling aligner based on a Bayesian IBM model (Östling and Tiedemann 2016). In spite of significantly larger computational cost we did not observe any substantial differences when evaluated on the task of inferring word order properties, as in Section 5.6.

- M_2: Whether w and u occur in a given Bible verse is decided by a draw from a categorical distribution with four outcomes (w only, u only, both, or neither).

In order to estimate our belief in M_2, a systematic co-occurrence,[7] we multiply our prior belief in M_2 with the Bayes factor of M_2 over M_1. Because a morpheme in one language is a translation equivalent of (very) approximately one morpheme in the other language, we use a prior of $1/V$ where V is the total number of unique subwords in L_1. We define a **subword** as any substring w of a token which has a higher frequency than any substring w' containing w. For instance, if the substring 'Jerusale' has the same frequency as 'Jerusalem', only the latter will be added to the subword vocabulary.

We use uniform Beta and Dirichlet priors, respectively, for M_1 and M_2. The resulting alignment score can thus be computed as follows, by combining the prior with the log-Bayes factor $\mathrm{BF}(M_2/M_1)$:

$$\begin{aligned} s(w,u) &= \log\frac{1}{V} + \mathrm{BF}(M_2/M_1) \\ &= \log\frac{1}{V} \\ &+ \log P(\langle n_w - n_{wu}, n_u - n_{wu}, n_{wu}, n - (n_w + n_u - n_{wu})\rangle|\mathbf{1}) \\ &- \log P(\langle n_w, n - n_w\rangle|\mathbf{1}) \end{aligned} \tag{1}$$

where n is the total number of verses that occur in both the L_1, and L_2 translations, n_w, n_u, and n_{wu} the number of verses containing w, u, and both, respectively. The Dirichlet-multinomial (and its special case, the Beta-binomial) likelihood function is given by

$$P(\mathbf{x}|\alpha) = \frac{\Gamma(\sum_i \alpha_i)}{\Gamma(\sum_i (x_i + \alpha_i))} \cdot \prod_k \frac{\Gamma(x_k + \alpha_k)}{\Gamma(\alpha_k)} \tag{2}$$

Note that Equation (1) gives a type-level score. In order to get token-level alignments, we greedily align each token in L_1 to the highest-scoring token in the corresponding verse of L_2. The score $s(w,u)$ is then used as a threshold to filter out tokens that should be left unaligned. In our experiments, we use the criterion $s(w,u) \geq 0$, in other words, that M_2 should be at least as credible as M_1. In addition, we use a few empirically determined thresholds for additional filtering: the log-Bayes factor $\mathrm{BF}(M_2/M_1)$ must be greater than $0.2n_{wu}$ *and* greater than $\min(100, 0.7n_{wu})$.

5.2 Multilingual Word Embeddings

The multilingual high-resource embeddings described in Section 4 cover only 32 languages in our sample, which corresponds to less than 3% of the languages in the Bible

7 Note that M_2 simply describes that w and u are not independently distributed, which could also mean that they have a complementary distribution. Since we still align on a token level, requiring instances of w and u to be present in the same verse, this is not a problem in practice.

corpus. In order to obtain multilingual word embeddings for all languages we study, we perform word alignment as described above, followed by naive multi-source projection by averaging over the embeddings of all aligned tokens. We use only one translation per language as source. When multiple translations exist for a given language, we have aimed to choose the one closest matching the relatively modern language that the high-resource embeddings have been trained on. In total, we project embeddings to 1,664 translations in 1,295 different languages.

5.3 Semantic Concepts

In order to obtain annotations of semantic concepts for each language, we use lexicons in 25 languages from the IDS (Key and Comrie 2015), which was described further in Section 4. In total 329 languages are available in the IDS, but we only use a subset of 25 languages where we have access to accurate lemmatizers. Each IDS lexicon entry is connected to a common inventory of semantic concepts from the Concepticon (List et al. 2022), such as TREE, WATER, and WOMAN. For each token we assign any concepts that are paired with its lemma in the IDS database. We choose a single semantic concept of a target-text token using a simple majority vote among all the aligned source text tokens, as long as at least 20% of source texts agree on the given concept label. This procedure is identical to the PoS and dependency relation projection described in Section 5.6.

5.4 Paradigms

For our reinflection model (Section 6.3) as well as the affixation type evaluation data (Section 5.5) we need examples of (partial) inflectional paradigms for each language. We approximate these using a combination of the PoS projections (Section 5.6) and semantic concept projections (Section 5.3). To obtain paradigm candidates for a given language, we perform the following heuristic procedure:

1. For each semantic concept, find the PoS tag most commonly associated with it.
2. Among the word forms with the given projected concept label and PoS tag, perform hierarchical clustering using mean pairwise normalized Levenshtein distance as the distance function.
3. Select clusters with at least two members, with at least one word form above 4 characters in length, and with a mean pairwise normalized Levenshtein distance below 0.3.

The normalized Levenshtein distance used is $d(s_1, s_2)/(|s_1| + |s_2|)$, where d is unweighted Levenshtein distance (Levenshtein 1966).

This method also means that we have an estimate of the part of speech for each paradigm, and in the present work we use this information to restrict our study to only noun and verb paradigms. Any part of speech with less than 50 partial paradigms identified is considered to lack inflection. Such low counts have been empirically determined to arise from noise in the alignment procedure.

5.5 Affixation Type

Using noun and verb paradigms estimated in Section 5.4, we can guess the proportion of prefixing and suffixing inflections by the following procedure. First, we sample 1,000 word pairs for each part of speech from each Bible translation, such that the word in each pair comes from the same paradigm, such as *annotate–annotating*. We then use the Levenshtein algorithm to compute the positions of the edit operations between the two words. If all operations are performed on the first half of each word, we count the pair as prefixing. If all operations are performed on the second half, we count it as suffixing. Otherwise, we count it as neither.

We evaluate the result of this heuristic by comparing against Dryer (2013f). To investigate the effect of avoiding ambiguous cases, we consider two cases. In the **Non-exclusive** condition, prefixing languages are those classified as weakly or strongly prefixing, or as being equally prefixing and suffixing. In the **Exclusive** condition, only languages that are weakly or strongly prefixing are counted, and all other languages (with either little affixation, or equally prefixing and suffixing) are discarded from the analysis. We define suffixing similarly.

Table 2 shows the level of agreement with Dryer (2013f). The table presents accuracy as well as F_1 scores. The F_1 score presented is the mean of both classes, positive and negative. Since our heuristic classifies all languages as either prefixing or suffixing, we mainly consider the **Exclusive** condition. Because our sample is strongly biased toward a few large language families, we focus on the **Family**-balanced scores, which weigh each doculect so that all top-level language families receive unit weight. For **Language** weighting, each ISO 639-3 language code receives unit weight, which is more easily comparable to previous work. We here achieve a family-balanced accuracy of 85.6% and a mean F_1-score of 0.798. This result is pulled down mainly by the low performance for identifying prefixing languages (recall 74% and precision 65%).

Concurrent work has confirmed that automatic estimation of affixation type is quite challenging (Hammarström 2021), for a variety of reasons including the difficulty of identifying productive patterns, and differentiating between inflectional and derivational morphology. Dryer (2013f) specifically concerns inflectional morphology, whereas our method is not able to fully separate inflectional morphology from derivational morphology, or affixes from clitics. We also note that while Dryer (2013f) counts the number of categories marked by affixes, we are counting the number of word forms with a given affix. Given the high agreement reported above, we do, however, consider our approximation to be good enough for further investigation.

5.6 Word Order Statistics

The typological databases used in our evaluation (described further in Section 7.1) have two shortcomings: they are sparse and categorical. Through multi-source projection is it possible to obtain reliable word order statistics (Östling 2015) for all of the languages in our data, which makes us able to compare how well our data (Bible translations) matches the typological databases used. It is also possible to use the projected features as classifier training data in the evaluation, and as a reference point for analyzing the classification results.

We use the token-level word alignments between each of the 35 Universal Dependencies-annotated translations (see Section 4) and the 1,664 low-resource translations to perform multi-source projection of PoS tags and dependency relations. Note

Table 2
Projected properties. The word classes ADJ* and NUM* are narrower versions of the corresponding UD word classes; see the main text for details. Accuracy and F_1 values are with respect to URIEL values from WALS and Ethnologue. **Exclusive** counts only languages where URIEL codes exactly one of a mutually exclusive set of options as true. **Non-exclusive** uses all available data. **Language** gives each ISO 639-3 language code equal weight, while **Family** gives each Glottolog family identifier equal weight.

		Language		Family	
Label	**Definition**	**Accuracy**	**F_1**	**Accuracy**	**F_1**
	Non-exclusive				
Object/verb order	NOUN/PROPN $\xleftarrow{\text{obj}}$ VERB	94.7%	0.945	87.0%	0.866
Oblique/verb order	NOUN/PROPN $\xleftarrow{\text{obl}}$ VERB	76.1%	0.640	71.7%	0.637
Subject/verb order	NOUN/PROPN $\xleftarrow{\text{nsubj}}$ VERB	81.4%	0.689	84.3%	0.578
Adjective/noun order	ADJ* $\xleftarrow{\text{amod}}$ NOUN	81.7%	0.799	78.7%	0.778
Relative/noun order	VERB $\xleftarrow{\text{acl}}$ NOUN	91.9%	0.850	86.5%	0.797
Numeral/noun order	NUM* $\xleftarrow{\text{nummod}}$ NOUN	92.6%	0.926	89.2%	0.889
Adposition/noun order	ADP $\xleftarrow{\text{case}}$ NOUN	94.8%	0.947	95.8%	0.955
Prefixing	Prefixes $\geq$ 50%	80.9%	0.766	83.5%	0.804
Suffixing	Suffixes $\geq$ 50%	70.7%	0.646	71.2%	0.619
	Exclusive				
Object/verb order	NOUN/PROPN $\xleftarrow{\text{obj}}$ VERB	95.8%	0.957	88.6%	0.880
Oblique/verb order	NOUN/PROPN $\xleftarrow{\text{obl}}$ VERB	76.1%	0.640	71.7%	0.637
Subject/verb order	NOUN/PROPN $\xleftarrow{\text{nsubj}}$ VERB	86.8%	0.735	92.3%	0.673
Adjective/noun order	ADJ* $\xleftarrow{\text{amod}}$ NOUN	85.8%	0.846	85.5%	0.850
Relative/noun order	VERB $\xleftarrow{\text{acl}}$ NOUN	92.4%	0.861	90.4%	0.851
Numeral/noun order	NUM* $\xleftarrow{\text{nummod}}$ NOUN	95.1%	0.951	92.0%	0.918
Adposition/noun order	ADP $\xleftarrow{\text{case}}$ NOUN	97.6%	0.975	98.1%	0.980
Prefixing	Prefixes $\geq$ 50%	87.3%	0.808	85.6%	0.798
Suffixing	Suffixes $>$ 50%	Identical to Prefix in this condition			

that for our purposes we do not need to produce full dependency trees, so dependency links are projected individually.[8] Each PoS tag, dependency head, or dependency label needs to be projected from at least 20% of the available source texts. Otherwise the projection is discarded, as a means of filtering out inconsistent translations and poorly aligned words.

For each language we count the proportion of head-initial orderings for each dependency label and head/dependent PoS combination, to obtain a word order feature matrix covering all languages. The projected word order properties are listed in Table 2.

8 We have experimented with using maximum spanning tree decoding to ensure consistency, but did not observe any improvement in word order estimation.

For instance, the well-studied typological parameter of object/verb order (where the object is headed by a noun) is captured by the head-initial ratio of NOUN/PROPN $\overset{\text{obj}}{\longleftarrow}$ VERB relations. A value of 0 would indicate strict object–verb order, while 1 indicates strict verb–object order, and 0.5 indicates that both orderings are equally frequent on a token basis.

A fundamental assumption in annotation projection is that grammatical relations are the same across translation equivalent words in different languages. While this does not hold in general, several things can be done to make the approximation closer. One source of disagreement is the differences in part-of-speech categories across languages. By focusing on core concepts of each category we can decrease the number of cases where translation equivalents participate in different syntactic relations because they belong to different parts of speech. Dixon (1982, pages 3–7) showed that a small set of concepts are most likely to be lexicalized across languages as true adjectives, that can be used attributively. When estimating adjective/noun order, we limit ourselves to this set in order to minimize the divergence of syntactic constructions across languages.[9] Östling and Wälchli (2019) showed that restricting the category of adjectives when projecting relations across Bible texts results in a much closer match to the adjective/noun order data from Dryer (2013a), as compared to using the Universal Dependencies ADJ tag. A similar approach was taken for numerals, where only the numerals 2–9 were chosen. This range was chosen to ensure that for the vast majority of languages with a numeral base of 10 or above (Comrie 2013), only atomic numerals would be chosen and the problem of complex numeral constructions for higher numbers can be avoided (Kann 2019). The word for the numeral 1 is often used for other functions (cf. the article *ein* in German), which would have posed additional challenges for accurate parsing and annotation projection.

One problem with using the core adjective concepts of Dixon is that these sometimes stand out from the larger class of adjectives with respect to word order. A familiar example is the Romance languages, where many of the core adjectives use adjective–noun order instead of the more productive noun–adjective order, but examples are spread across the world (Östling and Wälchli 2019). An alternative method would have been to automatically separate attributive constructions from other types of constructions, but this is a complex problem beyond the scope of this work.

For nouns and verbs, we simply use the Universal Dependencies NOUN and VERB tags, respectively. The high level of agreement with verb/object order data from Dryer (2013e) indicates that this approximation is accurate.

Table 2 shows the featues we project, their definition in terms of projected Universal Dependencies relations, and the level of agreement with WALS and Ethnologue (as aggregated and binarized by the URIEL database). All values are binarized so that a majority of head-initial projected relations give the value 1, otherwise 0. When multiple URIEL features describe the same phenomenon, the one with the head-initial interpretation is chosen (e.g., S_OBJECT_AFTER_VERB) for consistency. Projections are summarized at the level of a doculect, in our case a single Bible translation. Each language may have multiple translations, and a given language family may be represented by multiple languages. As mentioned in Section 5.5 above, we report results by weighting so that either languages or language families are given identical weight. We consider the latter to be more informative, since it approximates the expected performance on

9 We used the following Concepticon labels to define core adjectives: STRONG, HIGH, GOOD, BAD, SMALL, BIG, NEW, YOUNG, OLD, BEAUTIFUL.

a newly discovered language from a previously unknown family. Language-weighted numbers are included for ease of comparison with previous work, and to show the effect of using a language sample biased toward some families. We consider mean F_1 scores to be more informative, since several of the features are heavily biased toward one class, which often leads to inflated accuracy figures (e.g., subject/verb order) for methods biased toward the majority class.

Overall, there is a high level of agreement between the projected features and the classifications from WALS and Ethnologue. Looking at the **Exclusive** condition, which we use in our later experiments, the family-wise mean F_1 scores are 0.8 or above for all features except subject/verb and oblique/verb order. A thorough error analysis is beyond the scope of this work, but some previous work exists on projected typological parameters. Östling and Wälchli (2019) investigated projected adjective/noun order and found a varied number of causes for disagreements with typological databases, including coding errors in the databases themselves, and differences between the Bible translation and reference grammar doculects. In Section 8.4, we investigate a number of cases where the projections and databases disagree and find that those can be explained by languages with mostly free word order having been manually classified as having some dominant word order. This is also in line with the findings of Choi et al. (2021), who compared quantitative word order data from Universal Dependencies treebanks with WALS classifications.

At this point we should add that the classifications derived from projected data are never assumed to be correct in our evaluation. Instead, they are used as *training* data in some of our classification experiments, while only URIEL is used as a gold standard for comparison. We do, however, use projected labels as a complement in our error analysis in Section 8.4.

6. Language Representations

In order to capture different types of linguistic structure, we use a number of different neural models for creating language representations.[10] The model types are chosen to maximize the diversity of the learned representations, while requiring only the available data described in Sections 4 and 5.

We use the following models, which generate the language representations whose labels are in bold:

- Word-based language model with multilingual word embeddings (WORDLM)
- Character-based language model (CHARLM)
- Morphological reinflection of noun paradigms (REINFLECT-NOUN) or verb paradigms (REINFLECT-VERB)
- Word form encoder from characters of a word form to the multilingual word embedding space (ENCODER)

10 For languages with multiple Bible translations, we learn one representation per translation (doculect). The exception is the ASJP-based model and the language representations from previous work, which are all on the (ISO 639-3) language level. For simplicity, we use *language representation* for both levels of granularity.

- Neural machine translation models: many-to-English (NMTX2ENG) and English-to-many (NMTENG2X)
- Baseline representations from pairwise lexical similarity (LEXICAL and ASJP)

The models will be detailed in the following subsections.

6.1 Word-level Language Model

We train a language model to predict the embedding of the following word in the (fixed) multilingual embedding space described in Section 5.2. This consists of a simple left-to-right LSTM conditioned on the preceding word and a language embedding, whose output is projected through a fully connected layer to the multilingual word embedding space. As loss function, we use the cosine distance between the predicted word embedding and the actual word at that position in training data. This allows us to efficiently train the model with a vocabulary size of 18 million word types, where the computational cost of softmax normalization would be prohibitively high. While alternatives to our choice of cosine distance as a loss function could certainly be explored, the promising results obtained by this model in our evaluation makes us leave exploring alternatives for future work.

The choice of word-level vocabulary is due to our desire to keep all lexical information in the fixed embeddings. If a subword vocabulary of reasonable size was used over all 1,295 languages, the units would be relatively small and a large number of parameters would be required by the model just to memorize vocabulary across all languages, potentially reducing its ability to model syntax and semantics.

Only the LSTM parameters, the fully connected layer following it, and the language representations are updated during training. The word embeddings are fixed. Sentences of all languages are mixed, and presented in random order. In the experiments, we use 512-dimensional LSTM with 100-dimensional language embeddings. For the regularization, we use a dropout layer with probability 0.3 between the LSTM and the hidden layer.

Since semantic information is encoded in a language-independent way by the multilingual word embeddings, our intention with this model is for the LSTM to learn a language-agnostic model of semantic coherence, while relying on the language representations to decide how to order the information—that is, the syntax of each language. We refer the representations obtained from this model as WORDLM.

6.2 Character-based Language Model

We train a single LSTM language model over the characters making up each sentence in all languages. The model is conditioned at each time step only on the preceding character and a language embedding. The character embeddings are shared between languages. Sentences from all languages are mixed, and presented in random order. All parameters of the model are learned from scratch during training.

Ideally, we would want to train this model using an accurate transcription in, for instance, the International Phonetic Alphabet (IPA), but the Bible corpus is generally only available in the standard orthography (or orthographies) of each language. Because a number of very different writing systems are used, it is not possible to directly use the

raw text. To approximate a phonemic transcription, we use standard transliteration[11] into Latin script, followed by a few rules for phonemes generally represented by multi-grapheme sequences across Latin-based orthographies (e.g., *sh* → ʃ), as well as merging some vowels and voicing distinctions to reduce the size of the inventory. If accurate multilingual grapheme-to-phoneme (G2P) systems become available that cover most of the languages in the Bible corpus, that would of course be a much preferred solution since our approximations are not valid for all languages and orthographies.

This is roughly equivalent to the model of Östling and Tiedemann (2017), except that we use a pseudo-normalized Latin orthography rather than native writing systems. We refer to the model as CHARLM. We use a 128-dimensional LSTM, 100-dimensional character embeddings, and 100-dimensional language representations. We also use a dropout layer with probability 0.3 between the LSTM and the dense layer for regularization.

6.3 Multilingual Reinflection Model

We train an LSTM-based sequence-to-sequence model with attention to predict one form in an inflectional paradigm given another form. In spirit, this is similar to the reinflection task of Cotterell et al. (2016), except that we do not have access to accurate annotations of morphological features. Instead we simply pick random target forms without providing the model any further information. This model is implemented with OpenNMT (Klein et al. 2017) using default hyperparameters. We train two sets of language representations: (1) using only noun paradigms (REINFLECT-NOUN), (2) using only verb paradigms (REINFLECT-VERB).[12] The language of each example is encoded by a special language token, whose embedding becomes the language embedding for that language.

The model has direct access to the source form through the attention mechanism, and our intention is that it will learn to copy the lexical root of the source form to the target, needing only to learn which transformations to apply (e.g., removal and addition of affixes), and not to memorize the vocabularies of all languages. We expect the language representations to encode the necessary morphological information to perform this transformation. This is similar to the use of morphological inflection for fine-tuning language representations in Bjerva and Augenstein (2018a), except that we rely only on cross-lingual supervision and are thus able to directly train the model for the whole Bible corpus.

6.4 Word Encoder Model

The reinflection model described in the previous section is only concerned with predicting some other member of the same inflectional paradigm, without considering the properties of that form. It is therefore not possible for the model to connect a certain form with, say, number marking on nouns or tense marking on verbs. For this reason, we also train a model to encode word forms represented as transliterated character sequences into the multilingual word embedding space from Section 5.2. This model consists of a 2×128-dimensional BiLSTM encoder over a character sequence, followed

11 Using the transliteration tables from the Text::Unidecode library of Sean Burke.

12 As a sanity check, we have sampled from the model and as expected the *k*-best list of translations generally contains correct (but arbitrary) inflections of the lemma that the source form belongs to.

by an attention layer and a fully connected layer. We use cosine distance loss, as in the multilingual language model from Section 6.1. The target language is represented by a special token for each language, whose embedding becomes the language embedding for that language.

Our aim with this model is to capture not only general tendencies of inflectional morphology, but also the presence and location of specific markers (such as case suffixes, or number prefixes). We refer the representations obtained from this model as ENCODER.

6.5 Machine Translation Models

Inspired by Malaviya, Neubig, and Littell (2017), we train a many-to-English (NMTX2ENG) and an English-to-many (NMTENG2X) neural machine translation system. These are implemented in OpenNMT (Klein et al. 2017), using 512-dimensional LSTM models with a common subword vocabulary on the transliterated and normalized data described above in Section 6.2. For the many-to-English model, the source language is encoded using a unique token per language, while for the English-to-many model it is the target language that is encoded by a unique token. The embeddings of these tokens are used as language representations.

6.6 Lexical Similarity

For comparison purposes, we include two non-neural baselines which contain only *lexical* information about languages. The first is derived from the ASJP lexical database (Wichmann, Holman, and Brown 2018), which contains 40-item word lists of core vocabulary for a large number of languages. A total of 1,012 languages (unique ISO 639-3 codes) occur in the Bible corpus and have sufficiently complete (at least 30 items) word lists in ASJP. We follow Bakker et al. (2009, p. 171) in measuring the distance between two languages by taking the mean normalized Levenshtein distance between same-concept word forms, divided by the mean normalized Levenshtein distance between different-concept word forms.[13] If multiple varieties of the same (ISO 639-3) language are present in ASJP, the union of word forms over all varieties is used. We compute a 1012×1012 pairwise distance matrix, which we reduce to 100 dimensions using truncated SVD as implemented by Pedregosa et al. (2011).[14] We refer to this set of language representations as ASJP.

To further increase the variation, we also consider using a separate lexical dataset, namely, the Bible corpus itself. This also has the advantages of increasing the number of languages covered, and allowing representation at the doculect level (i.e., individual Bible translations). We use subword alignments to project our own multilingual word lists using the 2013 English New World translation as a pivot. In order to avoid proper nouns, only non-capitalized lemmas were used, and in order to ensure that the word list is not too sparse, only English lemmas that are reliably aligned to substrings in at least 75% of Bible translations are included. In total, 105 lemmas satisfy these criteria. All word forms are transliterated into the Latin alphabet and normalized as in Section 6.2, to allow for direct string comparison. Then the pairwise distance calculations and

13 For consistency with Bakker et al. (2009), we normalize by dividing by $\max(|s_1|, |s_2|)$.

14 We also attempted to use UMAP (McInnes et al. 2018), but found the structure of the resulting vectors to lead to instability during classifier training.

dimensionality reduction is performed as for the ASJP vectors. We refer to the resulting set of language representations as LEXICAL.

6.7 Pre-trained Multilingual Language Models

As discussed in Section 1.2.4, recent work has demonstrated that multilingual language models encode cross-lingual structural features. While these models are only trained on a relatively small and biased subset of our entire language sample (see Section 4 for a detailed discussion on these biases), we here apply them to all Bible translations with the exception of those with scripts that are not present in the language model training data (e.g., Coptic). This means that most of the languages evaluated on are not in the language model's training data. We choose this method for two reasons: First, our cross-validated classification evaluation method requires a sufficiently large and varied sample of languages, and second, multilingual language models have demonstrated the ability to transfer some knowledge even to languages outside of their training set through lexical similarity with in-training languages.

We obtain language representations from three pretrained multilingual language models:

- MBERT: Multilingual BERT (Devlin et al. 2019)
- XLM-R-BASE: XLM-RoBERTa base (Conneau and Lample 2019)
- XLM-R-LARGE: XLM-RoBERTa large (Conneau and Lample 2019)

For all of these, representations are obtained by first running each Bible verse individually through the model and computing the mean of the last-layer token representations for each verse. The vectors obtained from each verse are then averaged over the whole translation to obtain a representation of the language contained in the translation.

7. Experiments

As set out in the Introduction, we are interested in finding out to what extent we can control the type of information captured by language representations, and whether language embeddings from neural models make human-like typological generalizations. We do this by answering, for a large number of typological features f, how well a given set of language representations L capture f. Specifically, we find the extent to which f can be predicted from L alone using a logistic regression classifier. For ease of analysis, we train a binary logistic regression classifier for each feature with equal weights for the positive and negative class. This avoids biasing classifiers according to the data label distribution, which allows easier comparison between different subsets of the data, with different label distributions. In addition, our sampling procedure (described further below) gives equal weight to language families, regardless of how many members they contain.

7.1 Evaluation Data

The typological features used in this study are derived from two types of sources: traditional typological databases (following, for instance, Malaviya, Neubig, and Littell

2017), as well as a novel dataset consisting of word order features obtained from annotation projection in the Bible corpus.

7.1.1 Typological Databases. We use the URIEL typological database (Littell et al. 2017), specifically, the features derived from WALS (Dryer and Haspelmath 2013) and Ethnologue (Eberhard, Simons, and Fennig 2019). Features from these sources are used as gold standard labels for the evaluation. Note that the binarization of features in URIEL requires some simplification to the (already simplified) coding in the original data source. Features representing several mutually contradictory values may simultaneously be true. For instance, Irish is coded in URIEL as tending toward suffixing morphology, but also tending toward prefixing (it is coded as "Equal prefixing and suffixing" by Dryer [2013f]), while German according to URIEL has both object after verbs and object before verbs (it is coded as "No dominant order" by Dryer [2013d]). We resolve this by keeping only those instances in the data where exactly *one* of a set of mutually incompatible variables is true.

7.1.2 Projected Features. Five types of projected word order statistics described in Section 5.6 (object/verb order, subject/verb order, adjective/noun order, numeral/noun order, adposition/noun order) are used as training data for the classifiers, but never as gold standard labels for evaluation. This data has the advantage of being available for all languages in the Bible corpus, which allows more languages to be used for training than if we would restrict ourselves to the languages present in URIEL for the given feature. In addition, the morphological feature indicating whether prefixing or suffixing morphology dominates is used.

7.2 Cross-validated Classification

Our basic measure of whether a set of language representations encodes a specific typological feature is cross-validated classification performance, measured using F_1 score (the mean of the F_1 of the positive and negative classes). As described in Section 3, we use constrained leave-one-out cross-validation, taking care to exclude languages from the training fold that could be suspected to be non-independent of the evaluated language. All languages with gold standard labels available are classified, and the results are weighted in order to give either languages (defined according to ISO 639-3 codes) or language families (defined according to Glottolog family identifiers) equal weight. We consider family-weighted F_1 score to be the single most useful measure of classifier success, and this is what we report unless otherwise specified.

The uncertainty is estimated by Monte Carlo sampling, where 401 samples are drawn such that only one language from each family is chosen. As a dummy baseline, we train classifiers using the same parameters and data but with randomly shuffled target labels. This establishes a baseline range of F_1 and accuracy values that would be expected from a classifier that has not learned to predict the given feature at all.[15] The non-baseline classifier variance across Monte Carlo samples is due to different training folds being chosen each sample. When a single classification is extracted, the type value across all samples is used.

15 We find that this baseline chance level agrees well with a binomial (0.5) model, as expected. Computing this baseline empirically rather than relying on a theoretical model helped us to diagnose an issue with insufficient regularization.

If less than 50 language families are represented in the evaluation set for a particular feature, we skip evaluating it due to data sparsity. A total of eight features related to word order, and seven related to morphology, had sufficient sample sizes to be evaluated. The features are discussed in more detail in Section 8.1 and Section 8.2.

8. Results and Discussion

We will now describe the results of our evaluations for our own models (see Section 6) as well as of two previous studies. From Malaviya, Neubig, and Littell (2017) we use two sets of language representations derived from the same model: MTVEC (language embeddings) and MTCELL (averaged LSTM cell states). From Östling and Tiedemann (2017) we use the concatenated embeddings that were fed into the three LSTM layers, here labeled Ö&T. Some other authors have investigated language representations for smaller sets of languages, but our evaluation set-up is unsuitable for samples much smaller than a thousand languages.

In the figures below, we present the mean family-weighted F_1 for each set of language representations, for each feature of interest. Language representations are grouped in five groups that are visually distinguished in the figures:

1. Lexical baselines: ASJP and LEXICAL. These should, by design, not encode any structural features of language.
2. Neural Machine Translation (NMT): our NMTX2ENG and NMTENG2X models, as well as MTCELL and MTVEC from Malaviya, Neubig, and Littell (2017).
3. Character-level language models: our CHARLM and the previously published Ö&T (Östling and Tiedemann 2017).
4. Word-level language model: our WORDLM.
5. Word form models: our REINFLECT-NOUN, REINFLECT-VERB, and ENCODER.

Each figure has a dotted line indicating the 99th percentile of the shuffled-label baselines. This should be seen as a very rough baseline indicator, since we do not have a good way of modeling the complex distribution of classification results obtained from the (hypothetical) set of all possible language representations that do not encode typological features, given our sampling distribution of training languages. Language representations derived from lexical similarity exceed this baseline in two cases, though only by a small amount, so it likely represents an under-estimation of the actual baseline distribution. We do not interpret results exceeding this baseline as definite confirmations of typological features being encoded in the given language representations.

Some figures also have a dashed line, indicating the mean F_1 of projected labels with respect to the gold standard in URIEL. These correspond to the rightmost column in Table 2. We include the projection performance because it represents what can be done using hand-crafted methods on the same parallel text data as we have used for creating the language representations. Reaching this level indicates that the classifier has likely become about as good as can be expected given the underlying data.

Note that some representations (ASJP, MTCELL, MTVEC, Ö&T) are based on other data or other versions of the Bible corpus with a different set of languages, and thus have slightly different baselines. We have computed the baselines individually for each

set of language representations to confirm that our conclusions hold, but choose not to represent this in the figures for readability. The dotted and dashed lines in the figures are generated from the version of the Bible corpus used by us.

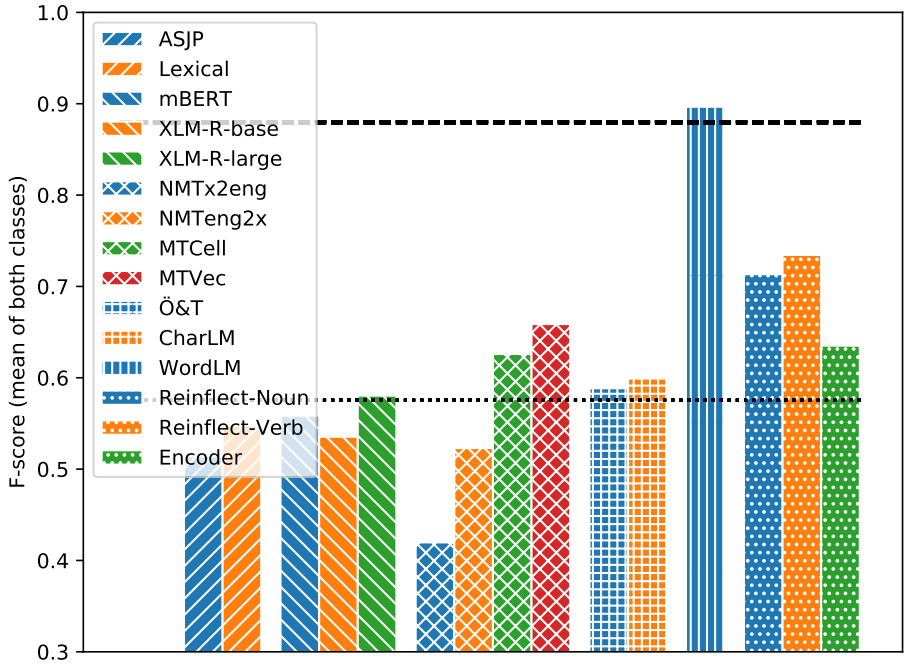
We wish to emphasize that if a set of language representations encode a typological feature in a useful way, given the hundreds of data points we use for training, we expect the classifier to be highly accurate. In contrast, with our evaluation set-up we expect classifiers to perform (approximately) randomly if there are no relevant typological features encoded in the language representations used to train them. Since the relevant differences in classification accuracy are very large, we present the main results as bar plots, complemented by exact numbers in the text only when we deem relevant. Differences between poorly performing classifiers are not relevant for our purposes, and we refrain from summarizing the complete data in a separate table. We should add that correlations between typological features somewhat complicate this binary distinction, but this is only relevant for the few language representations that actually seem to encode typological features, and those are analyzed in detail below.

8.1 Word Order Features

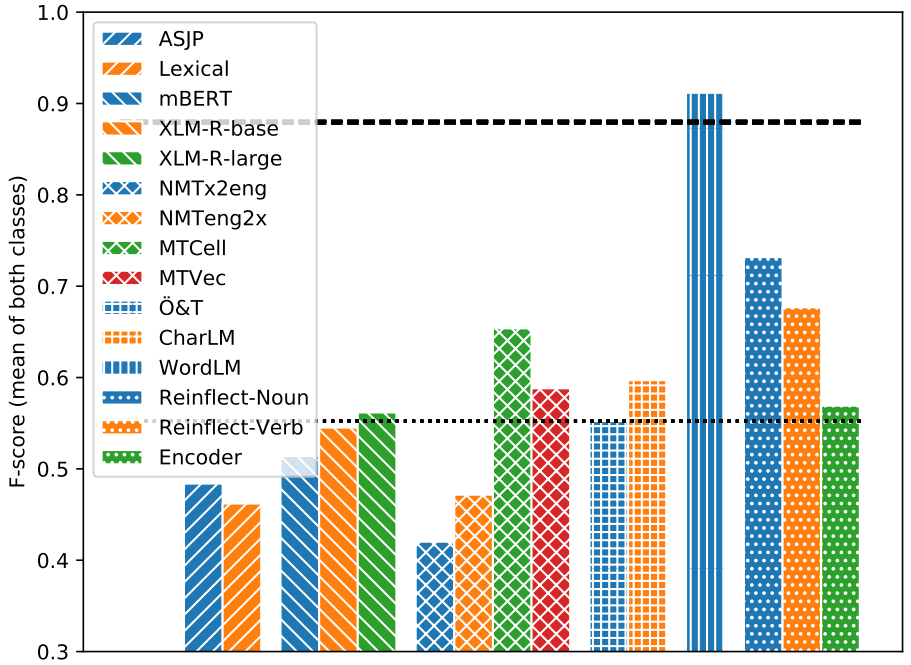
We start by looking at Figure 1a. The first thing to notice is that only the language representations from our word level language model (WORDLM) reach an F_1 score comparable to (and even slightly above) that of the projection method. This indicates that only the word level language model has managed to capture the order of object and verb, at least in a way that is separable by a linear classifier. The lexical baselines (ASJP and LEXICAL) encode lexical similarity between languages, and so are strongly correlated with word order properties within related languages or languages in contact. As intended, our evaluation set-up prevents these models from learning to identify even a clear and evenly distributed feature like the order of object and verb. Character-level language models (CHARLM and Ö&T) do not seem to encode word order properties, which indicates that they have not learned representations at the syntactic level. This is not surprising, since both models are relatively small and unlikely to learn enough vocabulary to generate to the level of syntax.

The word form models, in particular the reinflection models (REINFLECT-NOUN and REINFLECT-VERB), obtain moderately high F_1 values of around 0.7. Yet it is obvious that these models do not have sufficient data to conclude what the order of object and verb are in a language, since their input consists entirely of automatically extracted inflectional paradigms. We therefore suspect that the relative success in predicting may be due to the classifiers learning to predict *another* feature that correlates with the order of object and verb. To investigate whether this explanation is correct, we compute the corresponding F_1 scores for the classifier predictions with respect to each typological feature where we have data. In this case we find that classifications from both reinflection models are much better explained (REINFLECT-NOUN: +0.19 F_1, REINFLECT-VERB: +0.05 F_1) by the affix position (Dryer 2013f) feature.[16] In effect, the object/verb order labels we used for training were treated as noisy affix position labels, and the resulting classifier becomes much better at predicting affix position than object/verb order. An even clearer illustration of this can be found for the order of adposition and

16 Each pair of features has a unique set of overlapping languages, which we use in these comparisons in order to obtain comparable results. These F_1 differences from these head-to-head comparisons may not be equal to those obtained from using all available data for each feature, as we have presented in the figures.



(a) Order of object and verb, using gold standard labels for training.



(b) Order of object and verb, using projected labels for training.

Figure 1
Classification results for each set of language representations.

noun (see Figure 2), reflecting Greenberg's universal 27 (Greenberg 1963) on the cross-linguistic association of prepositions with prefixing morphology, and postpositions with suffixing.

There has been a long-standing debate on whether observed correlations between typological features are due to universal constraints on language, or simply due to genealogical and/or areal relations biasing the statistics (e.g., Dunn et al. 2011). We remain agnostic with regard to this question, but note that analyzing correlations between typological features is a challenging statistical problem. In this work we test all other features for which we have data, and mention which ones seem like plausible alternative explanations for a given classification result in terms of comparable or higher F_1 scores, without attempting to quantify their relative probability of the different explanations.

To summarize, we observe clear detections of the following typological features related to word order:

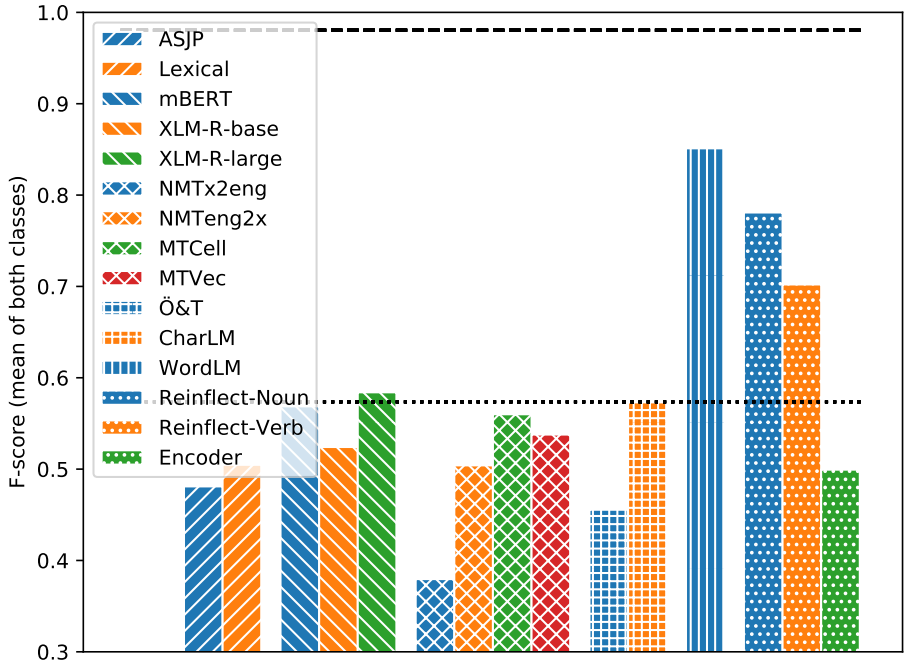
- Order of object and verb, for the WORDLM representations (Figure 1).
- Order of adposition and noun (prepositions/postpositions), for the WORDLM representations (Figure 2).
- Order of numeral and noun, for the WORDLM representations (Figure 3). Note that no representations obtained a mean F_1 above 0.7 when trained on URIEL data. As discussed in Section 8.4, this may be due to the much larger sample of languages with projected labels.
- Order of possessor and noun, for the WORDLM representations (Figure 4a). However, this result is about equally well explained (F_1 within 3 percentage points) by object and verb order, as well as adposition and noun order, so we consider this detection tentative.
- Order of numeral and noun, for the multilingual language models MBERT, XLM-R-BASE, and XLM-R-LARGE.

In addition to the features presented in the figures, we also examined all other features in URIEL with sufficiently large samples for our evaluation method. The following features that relate to word order or the presence of certain categories of words were examined:

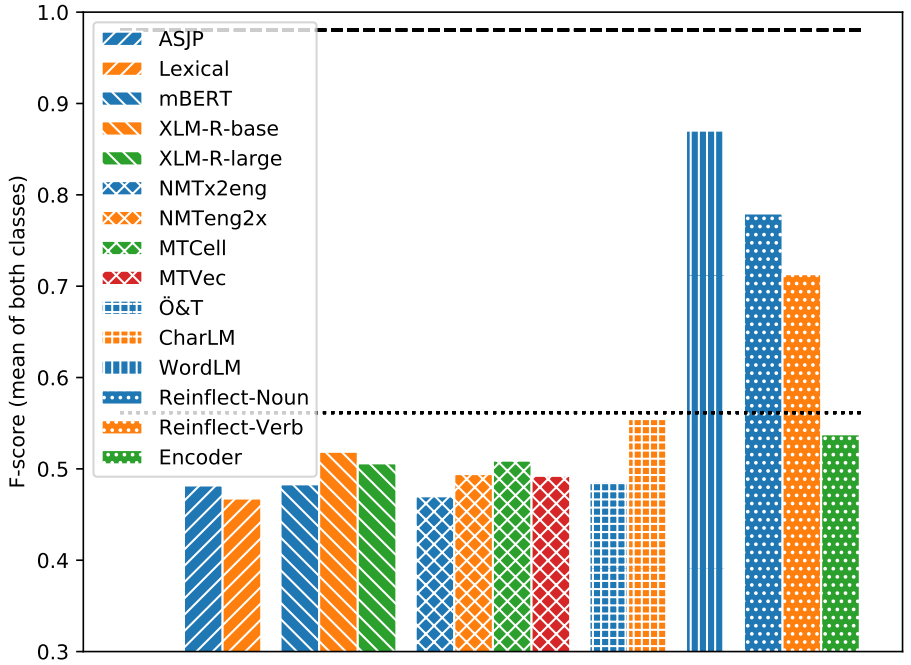
- Order of demonstrative word and noun
- Order of relative clause and noun (Figure 5a)
- Order of subject and object
- Existence of a polar question word

None of the language representations yielded classifications with an F_1 above 0.7 for either of the above features.

It is interesting to note where we did *not* see any clear indications of typological features encoded in the language representations. For at least some classical word order features, we see that there is sufficient information in the data to learn them, yet all models fail to do so.

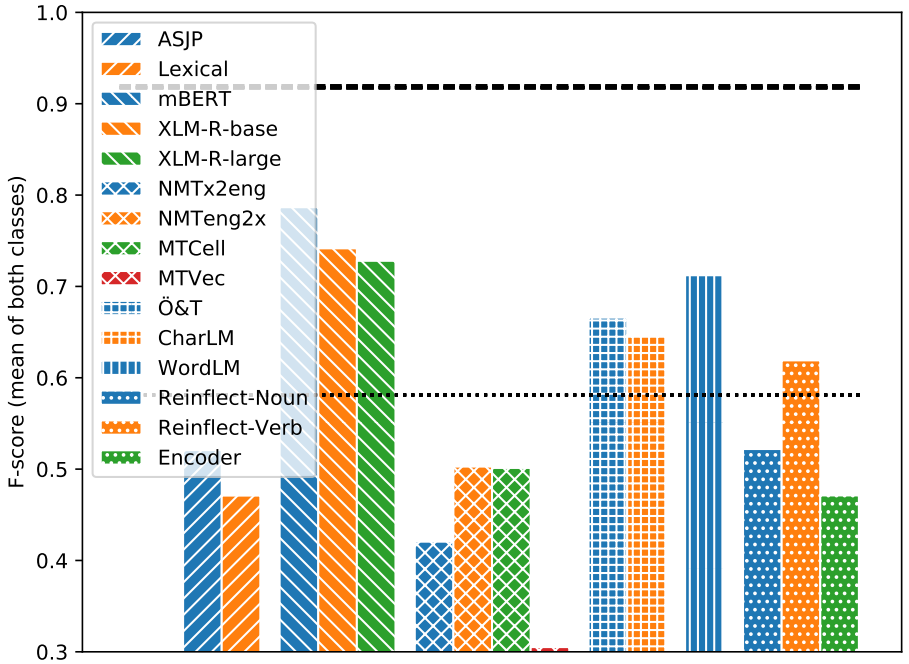


(a) Prepositions vs. postpositions, using gold standard labels for training.

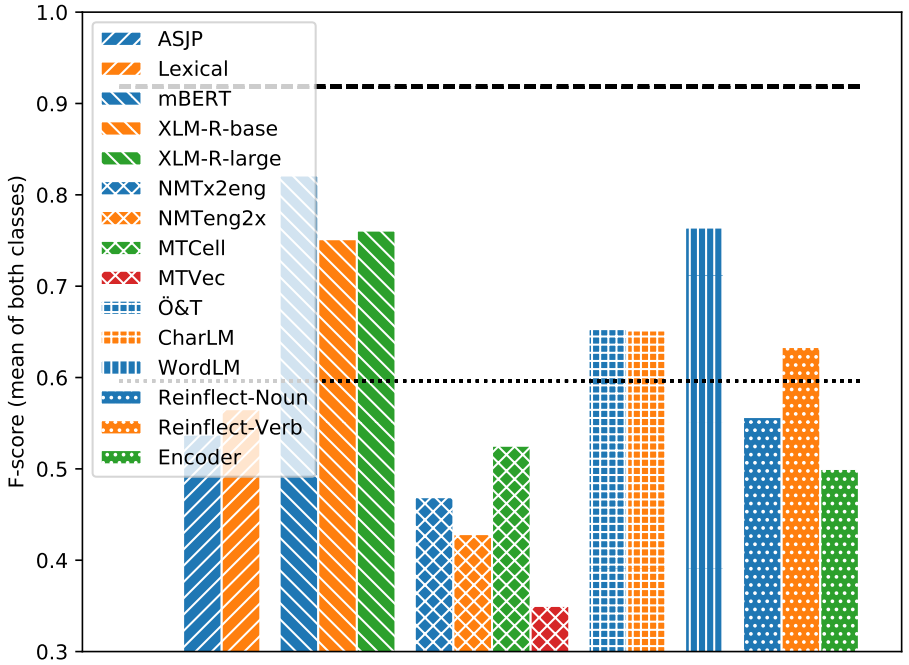


(b) Prepositions vs. postpositions, using projected labels for training.

Figure 2
Classification results for each set of language representations.

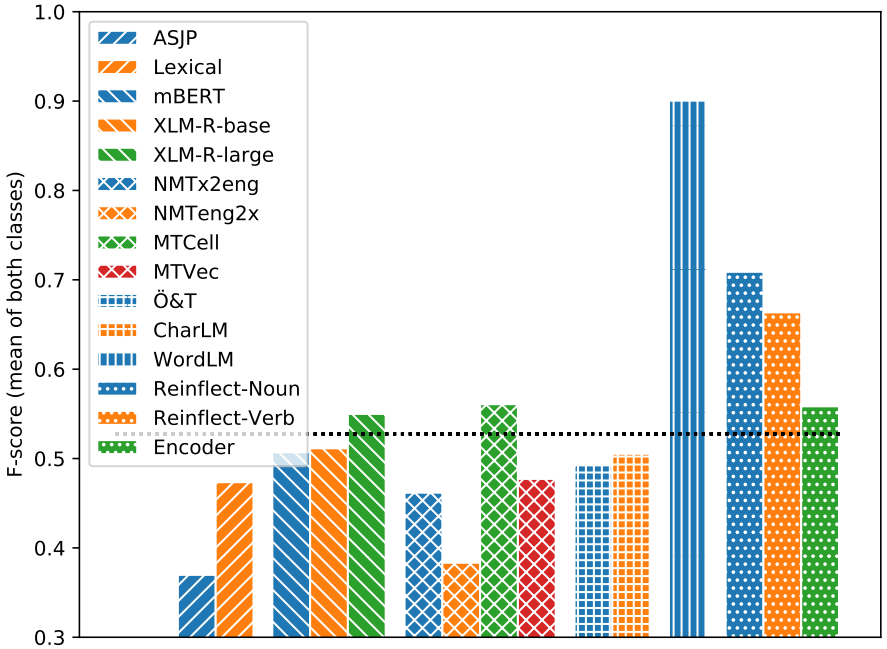


(a) Order of numeral and noun, using gold standard labels for training.

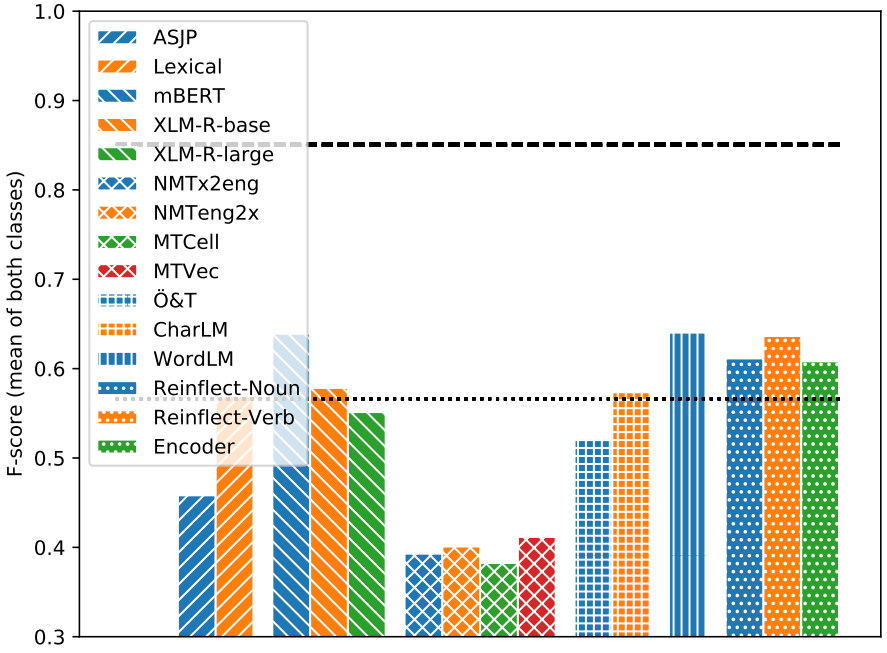


(b) Order of numeral and noun, using projected labels for training.

Figure 3
Classification results for each set of language representations.

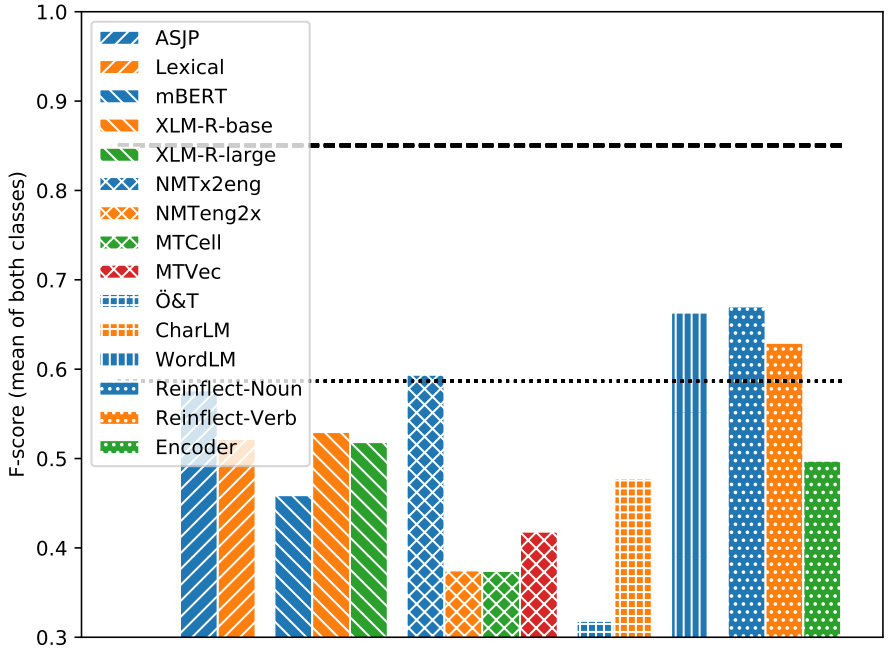


(a) Order of possessor and noun, using gold standard labels for training.

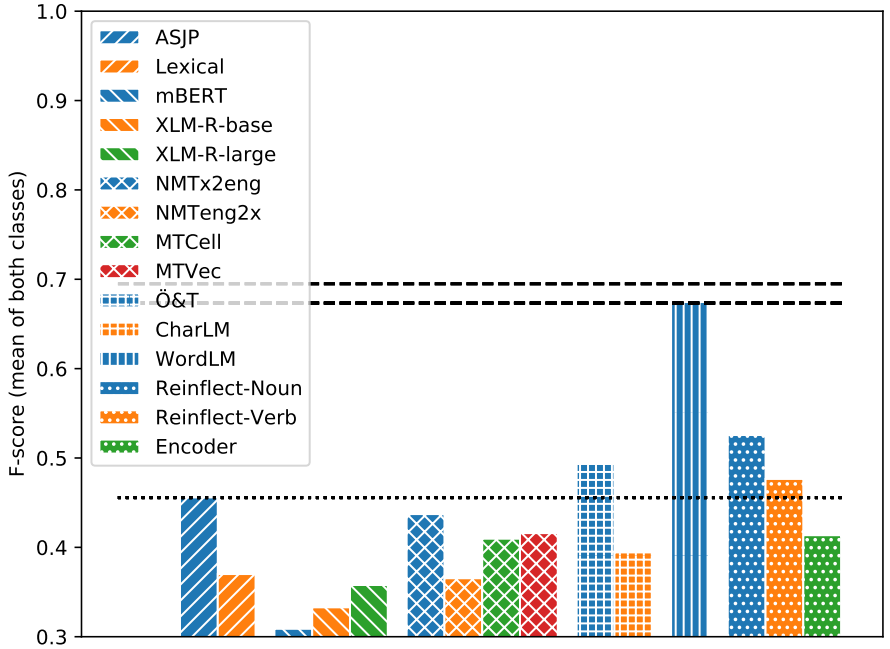


(b) Order of adjective and noun, using gold standard labels for training. Version with projected labels is omitted, but very similar.

Figure 4
Classification results for each set of language representations.



(a) Order of relative clause and noun, using gold standard labels for training.



(b) Order of subject and verb, using gold standard labels for training. Version with projected labels is omitted, but very similar.

Figure 5
Classification results for each set of language representations.

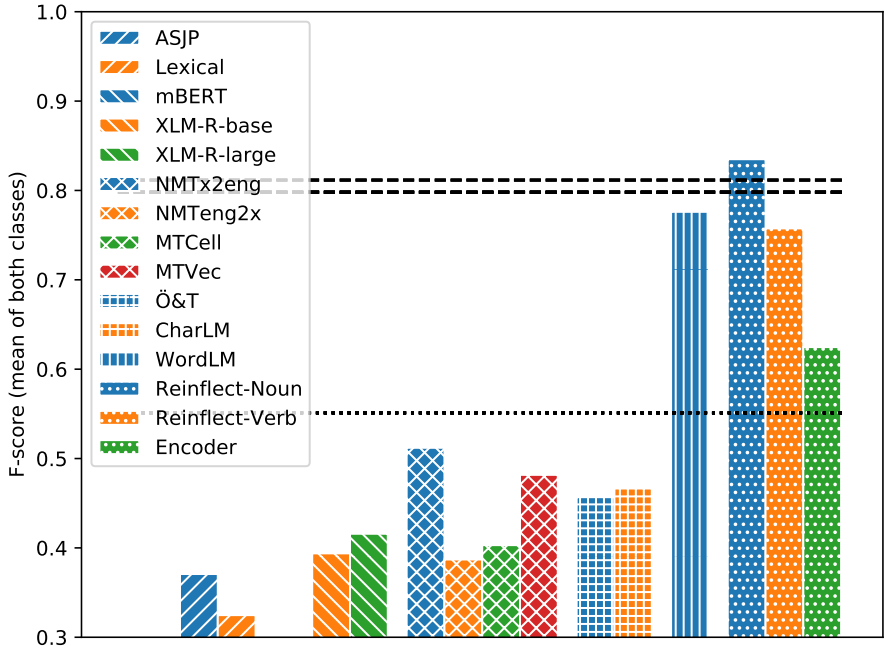
The order of adjective and noun can be accurately projected (mean F_1 of 0.850, see Table 2) but is not predictable with reasonable accuracy from even the WORDLM representations (Figure 4b). This also applies to the order of relative clause and noun, with a projection F_1 of 0.851 but poor classification results (F_1: 0.648). Classifiers trained on relative/noun order become most proficient (F_1: 0.881) at classifying adposition/noun order.

As can be seen in Table 2 and Figure 5b, the order of subject and verb is difficult to automatically extract through annotation projections in the data. The classification accuracy for the WORDLM representations on this feature is somewhat better (0.702) than the projection result (0.673). For reasons discussed in Section 8.4 below, we believe that this classifier has at least partly learned to identify subject/verb order.

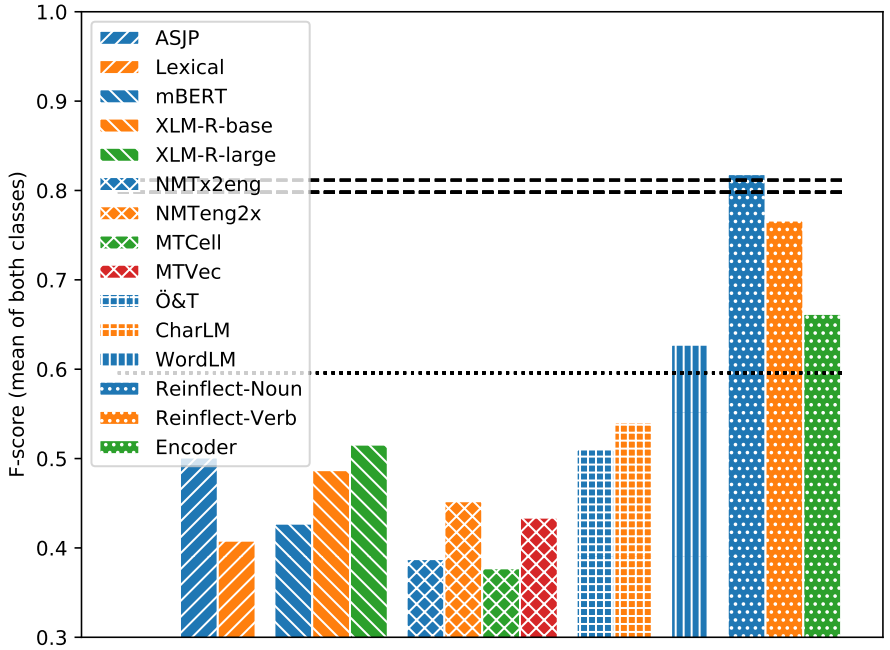
Apart from WORDLM and the REINFLECT models, none of the representations reach a mean F_1 of 0.7 for any of the features under investigation, with one intriguing exception. All three pretrained multilingual language models (MBERT, XLM-R-BASE, XLM-R-LARGE) obtain relatively high accuracy for the classification of numeral/noun order (Figure 3). This also happens to be the feature for which the character-based language model representations happen to achieve the highest accuracy. At a glance, these results seem odd, because the training data of the pretrained multilingual language models only cover a small minority of the language families present in the evaluation. The character-based language models, while trained on the full set of languages, instead have the problem of not being large enough to generalize to syntactic phenomena. We hypothesize these results are partly connected to the fact that simple text surface features are correlated with numeral/noun order. Because noun phrases are rarely broken up by punctuation, we can assume that languages where punctuation is commonly followed by digits use numeral–noun order, while digit–punctuation pairs should be more common in noun–numeral languages. Since only a minority of languages (125 out of 1,295) use digits to represent numerals in the New Testament, we do not expect this effect to be very strong. However, it highlights surface-level patterns as a potential source of errors in this type of research. Another potential explanation for the pretrained multilingual language models is the extreme bias toward numeral–noun order in their training data. Apart from some branches of Sino-Tibetan, noun–numeral order is very rare in Eurasia, and as far as we know only four languages in the multilingual BERT training data use this order (Malagasy, Swahili, Burmese, and Newari). On a global scale, noun–numeral order is however the most common, occurring in 608 out of 1,087 (56%) of the languages considered by Dryer (2013c) to have a dominant numeral/noun order. Thus, a fairly good heuristic is to assume that all languages recognized by the pretrained multilingual language models are numeral–noun languages, and the ones not recognized to use noun–numeral order.

8.2 Morphological Features

Figure 6 shows how well different language representations can be used to predict whether a language tends to use prefixes or suffixes (affixation type), according to the weighted affixation index of Dryer (2013f). Languages classified as not using affixation, or with equal use of prefixes and suffixes, are excluded from the sample. The language representations best able to predict this feature is the REINFLECT-NOUN, followed by REINFLECT-VERB and (when using gold-standard labels for training, Figure 6a) the WORDLM representations. However, with WORDLM representations, the object and verb order as well as adposition and noun order features both explain the classification results about equally well (F_1 within 1.5 percentage points). For the REINFLECT-



(a) Prefixing or suffixing in inflectional morphology, using gold standard labels for training.



(b) Prefixing or suffixing in inflectional morphology, using projected labels for training.

Figure 6
Classification results for each set of language representations.

VERB representations, the affixation type classification results can be explained by the negative affix position feature, which is not surprising given that it is included (along several other features) in the overall affixation position feature. The reinflection models have access only to word forms, without semantic or syntactic information, and so we do not expect them to differentiate between grammatical categories. In addition to overall prefixing/suffixing tendency, the following features related to morphology were examined:

- Whether case affixes are prefixes or suffixes
- Whether negative affixes are prefixes or suffixes
- Whether plural affixes are prefixes or suffixes
- Whether possessive affixes are prefixes or suffixes
- Whether TAM affixes are prefixes or suffixes
- Existence of a negative affix

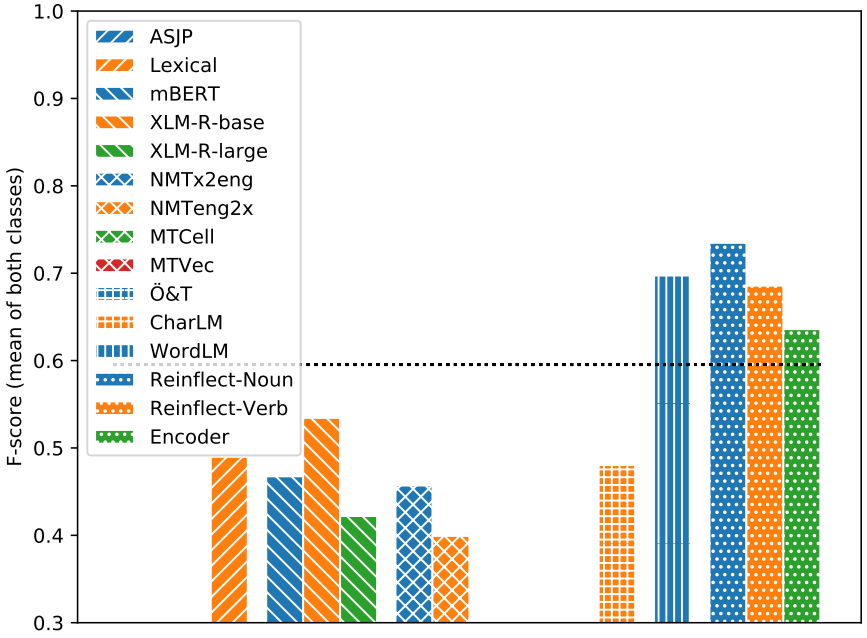
Some of them can be classified well using reinflection model representations, but they are all strongly correlated with each other and with the overall prefix/suffix feature, which is a weighted mean including most of the above features. This makes it difficult to conclusively determine which feature(s) a certain classifier has learned.

The ENCODER model does have access to both word form and semantics, in the form of projected word embeddings. In Figure 7a (whether negation is expressed with a prefix or a suffix) and Figure 7b (whether a possessive prefix or suffix is used), we see that this model does not seem to encode the position of these specific features any more clearly than the reinflection models, which likely only achieve high classification accuracy due to correlation with the position of other affixes in the same language. One reason for this failure to encode morphological information is that the model is faced with the difficult task of encoding the representations of 18 million vocabulary items. Unlike the reinflection models, the encoder model does not have the opportunity to copy information, but must store a mapping within its rather limited number of parameters (565,000). In future work, it may be worth investigating a model that predicts the word embeddings, rather than the form, given the embedding and form of another member of the same paradigm. Such a model could extract encoded lexical information directly from the source embedding, and could focus on identifying morphological information.

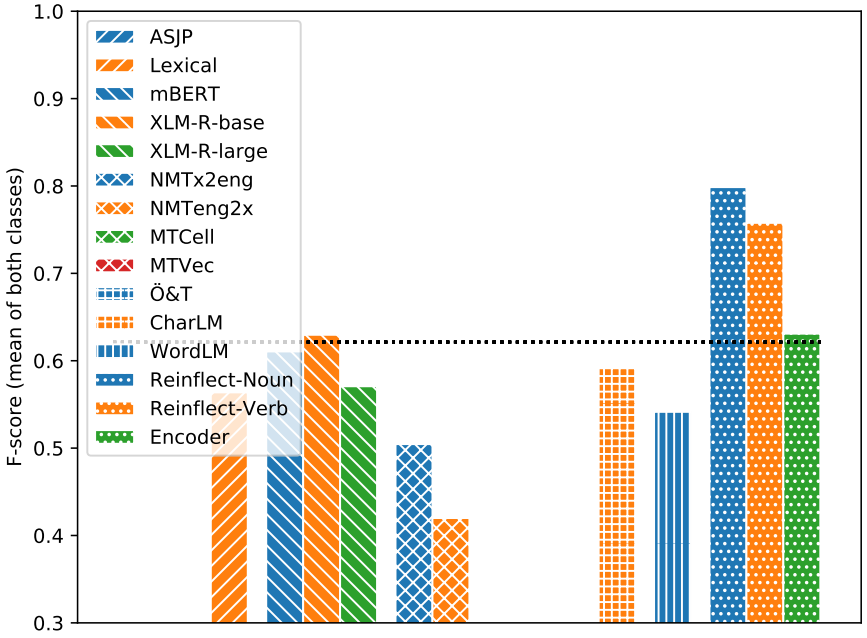
In summary, our reinflection models seem to encode the overall tendency toward prefixing or suffixing, while no models are able to single out the position of affixes for specific grammatical categories.

8.3 Naive Cross-validation Results

To illustrate the effect of not following our cross-validation set-up (Section 7.2), we now compare Figure 8a (naive cross-validation) with Figure 1a (linguistically sound cross-validation), and Figure 8b (naive) with Figure 4b (sound). Clear detections, such as object/verb order with the WORDLM representations, are not affected much by the cross-validation set-up and result in accurate classifiers in both cases. Language representations with baseline-level results, such as our NMT-based models (NMTENG2X

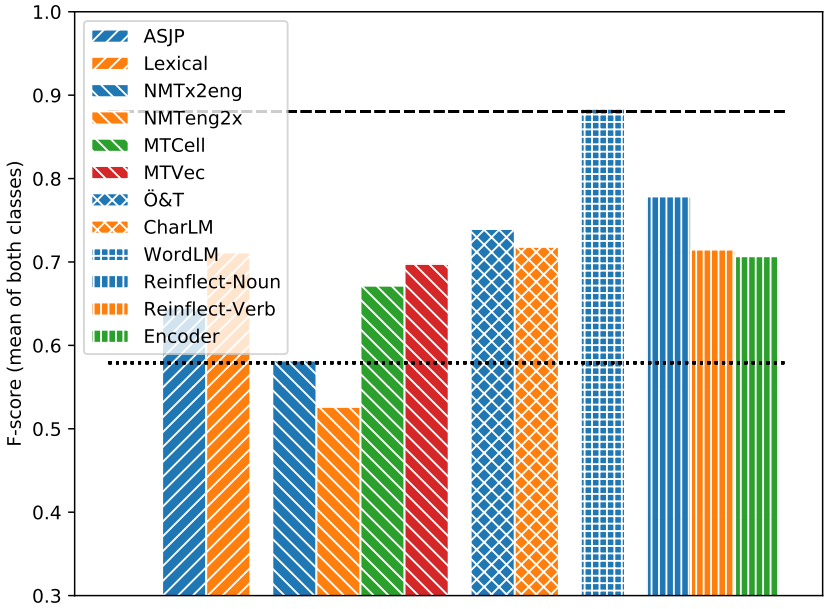


(a) Negative prefix or suffix, using gold standard labels for training.

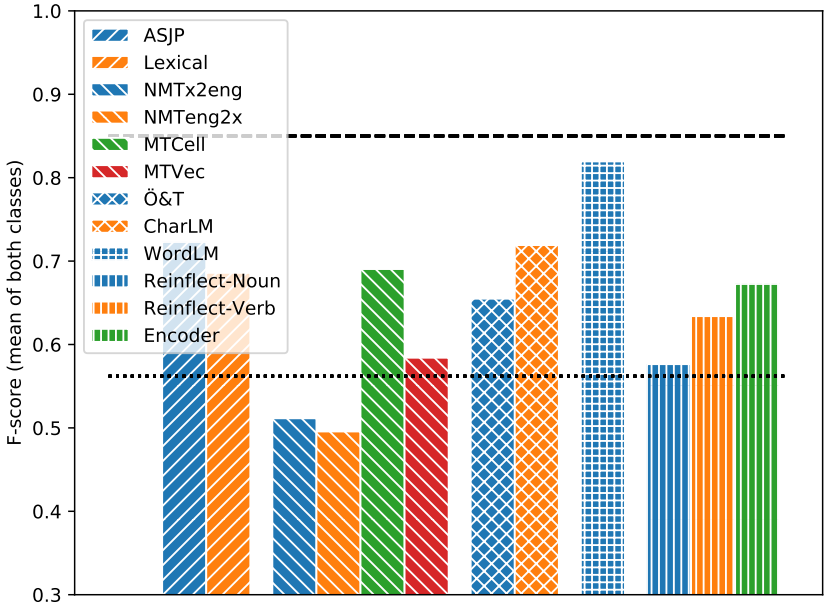


(b) Possessive prefix or suffix, using gold standard labels for training.

Figure 7
Classification results for each set of language representations. Note that some of the language representations contain too few languages in common with URIEL to be evaluated; the corresponding bars are omitted from the figures.



(a) Order of object and verb, using gold standard labels for training and naive cross-validation.



(b) Order of adjective and noun, using gold standard labels for training and naive cross-validation.

Figure 8
Classification results for each set of language representations, using naive cross-validation where languages related to the evaluated language are not excluded from the training fold. The point of this figure is to demonstrate how unsound evaluation methods give misleading results; see main text for details.

and NMTX2ENG), perform equally poorly in both cases, suggesting that they do not correlate well with any type of language similarity. For representations such as LEXICAL and ASJP, the naive cross-validation set-up results in much higher classification F_1 than the linguistically sound cross-validation. This is expected, since previous research has shown that the similarity metrics used to create these language representations can be used to reconstruct genealogical trees (Wichmann, Holman, and Brown 2018), which correlate well with typological features. The character-based language models (Ö&T and CHARLM) also show a similar increase in classification accuracy when naive cross-validation is used, which may indicate that they too use their language embeddings mainly to encode lexical similarity.

8.4 Analysis of Disagreements

For most classification experiments, we use URIEL data as a gold standard for both training and evaluation. However, for a few features we have access to projected labels. Here we apply these both as labels for training our classifiers, and as an additional source of information when analyzing the predictions of the classifiers we train.

To begin with, we compare the results when using URIEL labels for training (Figure 1a) with using projected labels (Figure 1b). The overall results are very similar, which indicates that the projected labels are useful for learning this feature, even though they diverge somewhat from the gold standard URIEL labels.

For a more detailed view of the results, we show 3-way confusion matrices for a number of features in Table 3, summarizing the three sets of labels we have:

1. Gold-standard URIEL labels (upper/lower matrix), index *i*

Table 3
3-way confusion matrices. We denote these matrices as $M_{i,j,k}$, where the sub-matrix *i* indicates the URIEL label, row *j* the projected label, and column *k* the classifier output. These all refer to the *evaluation* label. The header indicates whether URIEL or projected labels were used for *training*. All numbers are percentages of language families with a certain combination of labels. Language families with more than one doculect in the data contribute to multiple counts, but each family has equal total weight.

OV/VO URIEL	OV/VO projected	AdpN/NAdp URIEL	AdpN/NAdp projected
$\begin{pmatrix}\begin{pmatrix}54.6 & 0.4\\ 9.9 & 0.2\end{pmatrix}\\ \begin{pmatrix}1.3 & 0\\ 7.3 & 26.4\end{pmatrix}\end{pmatrix}$	$\begin{pmatrix}\begin{pmatrix}53.5 & 1.5\\ 8.7 & 1.4\end{pmatrix}\\ \begin{pmatrix}1.3 & 0\\ 3.9 & 29.8\end{pmatrix}\end{pmatrix}$	$\begin{pmatrix}\begin{pmatrix}35.8 & 5.4\\ 0.1 & 0.0\end{pmatrix}\\ \begin{pmatrix}0.0 & 1.8\\ 5.5 & 51.3\end{pmatrix}\end{pmatrix}$	$\begin{pmatrix}\begin{pmatrix}37.5 & 4\\ 0.1 & 0.0\end{pmatrix}\\ \begin{pmatrix}0.0 & 1.8\\ 5.8 & 51.1\end{pmatrix}\end{pmatrix}$
RelN/NRel URIEL	**NumN/NNum URIEL**	**AdjN/NAdj URIEL**	**SV/VS URIEL**
$\begin{pmatrix}\begin{pmatrix}15.2 & 0.1\\ 9.5 & 0.0\end{pmatrix}\\ \begin{pmatrix}0.0 & 0.0\\ 29.6 & 45.5\end{pmatrix}\end{pmatrix}$	$\begin{pmatrix}\begin{pmatrix}44.4 & 10.8\\ 0.7 & 2.2\end{pmatrix}\\ \begin{pmatrix}1.6 & 3.4\\ 8.5 & 28.3\end{pmatrix}\end{pmatrix}$	$\begin{pmatrix}\begin{pmatrix}29.0 & 4.9\\ 1.9 & 1.4\end{pmatrix}\\ \begin{pmatrix}8.7 & 2.5\\ 20.8 & 30.7\end{pmatrix}\end{pmatrix}$	$\begin{pmatrix}\begin{pmatrix}75.1 & 14.7\\ 0.0 & 1.1\end{pmatrix}\\ \begin{pmatrix}0.4 & 6.1\\ 0.2 & 2.2\end{pmatrix}\end{pmatrix}$

2. Projected label (row), index j
3. Predicted label from classifier (column), index k

To begin with, we can compare the matrices obtained for WORDLM when training on URIEL labels ($M^{\text{URIEL (OV/VO)}}$, top left in Table 3) and with projected labels ($M^{\text{projected (OV/VO)}}$, second from left). If disagreements between the language representation-based classifiers and the typological databases were mainly due to differences between the Bible doculects and those used by the WALS and Ethonogue database compilers, we would have expected a much higher agreement between projected and classified labels. On the contrary, the mean F_1 is actually somewhat lower when evaluated against projected labels, even when projected labels are used for training (mean F_1 is 0.851, compared to 0.910 when evaluated against URIEL).

The same pattern is present for another feature, order of adposition and noun (Figure 2), with confusion matrices in Table 3. The mean F_1 with respect to projected labels is nearly identical with URIEL-trained classifiers (0.887) as with classifiers trained on projected labels (0.869). We see occasional examples of the opposite case, where the mean F_1 is somewhat higher when evaluated against the projected labels, but our conclusion is that actual linguistic differences between the Bible corpus and URIEL do not alone explain the cases where our classifiers differ from the URIEL classifications.

A somewhat different result is shown in Figure 3 and Table 3 for the order of numeral and noun. Here, the mean F_1 is considerably higher (0.763) when trained on projected labels than on URIEL labels (0.684), where both figures are evaluated with respect to URIEL labels. This could be partly due to the fact that the projected labels are available for more languages, and the mean number of language families for each training fold is higher (101.1) for the projected labels than for URIEL labels (60.9). Recall that only one randomly sampled doculect per family is represented in each training fold, so the number of families corresponds to the number of training fold data points. The mean F_1 is not substantially different (difference is less than one percentage point) when evaluated on projected instead of URIEL labels, and this applies for both sets of training labels, which speaks against the hypothesis that the URIEL and projected labels represent substantially different interpretations of the feature.

One notable property of the confusion matrices in Table 3 is that $M_{0,1,1}$ and $M_{1,0,0}$ are generally very low, which means that when the projected feature value agrees with the classifier prediction, this consensus is very often correct according to URIEL. To quantify this, we can compute the F_1 for the subset of data where projected features and classifier predictions agree. Table 4 shows how the F_1 of WORDLM increases drastically when we evaluate on this subset alone, sometimes reaching perfect or near-perfect scores. This subset corresponds to the rows in Table 3 where $i = j$, covering the vast majority of language families.

The only apparent disagreement for the order of adposition and noun turns out to be an error in URIEL.[17] For the order of object and verb, URIEL disagrees in five cases: Mbyá Guaraní (Tupian), Purépecha (isolate), Koreguaje (Tucanoan), Luwo (Nilotic), Yine (Arawakan). We have located grammatical descriptions in languages readable to us for three of these, in addition to quantitative word order data for Mbyá Guaraní.

17 Strangely, URIEL codes Serbian as having postpositions, even though Dryer (2013b) correctly codes it as prepositional.

Table 4
Family-weighted mean F_1 scores of classifiers trained using WORDLM representations. The columns give values using **All doculects**, or only those doculects where the projected and the classifier-predicted value agrees (**Projected = Predicted**).

	Mean F_1 score	
Feature	**All doculects**	**Projected = Predicted**
Order of adjective and noun	0.639	0.880
Order of numeral and noun	0.762	0.947
Order of relative clause and noun	0.648	0.999
Order of adposition and noun	0.866	1.000
Order of object and verb	0.896	0.980
Order of subject and verb	0.702	0.865

Choi et al. (2021) compare basic word order obtained from Universal Dependencies corpora (Nivre et al. 2018) with those in WALS (Dryer and Haspelmath 2013) and Östling (2015), and question the classification of Mbyá Guaraní as SVO-dominant since SOV is nearly as common.[18]

Yine is classified by Ethnologue as an SOV language while our classification and projection methods both show a tendency toward VO order. Hanson (2010, page 292) states that "The relative order of predicate and arguments varies considerably under pragmatic and stylistic motivations [...] The predicate-first order is somewhat more common than argument-first in verbal clauses."

For Purépecha, Dryer (2013d) has SVO order. Friedrich (1984, pages 61–62) gives SOV order but adds that "the object–verb rule is weak" and further specifies that "Short objects and, often, pronominal ones are generally preverbal. [...] Objects with two or more words, especially long words, tend to be placed after the verb." There is no attempt at quantifying these statements.

From these examples, we see that when classifications from WALS or Ethnologue disagree with a classifier/projection consensus with regard to verb/object order, in all cases we have investigated this can be attributed to the languages having a flexible word order, where the identification of a single dominant word order can be called into question.

Our interpretation of the generally high agreement when the classifier and projections agree is that these two methods, at least for our WORDLM embeddings, complement each other. When both of them agree it is likely that the language is a clear example of the feature in question, and thus also likely to be classified as such by the database compilers. It is notable that we do *not* see a corresponding improvement of classification performance in the subset of languages where URIEL and the projections agree, which again indicates the observed divergences cannot only be explained by widespread grammatical differences between Bible doculects and URIEL sources.

In a few cases we observe the effects of different definitions of particular word order properties. The main exception to the pattern of high agreement between projected/classified consensus and URIEL classifications can be found for adjective/noun order, where 8.7% of families are classified as adjective–noun by both the projection

18 Choi et al. (2021) in fact compared Mbyá Guaraní with Paraguayan Guaraní (personal communication), which is coded as SVO by Dryer (2013e), citing Gregores and Suárez (1967, page 182) who describe Paraguayan Guaraní as having a rather free word order with SVO order being the most common, although they note that statements on word order should be taken as "very rough approximations, based on impressionistic evaluations of what is more frequent."

approach and the classifier, but are noun–adjective according to Dryer (2013a). In this group we find several Romance languages. As discussed earlier, these tend to use adjective–noun order for a set of very common core adjectives, whereas noun–adjective is more productive but may be less common on a token level. For several other language families we also find examples where the order between *core* adjective concepts and nouns differs from the order between Universal Dependencies ADJ-tagged words and nouns. However, a more careful analysis would be required to determine the cause of this discrepancy.

For the order of relative clause and noun, we see that the classifier has mediocre performance for the full sample but is near-perfect in the subset where projected and predicted labels agree. Looking at the full confusion matrix in Table 3, we see that the classifier is very good at classifying relative–noun languages, while the projection method instead excels at classifying noun–relative languages. This is mainly driven by the 29.6% of language families that are classified as noun–relative order by both URIEL and the projection method, while the classifier gives relative–noun order. The features that best explain (in terms of highest mean F_1) the classifications of the relative/noun classifier, are adposition/noun, possessor/noun and object/verb order. This is not surprising, since relative–noun languages are overwhelmingly postpositional, object–verb and possessor–noun. If the classifier has learned to use one or more of these features as a proxy for relative/noun order, we would expect the languages misclassified as relative–noun to also be mainly postpositional, object–verb and possessor–noun. This is precisely what we find, whereas languages correctly classified as noun–relative are overwhelmingly prepositional, verb–object and noun–possessor. In combination with high accuracy of the projection method for noun–relative order, this causes the classifier/projection consensus to be in nearly perfect agreement with Dryer (2013a) but partly due to reasons not directly related to relative clauses.

8.5 Evaluation Methodology

Malaviya, Neubig, and Littell (2017) reported identifying features of syntax and phonology in the language representations from a multilingual NMT system, and Bjerva and Augenstein (2018a) found features of syntax, morphology, and phonology in the language representations from the mulitilingual language model of Östling and Tiedemann (2017). Both relied on typological feature classification experiments. When strict separation of related languages between training and testing folds in the cross-validation is enforced, only a few solid identifications of typological features stand out, and these all come from our new models. Both Malaviya, Neubig, and Littell (2017) and Bjerva and Augenstein (2018a) did take some precautions to avoid correlations between features of close languages affecting their results. However, even though the precise cause for the discrepancy between our respective conclusions have not been conclusively determined, we believe that our identification of typological generalizations by neural models is much more robust and unambiguous than in previous work. In some cases, the accuracy obtained by our classifiers even exceeds that of hand-coded annotation projection. This makes us able to not only demonstrate that neural models can discover typological features, but also that they can be used in practice to classify languages according to those features. When combining the results of the language representation-trained classifier and our projection method, the agreement with manually coded features can be even further increased. In part we believe this is due to the methods being complementary. Our word-based language model uses projected word embeddings and cosine loss in order to train efficiently with the full 18 million word vocabulary of

all 1,295 languages, and is not limited by the Universal Dependencies annotations that our projection method relies on.

Perhaps the most important result of our work is that typological generalizations *can* be discovered by neural models solving NLP tasks, but only under certain circumstances. For word order features, the language representations from our multilingual word-based language model (WORDLM) result in highly accurate classifiers for a range of word order features, close to the accuracy of various hand-crafted approaches in previous work (Figure 9 in Ponti et al. 2019) as well as our projection-based approach (Section 5.6). The general tendency of languages to be prefixing or suffixing does also appear to be discovered by our reinflection models.

Apart from these examples, we do not find any clear evidence of typological features encoded in the 12 sets of language representations we investigated. In most cases classification results were consistent with random labels. In some cases, such as the WORDLM model being able to distinguish prefixing languages from suffixing, we show that the results can be better explained by the classifier learning a *different* but correlated typological parameter.

Through the representations from the word-level language model and reinflection models, as well as our features obtained through annotation projection, we establish estimates for how well a number of typological features can be extracted from our data. No other language representations, including those from previous work, even come close to this level. From this we conclude that the models have not encoded any of the syntactic or morphological features in our study, nor language features sufficiently correlated with the features studied to create an accurate classifier for any of them. It would be theoretically possible that some of the features are encoded in some language representations, but in a way not classifiable using a logistic regression classifier. This would however be difficult to verify, and our results show that at least the word-level language model and reinflection models do encode features that are identifiable by a linear classifier.

Several previous authors have showed that vector similarity between some set of language representations has a similar structure to traditional phylogenetic trees constructed by historical linguists (Östling and Tiedemann 2017; Oncevay, Haddow, and Birch 2020; Tan et al. 2019), or more generally cluster along family lines (Tiedemann 2018; He and Sagae 2019). While these observations are correct and can be of practical value in an NLP setting, they do not reveal much about whether linguistic *generalizations* are made by the model and encoded in the language representations.

Classification-based evaluations can be used to probe directly whether certain features are encoded in a set of language representations, assuming that correlations with genealogically and geographically close languages are properly controlled for. In Section 8.3, we showed that if care is not taken to make the testing set of each classifier model as independent as possible of the training set, it is very easy to obtain spurious results.

9. Conclusions

We expect that two types of readers will benefit from our work: those working with highly multilingual NLP applications, and those interested in using automatic means for studying the diversity of human languages.

From the NLP practitioner's point of view, we expect that some of our published data will be of particular interest. For instance, our artificially produced paradigms could be used as pre-training for morphological inflection models where annotated

data is sparse. There is also evidence that language representations can be useful for guiding multilingual NLP systems including machine translation (Oncevay, Haddow, and Birch 2020) and dependency parsing (Üstün et al. 2020), and our set of language representations with different properties provides a rich collection of representations for future experiments.

Broadening the perspective, we show that if a neural machine learning system is given the right kind of task to perform on a very large set of languages, and given only a small number of parameters to summarize the differences between languages, it uses those parameters to encode some of the same types of features that human linguists have long studied. From the point of view of the typologist, our research has resulted in a large amount of fine-grained data on several features related to word order and affix position. Not only do we find the dominant patterns, but also the amount of variation within each language. Such information has been used in token-based typological studies of word order variability (Levshina 2019), but restricted to much smaller samples of languages than what we now publish. We provide two complementary methods to obtain this information: annotation projection, and regression models from our learned language representations. As discussed in Section 8.4 and shown in Table 4, the agreement with typological databases is particularly high in the subset of languages where both of these sources point in the same direction. Importantly, unlike previous methods for typological feature prediction that utilize language relatedness and correlations between features for prediction (Murawaki 2019), we use entirely data-driven methods based on raw text data in each language, and are thus better positioned for finding *unexpected* properties of languages.

As an important direction of future work, we see that the granularity of the predicted features could be reduced. Due to lack of suitable token-level gold standard data for most of our language samples, we have been restricted to binary feature classifiers in this work (although token-based counts are still available from the projection method). In addition to a move toward fully token-based typology, we also see the need for investigating individual constructions and specific factors triggering variation within languages. For instance, our current work does not differentiate between word order in main clauses and in subordinate clauses. Achieving this level of detail for a geographically and typologically diverse sample of over a thousand languages would be a very valuable tool for typological research.

Acknowledgments

Thanks to Bernhard Wälchli, Mats Wirén, Dmitry Nikolaev, and our anonymous reviewers for valuable comments on this manuscript at different stages. This work was funded by the Swedish Research Council (2019-04129) and in part by the Swedish national research infrastructure Språkbanken and Swe-Clarin, funded jointly by the Swedish Research Council (2017-00626) and the 10 participating partner institutions. The computations were enabled by resources provided by the Swedish National Infrastructure for Computing (SNIC) at C3SE partially funded by the Swedish Research Council (2018-05973).

References

Ammar, Waleed, George Mulcaire, Miguel Ballesteros, Chris Dyer, and Noah Smith. 2016. Many languages, one parser. *Transactions of the Association for Computational Linguistics*, 4(1):431–444. https://doi.org/10.1162/tacl_a_00109

Artetxe, Mikel and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7(0):597–610. https://doi.org/10.1162/tacl_a_00288

Asgari, Ehsaneddin and Hinrich Schütze. 2017. Past, present, future: A computational investigation of the

typology of tense in 1000 languages. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 113–124. https://doi.org/10.18653/v1/D17-1011

Bakker, Dik, André Müller, Viveka Velupillai, Søren Wichmann, Cecil H. Brown, Pamela Brown, Dmitry Egorov, Robert Mailhammer, Anthony Grant, and Eric W. Holman. 2009. Adding typology to lexicostatistics: A combined approach to language classification. *Linguistic Typology*, 13(1):169–181. https://doi.org/10.1515/LITY.2009.009

Beinborn, Lisa and Rochelle Choenni. 2020. Semantic drift in multilingual representations. *Computational Linguistics*, 46(3):571–603. https://doi.org/10.1162/coli_a_00382

Bickel, Balthasar, Johanna Nichols, Taras Zakharko, Kristine Hildebrandt Alena Witzlack-Makarevich, Michael Rießler, Lennart Bierkandt, Fernando Zúñig, and John B. Lowe. 2017. The AUTOTYP typological databases. Version 0.1.0. https://github.com/autotyp/autotyp-data/tree/0.1.0

Bjerva, Johannes and Isabelle Augenstein. 2018a. From phonology to syntax: Unsupervised linguistic typology at different levels with language embeddings. 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 907–916. https://doi.org/10.18653/v1/N18-1083

Bjerva, Johannes and Isabelle Augenstein. 2018b. Tracking typological traits of Uralic languages in distributed language representations. In *Proceedings of the Fourth International Workshop on Computational Linguistics of Uralic Languages*, pages 76–86. https://doi.org/10.18653/v1/W18-0207

Bjerva, Johannes 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. https://doi.org/10.18653/v1/2021.eacl-main.38

Bjerva, Johannes, Yova Kementchedjhieva, Ryan Cotterell, and Isabelle Augenstein. 2019. A probabilistic generative model of linguistic typology. 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 1529–1540. https://doi.org/10.18653/v1/N19-1156

Bjerva, Johannes, Elizabeth Salesky, Sabrina J. Mielke, Aditi Chaudhary, Giuseppe G. A. Celano, Edoardo Maria Ponti, Ekaterina Vylomova, Ryan Cotterell, and Isabelle Augenstein. 2020. SIGTYP 2020 shared task: Prediction of typological features. In *Proceedings of the Second Workshop on Computational Research in Linguistic Typology*, pages 1–11. https://doi.org/10.18653/v1/2020.sigtyp-1.1

Blasi, Damian, Antonios Anastasopoulos, and Graham Neubig. 2022. Systematic inequalities in language technology performance across the world's languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5486–5505. https://doi.org/10.18653/v1/2022.acl-long.376

Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, and Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901, Curran Associates, Inc.

Chi, Ethan A., 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. https://doi.org/10.18653/v1/2020.acl-main.493

Choenni, Rochelle and Ekaterina Shutova. 2022. Investigating language relationships in multilingual sentence encoders through the lens of linguistic typology. *Computational Linguistics*, 48(3):635–672. https://doi.org/10.1162/coli_a_00444

Choi, Hee Soo, Bruno Guillaume, Karën Fort, and Guy Perrier. 2021. Investigating dominant word order on Universal Dependencies with graph rewriting. In *Proceedings of the International Conference on*

Recent Advances in Natural Language Processing (RANLP 2021), pages 281–290. https://doi.org/10.26615/978-954-452-072-4_033

Chowdhery, Aakanksha, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. PaLM: Scaling language modeling with Pathways. *arXiv preprint arXiv:2204.02311v5*.

Comrie, Bernard. 2013. Numeral bases (v2020.3). In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Zenodo.

Conneau, Alexis, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451. https://doi.org/10.18653/v1/2020.acl-main.747

Conneau, Alexis and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*, volume 32, Curran Associates, Inc.

Conneau, Alexis, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. *arXiv preprint arXiv:1710.04087v3*.

Cotterell, Ryan, Christo Kirov, John Sylak-Glassman, David Yarowsky, Jason Eisner, and Mans Hulden. 2016. The SIGMORPHON 2016 shared Task—Morphological reinflection. In *Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 10–22. https://doi.org/10.18653/v1/W16-2002

Croft, William. 2002. *Typology and Universals*, 2nd edition. Cambridge Textbooks in Linguistics. Cambridge University Press. https://doi.org/10.1017/CBO9780511840579

Cysouw, Michael and Jeff Good. 2013. Languoid, doculect and glossonym: Formalizing the notion 'language'. *Language Documentation and Conservation*, 7:331–359.

Dahl, Östen. 2007. From questionnaires to parallel corpora in typology. *STUF - Language Typology and Universals*, 60(2):172–181. https://doi.org/10.1524/stuf.2007.60.2.172

Devlin, Jacob, 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.

Dixon, Robert M. W. 1982. *Where have all the adjectives gone? and other essays in semantics and syntax*. Mouton, New York. https://doi.org/10.1515/9783110822939

Dryer, Matthew S. 2013a. Order of adjective and noun (v2020.3). In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Zenodo.

Dryer, Matthew S. 2013b. Order of adposition and noun phrase (v2020.3). In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Zenodo.

Dryer, Matthew S. 2013c. Order of numeral and noun (v2020.3). In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Zenodo.

Dryer, Matthew S. 2013d. Order of object and verb (v2020.3). In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Zenodo.

Dryer, Matthew S. 2013e. Order of subject, object and verb (v2020.3). In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Zenodo.

Dryer, Matthew S. 2013f. Prefixing vs. suffixing in inflectional morphology (v2020.3). In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Zenodo.

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

Dunn, Michael, Simon J. Greenhill, Stephen C. Levinson, and Russell D. Gray. 2011. Evolved structure of language shows lineage-specific trends in word-order universals. *Nature*, 473:79–82. https://doi.org/10.1038/nature09923, PubMed: 21490599

Eberhard, David M., Gary F. Simons, and Charles D. Fennig, editors. 2019. *Ethnologue: Languages of the World*, 22nd edition. SIL International, Dallas, Texas.

Ebrahimi, Abteen, Manuel Mager, Arturo Oncevay, Vishrav Chaudhary, Luis Chiruzzo, Angela Fan, John Ortega, Ricardo Ramos, Annette Rios, Ivan Vladimir Meza Ruiz, Gustavo Giménez-Lugo, Elisabeth Mager, Graham Neubig, Alexis Palmer, Rolando Coto-Solano, Thang Vu, and Katharina Kann. 2022. AmericasNLI: Evaluating zero-shot natural language understanding of pretrained multilingual models in truly low-resource languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6279–6299. https://doi.org/10.18653/v1/2022.acl-long.435

Friedrich, Paul. 1984. Tarascan: From meaning to sound. In Munro S. Edmonson, editor, *Supplement to the Handbook of Middle American Indians. Volume 2: Linguistics*. University of Texas Press, Austin, pages 56–82. https://doi.org/10.7560/775770-006

Greenberg, Joseph H. 1963. Some universals of grammar with particular reference to the order of meaningful elements. In Joseph H. Greenberg, editor, *Universals of Human Language*. MIT Press, Cambridge, Massachusetts, pages 73–113.

Gregores, Emma and Jorge A. Suárez. 1967. *A Description of Colloquial Guarani*. De Gruyter Mouton, Berlin, Boston. https://doi.org/10.1515/9783111349633

Grossman, Eitan, Elad Eisen, Dmitry Nikolaev, and Steven Moran. 2020. SegBo: A database of borrowed sounds in the world's languages. In *Proceedings of the Language Resources and Evaluation Conference (LREC 2020)*.

Hammarström, Harald. 2021. Measuring prefixation and suffixation in the languages of the world. In *Proceedings of the Third Workshop on Computational Typology and Multilingual NLP*, pages 81–89. https://doi.org/10.18653/v1/2021.sigtyp-1.8

Hammarström, Harald and Mark Donohue. 2014. Some principles on the use of macro-areas in typological comparison. *Language Dynamics and Change*, 4(1):167–187. https://doi.org/10.1163/22105832-00401001

Hammarström, Harald, Robert Forkel, Martin Haspelmath, and Sebastian Bank. 2022. *glottolog/glottolog: Glottolog database 4.7*. Zenodo.

Hanson, Rebecca. 2010. *A grammar of Yine (Piro)*. Ph.D. thesis, La Trobe University.

He, Taiqi and Kenji Sagae. 2019. Syntactic typology from plain text using language embeddings. In *Proceedings of the First Workshop on Typology for Polyglot NLP*.

Hewitt, John 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.

Hwa, Rebecca, Philip Resnik, Amy Weinberg, Clara Cabezas, and Okan Kolak. 2005. Bootstrapping parsers via syntactic projection across parallel texts. *Natural Language Engineering*, 11(3):311–325. https://doi.org/10.1017/S1351324905003840

Kanerva, Jenna, Filip Ginter, Niko Miekka, Akseli Leino, and Tapio Salakoski. 2018. Turku neural parser pipeline: An end-to-end system for the CoNLL 2018 shared task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 133–142.

Kann, Amanda. 2019. Ordföljdsvariation inom kardinaltalssystem: Extraktion av ordföljdstypologi ur parallella texter [Word order variation within cardinal number systems: Extraction of word order typology from parallel texts]. B.A. thesis, Department of Linguistics, Stockholm University.

Key, Mary Ritchie and Bernard Comrie, editors. 2015. *IDS*. Max Planck Institute for Evolutionary Anthropology, Leipzig. https://ids.clld.org/

Klein, Guillaume, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72.

Kudugunta, Sneha Reddy, Ankur Bapna, Isaac Caswell, Naveen Arivazhagan, and Orhan Firat. 2019. Investigating multilingual NMT representations at scale. *arXiv preprint arXiv:1909.02197v2*. https://doi.org/10.18653/v1/D19-1167

Levenshtein, Vladimir I. 1966. Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, 10:707.

Levshina, Natalia. 2019. Token-based typology and word order entropy: A study based on universal dependencies. *Linguistic Typology*, 23(3):533–572. https://doi.org/10.1515/lingty-2019-0025

List, Johann Mattis, Annika Tjuka, Christoph Rzymski, Simon Greenhill, and Robert Forkel, editors. 2022. *CLLD Concepticon 3.0.0*. Max Planck Institute for Evolutionary Anthropology, Leipzig. https://concepticon.clld.org/

Littell, Patrick, 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. https://doi.org/10.18653/v1/E17-2002

Malaviya, Chaitanya, Graham Neubig, and Patrick Littell. 2017. Learning language representations for typology prediction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2529–2535. https://doi.org/10.18653/v1/D17-1268

Mayer, Thomas and Michael Cysouw. 2014. Creating a massively parallel Bible corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3158–3163.

McDonald, Ryan, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, Claudia Bedini, Núria Bertomeu Castelló, and Jungmee Lee. 2013. Universal dependency annotation for multilingual parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 92–97.

McInnes, Leland, John Healy, Nathaniel Saul, and Lukas Grossberger. 2018. UMAP: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29):861. https://doi.org/10.21105/joss.00861

Moran, Steven and Daniel McCloy. editors. 2019. *PHOIBLE 2.0*. Max Planck Institute for the Science of Human History, Jena. https://phoible.org/

Murawaki, Yugo. 2019. Bayesian learning of latent representations of language structures. *Computational Linguistics*, 45(2):199–228. https://doi.org/10.1162/coli_a_00346

Naseem, Tahira, 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.

Nivre, Joakim, Mitchell Abrams, Željko Agić, Lars Ahrenberg, Lene Antonsen, Maria Jesus Aranzabe, Gashaw Arutie, Masayuki Asahara, Luma Ateyah, Mohammed Attia, Aitziber Atutxa, Liesbeth Augustinus, Elena Badmaeva, Miguel Ballesteros, Esha Banerjee, Sebastian Bank, Verginica Barbu Mititelu, John Bauer, Sandra Bellato, Kepa Bengoetxea, Riyaz Ahmad Bhat, Erica Biagetti, Eckhard Bick, Rogier Blokland, Victoria Bobicev, Carl Börstell, Cristina Bosco, Gosse Bouma, Sam Bowman, Adriane Boyd, Aljoscha Burchardt, Marie Candito, Bernard Caron, Gauthier Caron, Gülşen Cebiroğlu Eryiğit, Giuseppe G. A. Celano, Savas Cetin, Fabricio Chalub, Jinho Choi, Yongseok Cho, Jayeol Chun, Silvie Cinková, Aurélie Collomb, Çağrı Çöltekin, Miriam Connor, Marine Courtin, Elizabeth Davidson, Marie-Catherine de Marneffe, Valeria de Paiva, Arantza Diaz de Ilarraza, Carly Dickerson, Peter Dirix, Kaja Dobrovoljc, Timothy Dozat, Kira Droganova, Puneet Dwivedi, Marhaba Eli, Ali Elkahky, Binyam Ephrem, Tomaž Erjavec, Aline Etienne, Richárd Farkas, Hector Fernandez Alcalde, Jennifer Foster, Cláudia Freitas, Katarína Gajdošová, Daniel Galbraith, Marcos Garcia, Moa Gärdenfors, Kim Gerdes, Filip Ginter, Iakes Goenaga, Koldo Gojenola, Memduh Gökırmak, Yoav Goldberg,

Xavier Gómez Guinovart, Berta Gonzáles Saavedra, Matias Grioni, Normunds Grūzītis, Bruno Guillaume, Céline Guillot-Barbance, Nizar Habash, Jan Hajič, Jan Hajič jr., Linh Hà Mỹ, Na-Rae Han, Kim Harris, Dag Haug, Barbora Hladká, Jaroslava Hlaváčová, Florinel Hociung, Petter Hohle, Jena Hwang, Radu Ion, Elena Irimia, Tomáš Jelínek, Anders Johannsen, Fredrik Jørgensen, Hüner Kaşıkara, Sylvain Kahane, Hiroshi Kanayama, Jenna Kanerva, Tolga Kayadelen, Václava Kettnerová, Jesse Kirchner, Natalia Kotsyba, Simon Krek, Sookyoung Kwak, Veronika Laippala, Lorenzo Lambertino, Tatiana Lando, Septina Dian Larasati, Alexei Lavrentiev, John Lee, Phuong Lê Hồng, Alessandro Lenci, Saran Lertpradit, Herman Leung, Cheuk Ying Li, Josie Li, Keying Li, KyungTae Lim, Nikola Ljubešić, Olga Loginova, Olga Lyashevskaya, Teresa Lynn, Vivien Macketanz, Aibek Makazhanov, Michael Mandl, Christopher Manning, Ruli Manurung, Cătălina Mărănduc, David Mareček, Katrin Marheinecke, Héctor Martínez Alonso, André Martins, Jan Mašek, Yuji Matsumoto, Ryan McDonald, Gustavo Mendonça, Niko Miekka, Anna Missilä, Cătălin Mititelu, Yusuke Miyao, Simonetta Montemagni, Amir More, Laura Moreno Romero, Shinsuke Mori, Bjartur Mortensen, Bohdan Moskalevskyi, Kadri Muischnek, Yugo Murawaki, Kaili Müürisep, Pinkey Nainwani, Juan Ignacio Navarro Horñiacek, Anna Nedoluzhko, Gunta Nešpore-Bērzkalne, Luong Nguyễn Thị, Huyền Nguyễn Thị Minh, Vitaly Nikolaev, Rattima Nitisaroj, Hanna Nurmi, Stina Ojala, Adédayọ̀ Olúòkun, Mai Omura, Petya Osenova, Robert Östling, Lilja Øvrelid, Niko Partanen, Elena Pascual, Marco Passarotti, Agnieszka Patejuk, Siyao Peng, Cenel-Augusto Perez, Guy Perrier, Slav Petrov, Jussi Piitulainen, Emily Pitler, Barbara Plank, Thierry Poibeau, Martin Popel, Lauma Pretkalniņa, Sophie Prévost, Prokopis Prokopidis, Adam Przepiórkowski, Tiina Puolakainen, Sampo Pyysalo, Andriela Rääbis, Alexandre Rademaker, Loganathan Ramasamy, Taraka Rama, Carlos Ramisch, Vinit Ravishankar, Livy Real, Siva Reddy, Georg Rehm, Michael Rießler, Larissa Rinaldi, Laura Rituma, Luisa Rocha, Mykhailo Romanenko, Rudolf Rosa, Davide Rovati, Valentin Roca, Olga Rudina, Shoval Sadde, Shadi Saleh, Tanja Samardžić, Stephanie Samson, Manuela Sanguinetti, Baiba Saulīte, Yanin Sawanakunanon, Nathan Schneider, Sebastian Schuster, Djamé Seddah, Wolfgang Seeker, Mojgan Seraji, Mo Shen, Atsuko Shimada, Muh Shohibussirri, Dmitry Sichinava, Natalia Silveira, Maria Simi, Radu Simionescu, Katalin Simkó, Mária Šimková, Kiril Simov, Aaron Smith, Isabela Soares-Bastos, Antonio Stella, Milan Straka, Jana Strnadová, Alane Suhr, Umut Sulubacak, Zsolt Szántó, Dima Taji, Yuta Takahashi, Takaaki Tanaka, Isabelle Tellier, Trond Trosterud, Anna Trukhina, Reut Tsarfaty, Francis Tyers, Sumire Uematsu, Zdeňka Urešová, Larraitz Uria, Hans Uszkoreit, Sowmya Vajjala, Daniel van Niekerk, Gertjan van Noord, Viktor Varga, Veronika Vincze, Lars Wallin, Jonathan North Washington, Seyi Williams, Mats Wirén, Tsegay Woldemariam, Tak-sum Wong, Chunxiao Yan, Marat M. Yavrumyan, Zhuoran Yu, Zdeněk Žabokrtský, Amir Zeldes, Daniel Zeman, Manying Zhang, and Hanzhi Zhu. 2018. Universal dependencies 2.2. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

O'Horan, Helen, Yevgeni Berzak, Ivan Vulić, Roi Reichart, and Anna Korhonen. 2016. Survey on the use of typological information in natural language processing. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1297–1308.

Oncevay, Arturo, Barry Haddow, and Alexandra Birch. 2020. Bridging linguistic typology and multilingual machine translation with multi-view language representations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2391–2406. https://doi.org/10.18653/v1/2020.emnlp-main.187

Östling, Robert. 2015. Word order typology through multilingual word alignment. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 205–211. https://doi.org/10.3115/v1/P15-2034

Östling, Robert. 2016. Studying colexification through massively parallell corpora. In Päivi Juvonen and Maria

Koptjevskaja-Tamm, editors, *The Lexical Typology of Semantic Shifts*. De Gruyter. pages 157–176. https://doi.org/10.1515/9783110377675-006

Östling, Robert and Murathan Kurfalı. 2023. Parallel text typology dataset. https://doi.org/10.5281/zenodo.7506219

Östling, Robert and Jörg Tiedemann. 2016. Efficient word alignment with Markov Chain Monte Carlo. *Prague Bulletin of Mathematical Linguistics*, 106:125–146. https://doi.org/10.1515/pralin-2016-0013

Östling, Robert and Jörg Tiedemann. 2017. Continuous multilinguality with language vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 644–649. https://doi.org/10.18653/v1/E17-2102

Östling, Robert and Bernhard Wälchli. 2019. Word-order goes lexical typology: Adjective-noun order and massively parallel text. In *13th Conference of the Association for Linguistic Typology*, pages 378–380.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Platanios, Emmanouil Antonios, Mrinmaya Sachan, Graham Neubig, and Tom Mitchell. 2018. Contextual parameter generation for universal neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 425–435. https://doi.org/10.18653/v1/D18-1039

Ponti, Edoardo Maria, 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):1–43. https://doi.org/10.1162/coli_a_00357

Rama, Taraka, Lisa Beinborn, and Steffen Eger. 2020. Probing multilingual BERT for genetic and typological signals. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1214–1228. https://doi.org/10.18653/v1/2020.coling-main.105

Shopen, Timothy, editor. 2007. *Language Typology and Syntactic Description*, 2nd edition, volume 2. Cambridge University Press. https://doi.org/10.1017/CBO9780511619434

Smith, Samuel L., David H. P. Turban, Steven Hamblin, and Nils Y. Hammerla. 2017. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. *arXiv preprint arXiv:1702.03859v1*.

Søgaard, Anders, Ivan Vulić, Sebastian Ruder, and Manaal Faruqui. 2019. *Cross-Lingual Word Embeddings*, 2nd edition. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers, United States. https://doi.org/10.1007/978-3-031-02171-8

Stanczak, Karolina, Edoardo Ponti, Lucas Torroba Hennigen, Ryan Cotterell, and Isabelle Augenstein. 2022. Same neurons, different languages: Probing morphosyntax in multilingual pre-trained models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1589–1598. https://doi.org/10.18653/v1/2022.naacl-main.114

Sylak-Glassman, John, Christo Kirov, Matt Post, Roger Que, and David Yarowsky. 2015. A universal feature schema for rich morphological annotation and fine-grained cross-lingual part-of-speech tagging. In *Systems and Frameworks for Computational Morphology*, pages 72–93, Springer. https://doi.org/10.1007/978-3-319-23980-4_5

Tan, Xu, Jiale Chen, Di He, Yingce Xia, Tao Qin, and Tie-Yan Liu. 2019. Multilingual neural machine translation with language clustering. *arXiv preprint arXiv:1908.09324v1*. https://doi.org/10.18653/v1/D19-1089

Tiedemann, Jörg. 2011. Bitext Alignment. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers. https://doi.org/10.1007/978-3-031-02142-8_5

Tiedemann, Jörg. 2018. Emerging language spaces learned from massively multilingual corpora. *arXiv preprint arXiv:1802.00273v1*.

Üstün, Ahmet, Arianna Bisazza, Gosse Bouma, and Gertjan van Noord. 2020. UDapter: Language adaptation for truly Universal Dependency parsing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language*

Processing (EMNLP), pages 2302–2315. https://doi.org/10.18653/v1/2020.emnlp-main.180

Vastl, Martin, Daniel Zeman, and Rudolf Rosa. 2020. Predicting typological features in WALS using language embeddings and conditional probabilities: ÚFAL submission to the SIGTYP 2020 shared task. In *Proceedings of the Second Workshop on Computational Research in Linguistic Typology*, pages 29–35. https://doi.org/10.18653/v1/2020.sigtyp-1.4

Wälchli, Bernhard and Michael Cysouw. 2012. Lexical typology through similarity semantics: Toward a semantic map of motion verbs. *Linguistics*, 50(3):671–710. https://doi.org/10.1515/ling-2012-0021

Wang, Xinyi, Sebastian Ruder, and Graham Neubig. 2022. Expanding pretrained models to thousands more languages via lexicon-based adaptation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 863–877. https://doi.org/10.18653/v1/2022.acl-long.61

Wichmann, Søren, Eric W. Holman, and Cecil H. Brown. 2018. The ASJP database (version 18). https://asjp.clld.org/

Xue, Linting, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498. https://doi.org/10.18653/v1/2021.naacl-main.41, PubMed: 33576803

Yarowsky, David and Grace Ngai. 2001. Inducing multilingual POS taggers and NP bracketers via robust projection across aligned corpora. In *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics on Language Technologies*, NACCL'01, pages 1–8. https://doi.org/10.3115/1073336.1073362