Using Cross-Lingual Part of Speech Tagging for Partially Reconstructing the Classic Language Family Tree Model

Anat Samohi[*]
Efi Arazi School of Computer Science
Reichman University, Israel
anatsamohi@gmail.com

Daniel Weisberg Mitelman[*]
The Data Science Institute
Reichman University, Israel
dwmitelman@gmail.com

Kfir Bar
The Data Science Institute
Reichman University, Israel
barkfir@yahoo.com

Abstract

The tree model is well known for expressing the historic evolution of languages. This model has been considered as a method of describing genetic relationships between languages. Nevertheless, some researchers question the model's ability to predict the proximity between two languages, since it represents genetic relatedness rather than linguistic resemblance. Defining other language proximity models has been an active research area for many years. In this paper we explore a part-of-speech model for defining proximity between languages using a multilingual language model that was fine-tuned on the task of cross-lingual part-of-speech tagging. We train the model on one language and evaluate it on another; the measured performance is then used to define the proximity between the two languages. By further developing the model, we show that it can reconstruct some parts of the tree model.

1 Introduction

Language families are defined by the evolution of languages over the history, providing indications regarding the proximity between them. The tree model, which was first introduced by Augustus Schleicher (Schleicher, 1853) is considered as the consensual language-family model. For example, Figure 1 shows the Indo-European branch of the tree model; a full version of the model is nicely presented on *Ethnologue*[1]. In this paper, we refer to this source as a reference for the classic family tree model.

Concomitantly, there have been theories that question the tree model as being an indicator for language proximity, since it represents genetic relatedness rather than lexical resemblance. Loanwords, as well as other lexical influences are usually not expressed in the classic tree model. Representing historical relatedness, the tree is agnostic to various linguistic influences. Consequently, some claim that language families should be defined by alternative models (Geisler and List, 2013).

The Universal Grammar, introduced by Noam Chomsky, is usually defined as the "system of categories, mechanisms and constraints shared by all human languages and considered to be innate" (Dobrovolsky et al., 2016). In other words, a human language is derived from a set of structural rules, typically referred to as *generative grammar*, which we are usually totally unaware of. We can intuitively distinguish between nouns and verbs; children can phrase a sentence they have not heard before by ordering parts of speech they are familiar with in a valid grammatical order. A child can identify a noun without knowing what a noun is, or without even understanding the meaning of that specific noun.

It may be assumed that rather than this aspect of universal grammar being specific to language, it is more generally a part of human cognition, and there might be a common structure for different languages. Still, the ability to classify words into parts of speech requires some knowledge of the structure of the specific language.

The hypothesis we examine in this paper relies on the assumption that historically close languages, like French and Spanish, share some information that may help the classification of words into part-of-speech (POS) tags. While identifying this type of information is out of scope for this paper, we will show that this information can be used by a neural network for predicting POS tags of one language only using examples from another language.

Our goal is to redefine the proximity between languages to achieve a comparable model to the classic tree model, by considering only POS tags.

[*]Equal contribution.
[1]https://www.ethnologue.com/browse/families

Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change, pages 78 - 88

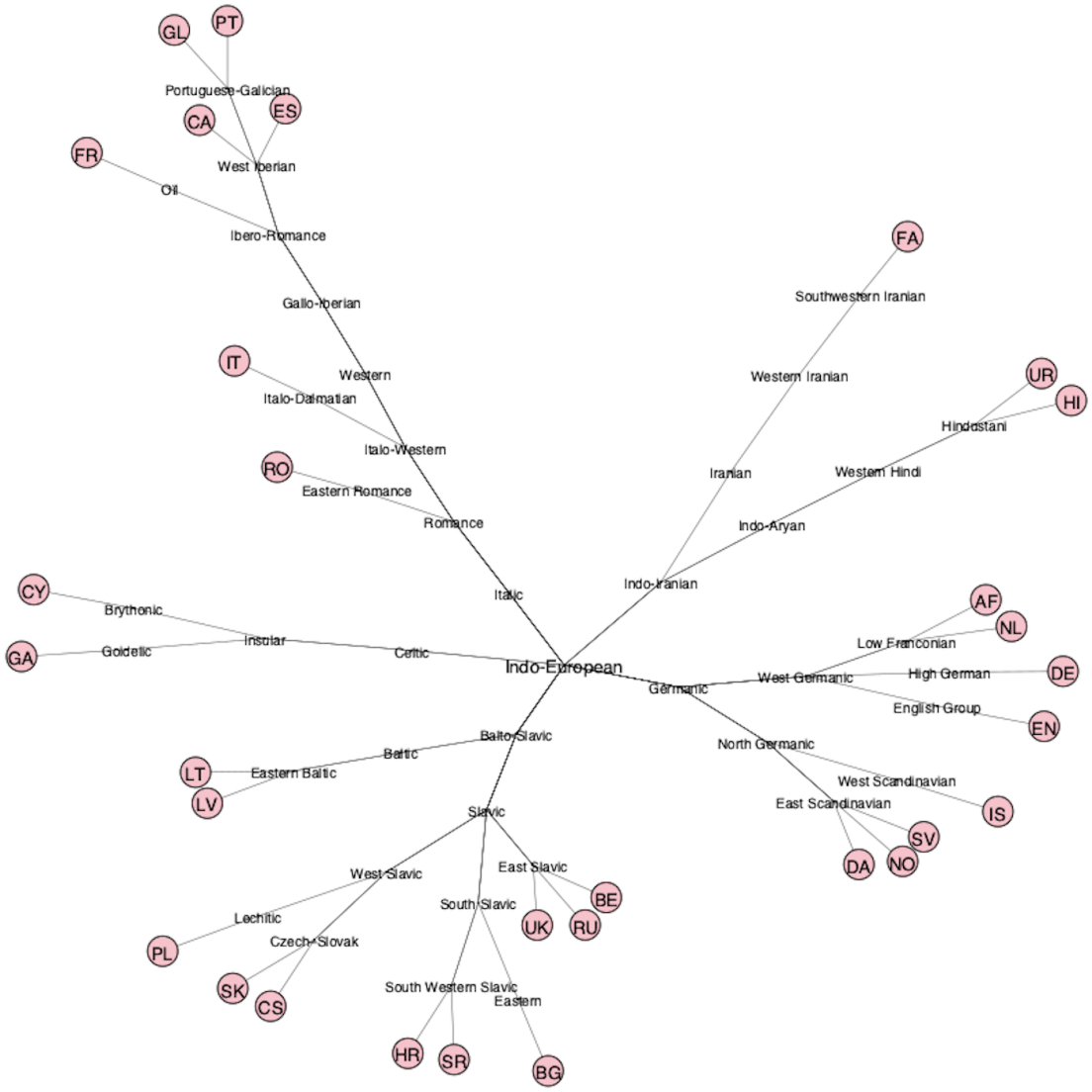

Figure 1: The Indo-European language branch; the graph was created by the `igraph` package for Python (Csardi and Nepusz, 2006). The languages are represented using their equivalent two-letter ISO 639-1 code.

To enable transferability between languages, we suggest using a multilingual pre-trained language model (MPLM), fine-tuned for the POS tagging task in a multilingual environment. Specifically, we take a multilingual zero-shot training approach by fine-tuning an MPLM to predict POS tags for texts written in one language, the source language, and evaluating it on texts written in another language, the target language. The performance metrics are then used to estimate the similarity between the source and target languages. As a final step, we generate a language-similarity graph, which we describe as an approximation for the classic tree model. We make two main contributions: (1) Reconstructing part of the classic tree model using POS-based similarity scores; and, (2) Providing some insights into the cross-lingual generalization of MPLMs.

We proceed as follows: In Section 2 we cite some related studies, following by a detailed description of our method, provided in Section 3. We end Section 3 with reporting on some results. We discuss the results in Section 4 and make some conclusions.

2 Related Work

There have been some prior studies on measuring distance between languages. In their paper, Chiswick and Miller (2005) presented some empirical observations of how rapidly speakers of

a given language gained proficiency in another tongue. Specifically, they measured the speed of English acquisition by immigrants of various linguistic backgrounds in the United States and Canada. Their first languages were ranked for the distance from English, on a scale from 1.0 (very different than English) to 3.0 (closest to English). It has been found empirically that the greater the distance between an immigrant's origin language and English, the lower is the level of the immigrant's English language proficiency.

There have been many attempts to use computational tools to infer the relations between languages; the dominant approach is known as *phylogenetic linguistics*. Phylogenetic linguistics is about establishing historical relationships among languages, by considering the evolutionary nature of human languages. In computational phylogenetics, words and/or phonemes of what counts as the same language over time, are analyzed and compared among languages.

Specifically, Swadesh (1950) was first to introduce a computational phylogenetic technique called *lexicostatistics* for comparing between two languages. In lexicostatistics, the similarity between two languages is calculated by a function of the percentage of cognates found in a predefined list of words of the two languages. Swadesh's work has been followed by a number of studies that use lexicostatistics or a minor variation of it (Nakhleh et al., 2005; Holman et al., 2008; Bakker et al., 2009; Petroni and Serva, 2010; Barbançon et al., 2013).

Instead of measuring the percentage of cognates, Petroni and Serva (Petroni and Serva, 2008; Serva and Petroni, 2008) proposed to calculate a normalized Levenstein distance among words with the same meaning and then to take the average over the words contained in a cross-lingual list. Müller et al. (2010) conducted a lexical comparison using the Levenstein distance approach, between 4,350 languages of the ASJP database (Brown et al., 2008), and created a full diagram of lexical proximity. They showed that lexical resemblance is related to genetic affiliation. However, some of the languages that have been found as lexically similar, according to their technique, are not closely genetically associated.

Another computational approach for measuring language similarity is based on corpus analysis. Gamallo et al. (2017) used the known perplexity score of a probabilistic n-gram language model to measure the distance between European languages. Asgari and Mofrad (2016) compared 50 languages from different families by training a monolingual language model on each language individually, using a parallel corpus of the Bible (Christodouloupoulos and Steedman, 2015), and apply them to calculate perplexity on all the other languages. In some of the works that are mentioned above, the proximity between languages is not perfectly aligned with the classic tree model.

While the main focus has always been on lexical similarity, some attempts were made to compare languages on the syntactic level. Longobardi and Guardiano (2009) characterized 28 languages, mostly Indo-European ones, using a set of 63 predefined morpho-syntactic parameters. They calculated a normalized Hamming distance over those parametric representations, with which they were able to generate a language tree. They showed that this tree is equivalent to a tree that was generated based on a traditional lexicostatistics approach, suggesting that syntactic characteristics are sufficiently robust to reconstruct a plausible historical language tree. The same method was re-used in (Longobardi et al., 2013), which was concluded in a similar way. In a recent work, Shu et al. (2021) applied a different comparison technique on the same syntactic characteristics, using Markov models. In all of those works, the selection of the syntactic characteristics to be used for comparison, plays an important role in the creation of a language proximity model.

To the best of our knowledge, there have not been attempts to compare languages using syntactic information in a non-parametric way. In this work, we take a corpus-based approach to automatically extract part-of-speech tags from a given text in order to generate a language-proximity model. In that sense, we consider our approach as a non-parametric estimation method, since we do not need to manually define specific syntactic parameters to consider for calculating similarity between languages.

To transfer information across languages, we use mBERT, a multilingual version of BERT (Devlin et al., 2019), that was pre-trained on texts written in over 100 languages based on a shared vocabulary.[2] During pre-training, the training documents are given to mBERT without any indication on the language that they have been written with. Like

[2]Similar to BERT, mBERT's tokens are subwords.

every other pre-trained language model (PLM), the pre-trained mBERT model is typically fine-tuned on a training set of a specific downstream task, which could be either monolingual or multilingual. This unique multilingual design allows mBERT to handle multilingual tasks in a transfer-learning way. In another study, Wu and Dredze (2019) reported an impressive performance using mBERT in a zero-shot cross-lingual transfer learning setting on several NLP tasks, including POS tagging. They claimed that mBERT may learn a cross-lingual representation by generalizing and abstracting some language-specific information. A similar observation was made by Gonen et al. (2020) who claimed that mBERT learns information by two components, one that encodes the language and another that encodes some abstract information that can be used in a cross-lingual way.

3 Methodology

3.1 Language Similarity Score

For every pair of languages, source language and target language, we measure their similarity as the performance of an mBERT-based POS tagger fine-tuned on the source language, and evaluated on the target language. For training and evaluation, we use treebanks from Universal Dependencies (UD).[3] In particular, we use the Universal POS labels[4] assigned for every syntactic word in the text. The Universal POS tagset contains the following core part-of-speech categories that can be used for any UD language: adjective, adposition, adverb, auxiliary, coordinating conjunction, determiner, interjection, noun, numeral, particle, pronoun, proper noun, punctuation, subordinating conjunction, symbol, verb and other. Each treebank is divided to train and test sets. Therefore, we fine-tune mBERT on the UPOS (universal POS) tagging task using the source language's training set, and evaluate it on the target language's testing set.

Our selected evaluation metric is the micro average F1 score. Clearly, for every pair of languages we calculate two F1 scores, one for each direction. The two scores are not necessarily equivalent.

In all our experiments, we use the commonly used pre-trained language model `bert-base-multilingual-cased`, provided by the Hugging Face transformers library (Wolf et al., 2019). For every language we fine-tune the model for the standard token classification downstream task for three epochs, using a learning rate value of $5e-5$.

We include 36 languages in our study, taken from a diversity of language families and subfamilies. The full list of languages is provided in Figure 2. For each language we indicate its two-letter ISO 639-1 code, which we use throughout the paper. All the 36 languages we process are covered by mBERT.

Overall, we calculate the F1 score for every pair of languages, resulting in $36^2 = 1296$ scores. A partial list of the scores is provided in Table 1, while the full set of results is added as Appendix A. Clearly, the model that is trained on English performs better on Spanish than on Russian and Hindi.

Src/Trgt	EN	ES	RU	HI
EN	0.97	0.84	0.80	0.64
ES	0.80	0.99	0.80	0.58
RU	0.74	0.81	0.98	0.64
HI	0.61	0.57	0.67	0.97

Table 1: F1 scores for some of the language pairs. Rows represent source languages, while columns represent the target languages. For example, the first row represents the F1 scores resulted from evaluating mBERT on the UPOS tagged test sets in English, Spanish, Russian and Hindi, after previously fine-tuned on the English UPOS tagged train set.

As mentioned before, the two F1 scores that were calculated for each pair of different languages, are not necessarily equal. In fact, they are very unlikely to be equal, since the performance of the tagger is affected not only by the difference between the languages, but also by the size and the quality of the training sets, as well as the volume and quality of the texts in each relevant language, which were used for training mBERT.

The average of the absolute difference between all language pairs is 0.0874 and the standard deviation is 0.074. While some pairs have relatively similar scores in both directions, some other have significantly different ones. However, as we show later, we do not use the F1 scores directly as some sort of a distance function between the languages. Instead, we represent each language l by a vector of F1 scores calculated by all other models during evaluation on l's testing set, and use a clustering

[3] https://universaldependencies.org

[4] https://universaldependencies.org/u/pos

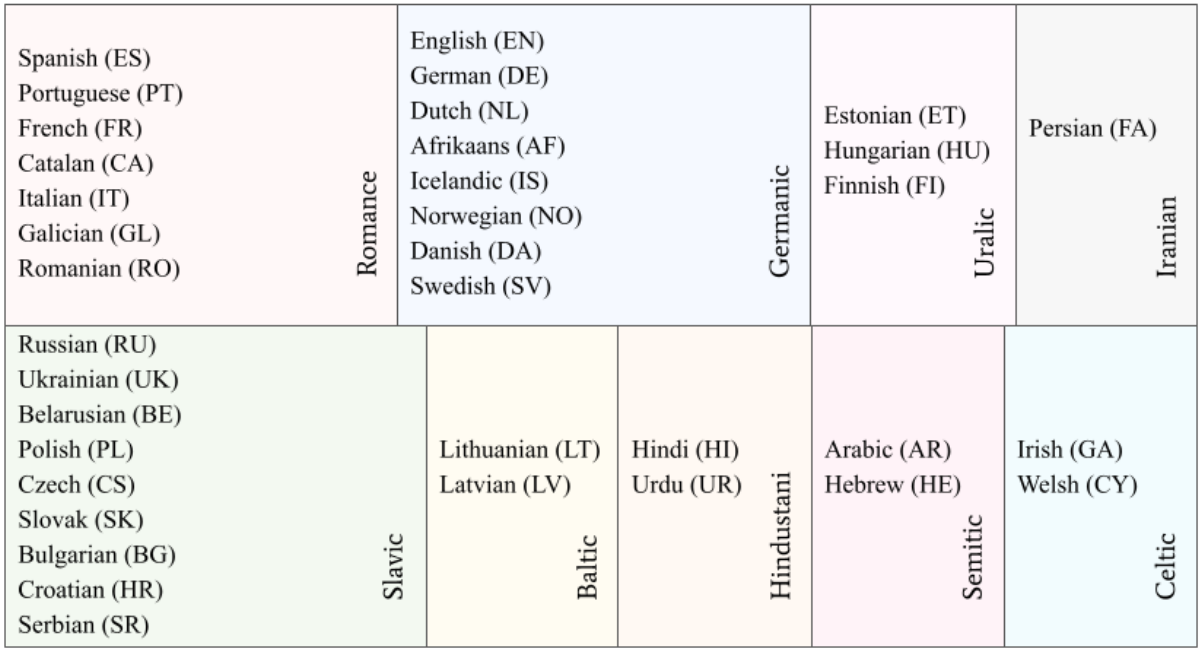

Figure 2: The 36 languages we include in our study. We chose languages from different families and subfamilies. The two-letter ISO 639-1 code is provided in parentheses next to each language name.

algorithm to organize these vectors into language families.

Before we show how we do that, first, we argue that our cross-lingual F1 score is an important piece of information for reconstructing the classic tree model. Our argument is based on the correlation between our cross-lingual F1 score and the proximity of language pairs in the classic tree model. In order to measure the proximity between two languages in the classic tree model, we use the Wu-Palmer similarity (Wu and Palmer, 1994) metric, which was originally invented for measuring relatedness of two synsets in a WordNet taxonomy. For the context of using Wu-Palmer, the tree model has the same characteristics as WordNet; language family names are represented by intermediate nodes, while language names are represented by the leaves. Therefore, the Wu-Palmer score for two languages L_1, L_2 is calculated as follows:

$$2 \cdot \frac{depth(lcs(L_1, L_2))}{depth(L_1) + depth(L_2)}$$

with lcs representing the least common subsumer, that is, the first common ancestor of the two languages in the language-family tree. The score ranges between 0 and 1, but it can never go to zero since the depth of $lcs(L_1, L_2)$ is never zero (the model tree has a single root).

We denote the Wu-Palmer score as WP. As opposed to our cross-lingual F1 score, WP is symmetric.

We calculate WP for every language pair, and compare with our F1; the results are shown in Figure 3. Every data point in this chart represents a single language pair out of the 36^2 pairs. Overall, we learn that the F1 score increases along with WP, except maybe on relatively small WP values, representing pairs of languages taken from significantly different branches of the language-family tree.

Furthermore, we measure the correlation between the two metrics using Pearson (for linear correlation) and Spearman (for monotonic correlation) and realize that both are strongly correlated with Pearson$= 0.64$ (at $p < 0.001$), and Spearman$= 0.59$, (at $p < 0.001$).

Figure 4 visualizes the F1 scores of all language pairs as a heatmap, with target languages provided as rows and source languages as columns. For each target language, all the 36 source languages are sorted according to the F1 scores (from the highest to the lowest). The color represents the proximity, as calculated by the WP score; a lighter color is equivalent to a higher proximity. For example in the fifth row, the best performance on the Spanish test set is observed by the Spanish model, followed by other Romance languages, Catalan, Italian, Portuguese and so on. The worst performance was recorded by the Welsh model. Evidently, higher

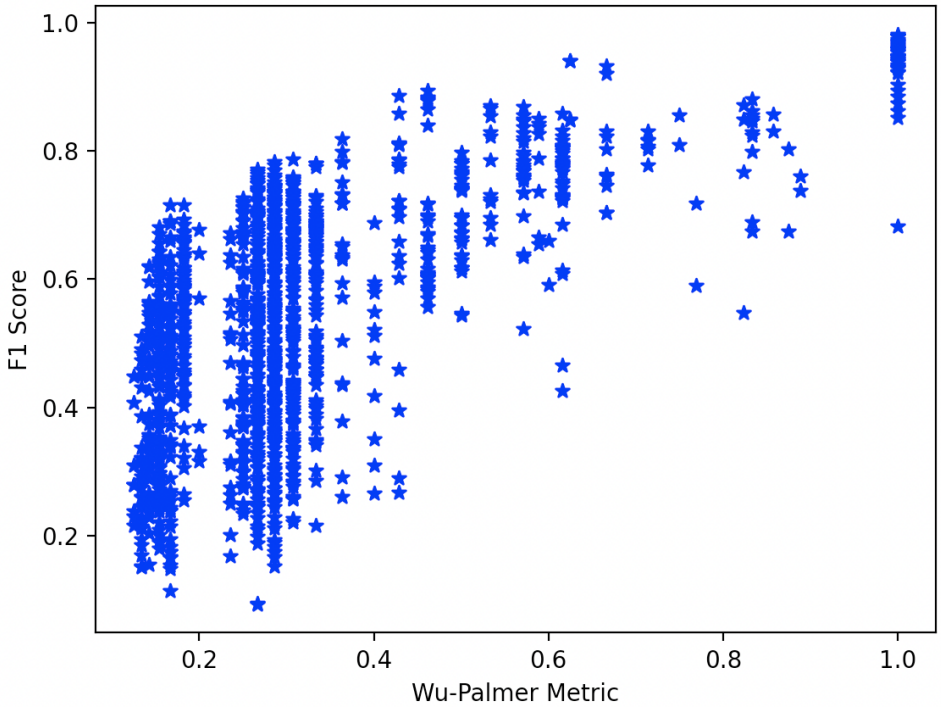

Figure 3: UPOS F1 scores compared with WP scores.

proximity values (light boxes on the left side of the heatmap) derive higher performance on the cross-lingual POS task, indicating that the closer two languages are, the encoded information in their corresponding models tends to be more helpful for POS tagging.

3.2 Reconstructing Language Families

In this section we show how we use the resulting F1 scores, calculated for every language pair, to reconstruct the language-family tree.

We represent every language l by a 36-dimensional vector consisted of the F1 scores of the models that have been trained on all other languages, evaluated on l. We generate exactly 36 vectors, one for each language. Conceptually, the vector of language l represents the similarity of l to all the other languages, by considering only cross-lingual UPOS information, as captured by mBERT.

To identify families and subfamilies of languages, we use k-means (Lloyd, 1982) to cluster the 36 vectors. In addition to the collection of vectors, k-means receives as input a parameter k that denotes the number of clusters.

According to Figure 2, the tree model organizes the 36 languages into 9 families; therefore, we run k-means with value of $k = 9$. In Figure 5 we visualize the resulting clusters. The color of the circle next to the language name marks the cluster. Note that while the k-means algorithm works with 36-dimensional vectors given as an input, we visualize the vectors on a 2-dimensional axis, which we calculate using the principal component analysis (PCA) algorithm for reducing dimensions. The clusters are summarized in Table 2. We discuss the results in the following section.

3.3 Results

The alternative partitioning for language families that we get, partially align with the classic tree model.

Cluster 1 contains only Romance languages. All languages in cluster 2, except Romanian, are considered as Germanic in the classic tree model. Cluster 3 contains all Slavic languages excluding the Baltic languages. Cluster 4 includes the Baltic languages (Lithuanian and Latvian) as well as two Uralic languages (Finnish and Estonian). Those four languages are spoken in the geographically close countries Lithuania, Latvia, Finland and Estonia, respectively, suggesting that there might be a geographical dimension in our POS-based language proximity method. We plan to further investigate this discovery as one of our future directions. Cluster 6 contains only Hindustani languages. The two Semitic languages (Hebrew and Arabic) are grouped together in cluster 7, which also includes

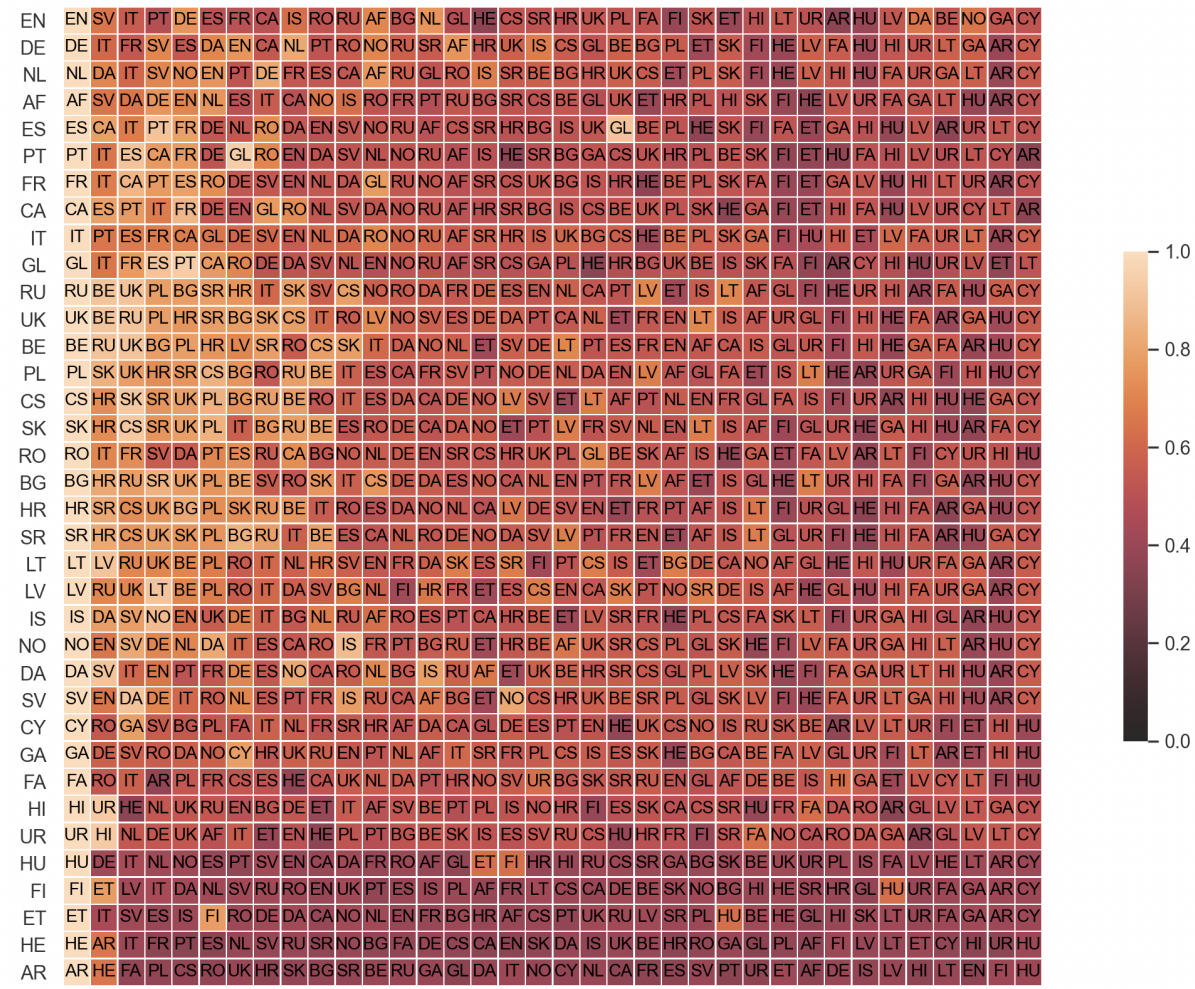

Figure 4: A heatmap of the WP scores calculated for all language pairs, sorted according to F1 scores. For more information about this arrangement see the text.

Cluster	Languages	Family
1	Spanish, Portuguese, French, Catalan, Italian, Galician	Romance
2	English, German, Dutch, Afrikaans, Icelandic, Norwegian, Danish, Swedish, Romanian	Mostly Germanic
3	Russian, Ukrainian, Belarusian, Polish, Czech, Slovak, Bulgarian, Croatian, Serbian	Slavic
4	Lithuanian, Latvian, Finnish, Estonian	Baltic and Uralic
5	Hungarian	Uralic
6	Hindi, Urdu	Hindustani
7	Persian, Hebrew, Arabic	Iranian and Semitic
8	Irish	Celtic
9	Welsh	Celtic

Table 2: The clusters obtained by running k-means with $k = 9$. We provide some information about the language families of each cluster in the third column.

Persian probably due to historical influences. Hungarian is the only language in cluster 5. Clusters 8 and 9 represents two languages of the Celtic family. They should have probably been clustered together.

Overall although there are a few misplacements, our clustering method was able to reconstruct parts of the tree model. 31 out of 36 languages were classified correctly according to the classic model.

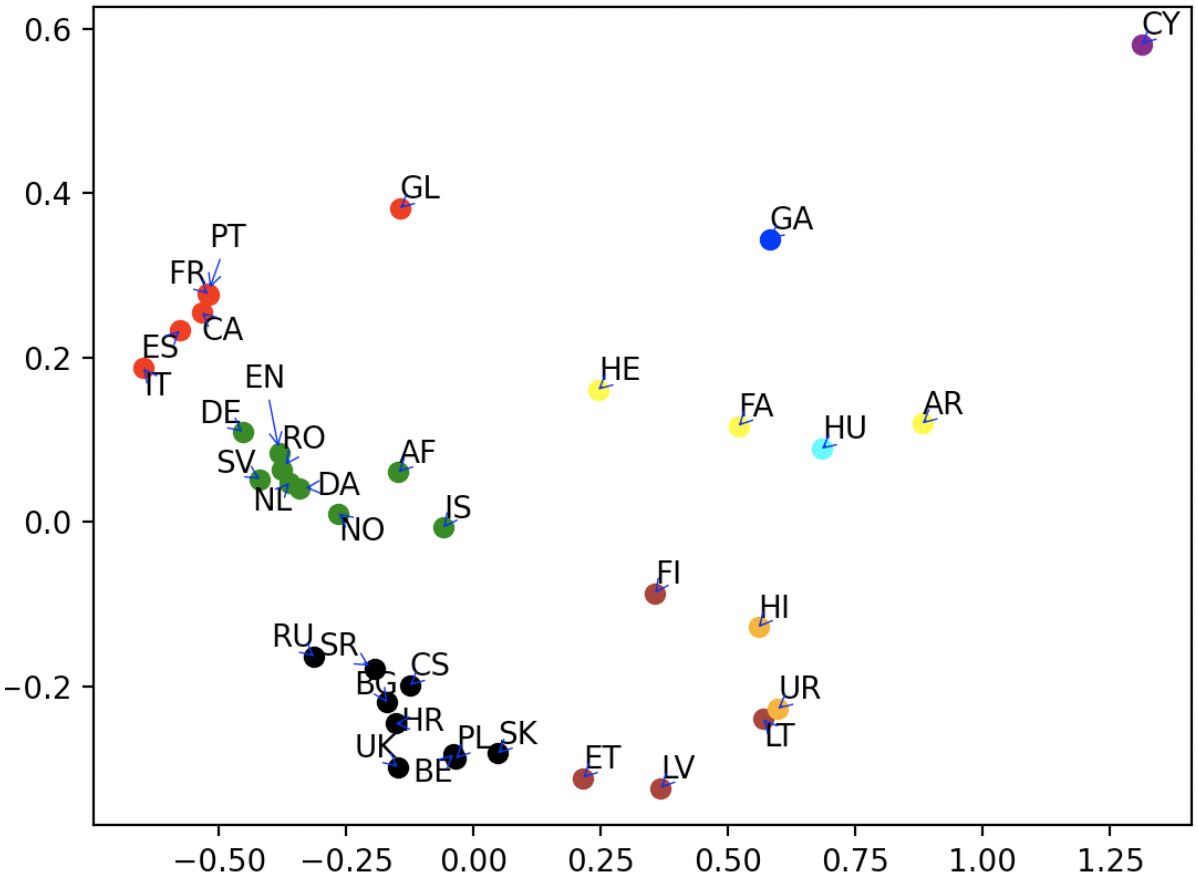

Figure 5: The 9 clusters resulted from k-means. The original 36-dimensional vectors are visualized using their first two principle components.

4 Discussion and Conclusions

In this work we used a cross-lingual model trained on UPOS for measuring the proximity between languages. We showed that our new language-proximity model can reconstruct families of genetically related languages, suggesting that POS information plays a major role in modelling similarity between languages.

We believe that we have demonstrated the potential of a fine-tuned mBERT model to capture some cross-lingual information that is needed for assigning UPOS tags to a text written in an unseen language. On average, models of genetically related languages perform better on each other in this task, even if they are not written in the same script. For example, in Table 1 we show that a Spanish (ES) model performs similarly on English (EN) and Russian (RU), although both Spanish and English are written in the Latin script while Russian is written in the Cyrillic script.

There are a few caveats to this research to note. mBERT was pre-trained on the full collections of Wikipedia articles in the relevant languages. Therefore, the size of those collections varies proportionally to the number of active speakers. To handle that bias, the authors of mBERT had decided to upsample the Wikipedia collections of the less dominant languages, in the main training loop. Wu and Dredze (2020) have recently addressed that problem and showed that mBERT performs better on cross-lingual zero-shot tasks on languages that have large Wikipedia collections. In our work, we handle that bias by designing each individual language vector to have F1 scores from all other languages, including both high-resource and low-resource languages. Therefore, every language is represented by F1 scores achieved by models trained on exactly the same language set.

Another caveat is the size and quality of the treebanks we use for training and testing our models. As noted before, we believe that our approach to represent a language using scores from models trained on all the 36 language included in this

research, mitigates this risk.

We make a final practical observation. The results of our study suggest that for UPOS tagging, mBERT may benefit from training on texts written in languages that are genetically similar to the target language, based on the classic tree model. These results are aligned with what have been reported by Wu and Dredze (2020).

References

Ehsaneddin Asgari and Mohammad R.K. Mofrad. 2016. Comparing fifty natural languages and twelve genetic languages using word embedding language divergence (WELD) as a quantitative measure of language distance. In *Proceedings of the Workshop on Multilingual and Cross-lingual Methods in NLP*, pages 65–74, San Diego, California. Association for Computational Linguistics.

Dik Bakker, 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.

François Barbançon, Steven N Evans, Luay Nakhleh, Don Ringe, and Tandy Warnow. 2013. An experimental study comparing linguistic phylogenetic reconstruction methods. *Diachronica*, 30(2):143–170.

Cecil H Brown, Eric W Holman, Søren Wichmann, and Viveka Velupillai. 2008. Automated classification of the world s languages: a description of the method and preliminary results. *Language Typology and Universals*, 61(4):285–308.

Barry R Chiswick and Paul W Miller. 2005. Linguistic distance: A quantitative measure of the distance between English and other languages. *Journal of Multilingual and Multicultural Development*, 26(1):1–11.

Christos Christodouloupoulos and Mark Steedman. 2015. A massively parallel corpus: The Bible in 100 languages. *Language resources and evaluation*, 49(2):375–395.

Gabor Csardi and Tamas Nepusz. 2006. The igraph software package for complex network research. *InterJournal*, Complex Systems:1695.

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

Michael Dobrovolsky, William Delaney O'Grady, and Francis Katamba. 2016. *Contemporary Linguistics*. Longman.

Pablo Gamallo, José Ramom Pichel, and Iñaki Alegria. 2017. From language identification to language distance. *Physica A: Statistical Mechanics and its Applications*, 484:152–162.

Hans Geisler and Johann-Mattis List. 2013. Do languages grow on trees? The tree metaphor in the history of linguistics. *Classification and evolution in biology, linguistics and the history of science. Concepts–methods–visualization*, pages 111–124.

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

Eric W Holman, Søren Wichmann, Cecil H Brown, Viveka Velupillai, André Müller, and Dik Bakker. 2008. Explorations in automated lexicostatistics. *Folia Linguistica*, 42(2):331–354.

Stuart Lloyd. 1982. Least squares quantization in PCM. *IEEE transactions on information theory*, 28(2):129–137.

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

Giuseppe Longobardi, Cristina Guardiano, Giuseppina Silvestri, Alessio Boattini, and Andrea Ceolin. 2013. Toward a syntactic phylogeny of modern Indo-European languages. *Journal of Historical Linguistics*, 3(1):122–152.

André Müller, Søren Wichmann, Viveka Velupillai, Cecil H Brown, Pamela Brown, Sebastian Sauppe, Eric W Holman, Dik Bakker, Johann-Mattis List, Dmitri Egorov, et al. 2010. ASJP world language tree of lexical similarity: Version 3 (July 2010). *Retrieved*, 10(19):2015.

Luay Nakhleh, Don Ringe, and Tandy Warnow. 2005. Perfect phylogenetic networks: A new methodology for reconstructing the evolutionary history of natural languages. *Language*, pages 382–420.

Filippo Petroni and Maurizio Serva. 2008. Language distance and tree reconstruction. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(08):P08012.

Filippo Petroni and Maurizio Serva. 2010. Measures of lexical distance between languages. *Physica A: Statistical Mechanics and its Applications*, 389(11):2280–2283.

August Schleicher. 1853. Die ersten spaltungen des indogermanischen urvolkes. *Allgemeine Monatsschrift für Wissenschaft und Literatur*, 3:786–787.

Maurizio Serva and Filippo Petroni. 2008. Indo-European languages tree by Levenshtein distance. *EPL (Europhysics Letters)*, 81(6):68005.

Kevin Shu, Andrew Ortegaray, Robert C Berwick, and Matilde Marcolli. 2021. Phylogenetics of Indo-European language families via an algebro-geometric analysis of their syntactic structures. *Mathematics in Computer Science*, pages 1–55.

Morris Swadesh. 1950. Salish internal relationships. *International Journal of American Linguistics*, 16(4):157–167.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

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

Shijie Wu and Mark Dredze. 2020. Are all languages created equal in multilingual BERT? In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 120–130, Online. Association for Computational Linguistics.

Zhibiao Wu and Martha Palmer. 1994. Verb semantics and lexical selection. In *Proceedings of the 32nd Annual meeting of the Associations for Computational Linguistics*, pages 133–138.

A Appendix: F1 Scores

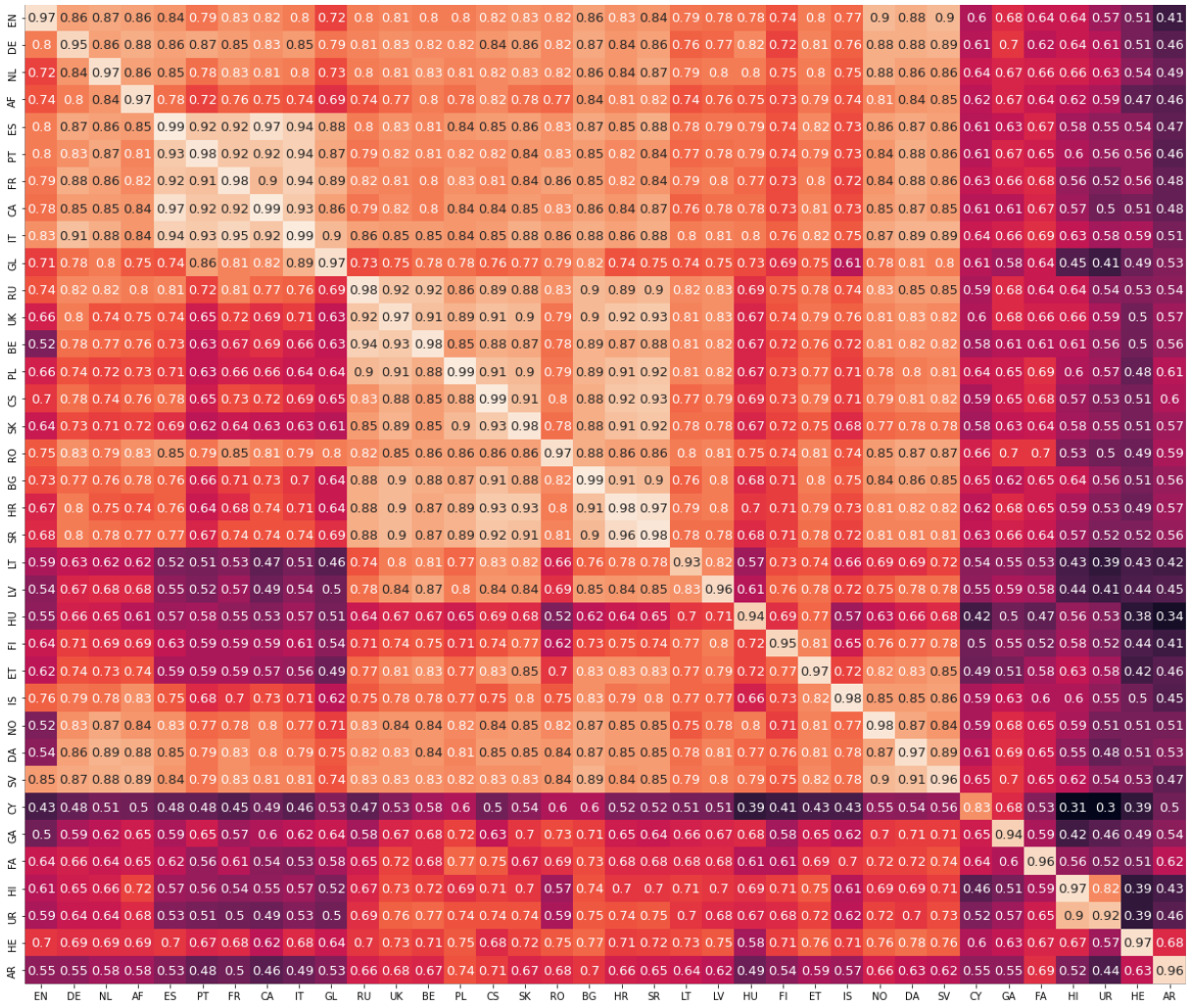

Figure 6: F1 Scores for all language pairs.