Probing Multilingual BERT for Genetic and Typological Signals

Taraka Rama
University of North Texas
taraka.kasi@gmail.com

Lisa Beinborn
VU Amsterdam
l.m.beinborn@vu.nl

Steffen Eger
TU Darmstadt
eger@aiphes.tu-darmstadt.de

Abstract

We probe the layers in multilingual BERT (mBERT) for phylogenetic and geographic language signals across 100 languages and compute language distances based on the mBERT representations. We 1) employ the language distances to infer and evaluate language trees, finding that they are close to the reference family tree in terms of quartet tree distance, 2) perform distance matrix regression analysis, finding that the language distances can be best explained by phylogenetic and worst by structural factors and 3) present a novel measure for measuring diachronic meaning stability (based on cross-lingual representation variability) which correlates significantly with published ranked lists based on linguistic approaches. Our results contribute to the nascent field of typological interpretability of cross-lingual text representations.

1 Introduction

Cross-lingual text representations have become extremely popular in NLP, since they promise universal text processing in multiple human languages with labeled training data only in a single one. They go back at least to the work of Klementiev et al. (2012), and have seen an exploding number of contributions in recent years. Recent cross-lingual models provide representations for about 100 languages and vary in their training objectives. In offline learning, cross-lingual representations are obtained by projecting independently trained monolingual representations into a shared representational space using bilingual lexical resources (Faruqui and Dyer, 2014; Artetxe et al., 2017). In joint learning (Wang et al., 2020), the cross-lingual representations are learned directly, for example as a byproduct of large-scale machine translation (Artetxe and Schwenk, 2018).

As parallel data is scarce for less frequent language pairs, the multilingual BERT model (mBERT) simply trains the BERT architecture (Devlin et al., 2019) on multilingual input from Wikipedia. The cross-lingual signal is thus only learned implicitly because mBERT uses the same representational space independent of the input language. This naive approach yields surprisingly high scores for cross-lingual downstream tasks, but the transfer does not work equally well for all languages. Pires et al. (2019) show that the performance differences between languages are gradual and that the representational similarity between languages seem to correlate with typological features. These relationships between languages remain opaque in cross-lingual representations and pose a challenge for the evaluation of their adequacy. Evaluations in down-stream tasks are an unreliable approximation because they can often be solved without accounting for deep linguistic knowledge or for interdependencies between subgroups of languages (Liang et al., 2020).

While more language-agnostic representations can be beneficial to improve the average performance in task-oriented settings and to smooth the performance differences between high- and low-resource languages (Libovickỳ et al., 2019; Zhao et al., 2020a), linguists are more interested in the representational differences between languages. The field of computational historical linguistics, for example, examines subtle semantic and syntactic cues to infer phylogenetic relations between languages (Rama and Borin, 2015; Jäger, 2014). Important aspects are the diachronic stability of word meaning (Pagel et al., 2007;

Proceedings of the 28th International Conference on Computational Linguistics, pages 1214–1228
Barcelona, Spain (Online), December 8-13, 2020

Holman et al., 2008) and the analysis of structural properties for inferring deep language relationships (Greenhill et al., 2010; Wichmann and Saunders, 2007).

Traditionally, these phenomena have been approximated using hand-selected word lists and typological databases. Common ancestors for languages are typically inferred based on cases of shared word meaning and surface form overlap and it can be assumed that these core properties are also captured in large-scale cross-lingual representations to a certain extent. For example, Beinborn and Choenni (2019) find that phylogenetic relations between languages can be reconstructed from cross-lingual representations if the training objective optimizes monolingual semantic constraints for each language separately as in the multilingual MUSE model (Conneau et al., 2017). MUSE is restricted to only 29 frequent languages, however. While mBERT is a powerful cross-lingual model covering an order of magnitude more languages (104), a better understanding of the type of signal captured in its representations is needed to assess its applicability as a testbed for cross-lingual or historical linguistic hypotheses. Our analysis quantifies the representational similarity across languages in mBERT and disentangles it along genetic, geographic, and structural factors.

In general, the urge to improve the interpretability of internal neural representations has become a major research field in recent years. Whereas dense representations of images can be projected back to pixels to facilitate visual inspection, interpreting the linguistic information captured in dense representation of languages is more complex (Alishahi et al., 2019; Conneau and Kiela, 2018). Diagnostic classifiers (Hupkes et al., 2018), representational stability analysis (Abnar et al., 2019) and indirect visualization techniques (Belinkov and Glass, 2019) are only a few examples for newly developed probing techniques. They are used to examine whether the representations capture part-of-speech information (Zhang and Bowman, 2018), syntactic agreement (Giulianelli et al., 2018), speech features (Chrupała et al., 2017), and cognitive cues (Wehbe et al., 2014). However, the majority of these interpretability studies focus solely on English. Krasnowska-Kieraś and Wróblewska (2019) perform a contrastive analysis of the syntactic interpretability of English and Polish representations and Eger et al. (2020) probe representations in three lower-resource languages. Cross-lingual interpretability research for multiple languages focuses on the ability to transfer representational knowledge across languages for zero-shot semantics (Pires et al., 2019) and for syntactic phenomena (Dhar and Bisazza, 2018). In this work, we contribute to the nascent field of typological and comparative linguistic interpretability of language representations at scale (Kudugunta et al., 2019) and analyze representations for more than 100 languages.

Our contributions: We probe the representations of one of the current most popular cross-lingual models (mBERT) and find that mBERT lacks information to perform well on cross-lingual semantic retrieval, but can indeed be used to accurately infer a phylogenetic language tree for 100 languages. Our results indicate that the quality of the induced tree depends on the inference algorithm and might also be the effect of several conflated signals. In order to better disentangle phylogenetic, geographic, and structural factors, we go beyond simple tree comparison and probe language distances inferred from cross-lingual representations by means of multiple regression. We find phylogenetic similarity to be the strongest and structural similarity to be the weakest signal in our experiments. The phylogenetic signal is present across all layers of mBERT. Our analysis not only contributes to a better interpretation and understanding of mBERT, but may also help explain its cross-lingual behavior in downstream tasks (Pires et al., 2019).[1]

2 Related work

Representational distance between two languages refers to the (averaged) differences between model representations for selected concepts in the two languages. Interpretability analyses attempt to disentangle the typological factors that influence the representational distance. In this work, we distinguish between phylogenetic, geographic and structural factors. Two languages are considered to be *phylogenetically* close if they descend from a common ancestor language. *Geographically* close languages are languages which are primarily spoken in regions with a small physical distance on Earth. *Structural* similarity between languages refers to shared syntactic and morphological features. For many languages,

[1] Our code and data are available from `https://github.com/PhyloStar/mBertTypology`.

the three categories overlap, but they are not necessarily linked. For example, Spanish and Basque are geographically close, but structurally and phylogenetically quite distant.

Previous approaches differ in the type of cross-lingual representations, the number of languages, and the methodology for determining representational distance and for interpreting the typological signal. Kudugunta et al. (2019) obtain representations for 102 language pairs (English ↔ language X) using neural machine translation and then visualize the representations using dimensionality reduction. They explore the visualization qualitatively and find clusters which resemble language families. However, when zooming into the clusters, it becomes evident that a mixture of genetic and geographic factors contributes to the representational distance. For instance, Dravidian and Indo-Aryan languages overlap completely.

Eger et al. (2016) induce bilingual vector spaces for 21 Europarl languages and quantify representational distance between languages by averaging over the pairwise similarity between word representations. They find that the differences can be better explained by geographic than by phylogenetic factors. Conversely, Rabinovich et al. (2017) analyze English translations of sentences in 17 Europarl languages and find that syntactic traces of the native language of the translator can best be explained by language genetics. Bjerva et al. (2019) use the same dataset and train language representations on the linguistic structure of the sentences. They find that the representational distance between languages can be better explained by structural similarity (obtained from dependency trees) than by language genetics. Pretrained cross-lingual models are optimized for tasks such as bilingual lexicon induction and machine translation. Even if linguistic information is not explicitly provided during training, recent interpretability research indicates that phylogenetic properties are encoded in the resulting representations. Beinborn and Choenni (2019) obtain representations for Swadesh word lists from the MUSE model (Conneau et al., 2017) which jointly optimizes monolingual and crosslingual semantic constraints. They find that hiearchical clustering over the representational distance between languages yields phylogenetically plausible language trees. Interestingly, they cannot trace the phylogenetic signal in representations from the sentence-based LASER model (Artetxe and Schwenk, 2018) which is trained to learn language-neutral representations for machine translation. Libovickỳ et al. (2019) analyze representations from mBERT and find that clustering over averaged representations for the 104 languages yields phylogenetically plausible language groups. They argue that mBERT is not language-neutral and that semantic phenomena are not modeled properly across languages. In our analysis, we further quantify the representational distance and disentangle it along phylogenetic, geographic, and structural factors.

Bjerva and Augenstein (2018) train cross-lingual representations in an unsupervised way for different linguistic levels including phonology, morphology, and syntax. They use them to infer missing typological features for more than 800 languages. Malaviya et al. (2017) also infer missing features in typological databases from cross-lingual representations. Inferring such missing features can be considered a form of probing. Indeed, in contemporaneous work, Choenni and Shutova (2020) predict typological properties from representations of four different recent state-of-the-art cross-lingual encoders using probing classifiers. We do not use probing classifiers in our work because the choice of classifier and the size of its training data may affect the probing outcomes (Eger et al., 2020).

Table 1 summarizes selected interpretability approaches analyzing the typological signal in crosslingual representations. The findings for the dominant signal type vary strongly due to different choices for the representational model, the analysis unit, and the number of languages. Our work differs from previous work mainly in terms of the battery of tests probing for genetic and typological signals and the preciseness in teasing apart the different typological components.

3 Methodology

In the following, we first briefly describe the two cross-lingual embedding spaces analyzed in this work, mBERT and FastText (§3.1). Then, we detail how we compute distances between languages using representations from these spaces (§3.2) and concept lists developed in historical linguistics (§3.3). Once we have language distances, we infer trees from distance matrices and compare these trees to gold standard phylogenetic trees (§3.4) to evaluate how strong a historical linguistic tree signal is contained in our

Approach	Representational Target	Unit	# Languages	Dominant Signal
Malaviya et al. (2017)	translation	sentences	1,017	structural
Rabinovich et al. (2017)	syntax	sentences	17	phylogenetic
Bjerva et al. (2019)	syntax	sentences	20	structural
Eger et al. (2016)	semantics	words	21	geographical
Beinborn and Choenni (2019)	semantics	words	28	geographical
Libovický et al. (2019)	semantics	sentences	104	phylogenetic

Table 1: Summary of related work on typological interpretability of crosslingual representations.

cross-lingual representations (§4).

3.1 Cross-lingual Representations

We compare two different models: mBERT ia a cross-lingual model trained with a language-neutral contextualized objective and FastText stands for static monolingual word representations that have been aligned into a joint multilingual space.

mBERT mBERT is based on the multi-layer bidirectional transformer model BERT (Devlin et al., 2019). It is trained on the task of masked language modeling and next sentence prediction. The base model consists of 12 representational layers. mBERT is trained on the merged Wikipedias of 104 languages, with a shared word-piece vocabulary. It does not use explicit alignments across languages, thus has no mechanism to enforce that translation equivalent word pairs have similar representations. Recently, there has been a vivid debate regarding the quality of representations produced by mBERT. Pires et al. (2019) claim that it is surprisingly good at zero-shot cross-lingual transfer, and works best for structurally similar languages. Cao et al. (2020) find that mBERT exhibits vector space misalignment across languages and zero-shot cross-lingual transfer is improved after their suggested re-mapping. K et al. (2020) show that lexical overlap plays no big role in cross-lingual transfer for mBERT, but the depth of the network does, with deeper models having better transfer. Zhao et al. (2020b) find that mBERT lacks fine-grained cross-lingual text understanding and can be fooled by adversarial inputs produced by the corrupt input produced by MT systems.

FastText FastText (Bojanowski et al., 2017) builds static word representations on the basis of a word's characters. This allows it to induce better representations for infrequent and unknown words. We use a joint multilingually aligned vector space spanning 44 languages using the RCLS method described in Joulin et al. (2018) and refer to it as mFastText.[2]

3.2 Representational Distance

Assume we have M languages and N concepts (illustrated in Table 4 in the appendix). Assume further that each concept is expressed as a word in each language which is represented by a d-dimensional vector.

If all the vectors reside in a cross-lingually shared space, then the representational distance between two languages can be obtained by averaging the pairwise distances between all word vectors in the two languages for the N concepts. That means one computes:

$$\text{dist}(i, j) = \frac{1}{N} \sum_{k=1}^{N} d(\mathbf{v}_k(i), \mathbf{v}_k(j)) \tag{1}$$

where $\mathbf{v}_k(i)$ and $\mathbf{v}_k(j)$ stand for the vectors corresponding to the k-th concept for languages i and j, respectively (with words $v_k(i)$ and $v_k(j)$). In our experiments, we use cosine distance, but d may in principle refer to any suitable distance measure, e.g., Euclidean distance or Spearman correlation.[3],[4]

[2] `https://fasttext.cc/docs/en/aligned-vectors.html`

[3] Note that we compute the average of the distances, while it is also possible to compute the distance of the average representations (Libovický et al., 2019).

[4] An alternative to this direct comparison of the word vectors is a 'second-order' encoding where the representation $\hat{\mathbf{v}}_k(i)$ for a word is determined by the distances of its vector $\mathbf{v}_k(i)$ to the vectors for the N concepts (Eger and Mehler, 2016; Beinborn

When the corresponding words for each concept are not available in all languages, but only in one language (e.g., English), Beinborn and Choenni (2019) instead set $\mathbf{v}_k(i)$ to be the nearest neighbor of the English word for concept k in language i. This has the advantage that one can infer language distances without translation data in target languages. A drawback of this approach is that the relation between nearest neighbors in a vector space may not be that of similarity but of relatedness, e.g., *nose* is related to *mouth*, but it is not a synonym (meaning-equivalent).

In our experiments below, words for all concepts k are available in all languages, thus we do not need to resort to nearest neighbors of the English words.

3.3 Concept Lists

All our experiments are based on multilingual word lists obtained from linguistic databases. NorthEuraLex[5] features word lists for 1,016 concepts in 100 languages spoken in Northern Eurasia which have been transcribed by linguists (Dellert et al., 2020). The database is known for its high quality, but unfortunately covers only 54 of the 104 languages in mBERT. In order to analyze more languages, we additionally use PanLex[6] which contains lists for 207 concepts in more than 500 languages (Kamholz et al., 2014). It covers 99 languages in mBERT, but the quality of the word lists is not uniform across languages. PanLex sometimes includes multiple word lists written in different scripts for the same language, e.g. for Greek. In such a case, we include all available word lists for the language in our analysis.

3.4 Evaluating language trees

Historical differences between languages are commonly represented in phylogenetic trees which group languages by their evolution from common ancestors. We want to examine to which extent these phylogenetic differences can explain the observed representational distance in the cross-lingual model. We calculate all pairwise representational distances between languages as in Eq. (1). From this distance matrix, we infer a language tree using two inference techniques that are widely popular in computational biology for inferring species trees: 1) The unweighted pair group method with arithmetic mean (UPGMA) (Sokal and Michener, 1958) initially assumes that each language forms an individual cluster and then successively joins the two clusters with the smallest average distance. 2) The iterative Neighbor Joining (Saitou and Nei, 1987) algorithm starts with an unstructured star-like tree and iteratively adds nodes to create subtrees.

Reference tree Most automatic clustering methods produce binary trees due to computational simplifications whereas phylogenetic trees by linguistic experts are usually m-ary. To faciliate the evaluation of the inferred tree, previous work used a binary Levenshtein-based approximation (Serva and Petroni, 2008) as reference tree. This approximation provides an acceptable reference for a small subset of languages, but does not accurately reflect the more fine-grained differences for the Indo-European language family (Fortson, 2004). As we are evaluating a much larger set of languages here, we use the more reliable reference trees compiled by linguistic experts available in Glottolog (Hammarström et al., 2020).

Tree evaluation In order to compare the m-ary reference tree to the binary inferred tree, we apply a variant of quartet distance known as generalized quartet distance (Pompei et al., 2011). This metric evaluates the quality of the whole tree by comparing subgroups of four languages (quartets) which form so-called butterfly structures. A butterfly quartet refers to a quartet in which the four languages can be structured as two pairs of languages belonging to the same subfamily. For example, the pairs Spanish/Italian and Russian/Ukrainian form a butterfly structure whereas the four languages Hindi-German-

and Choenni, 2019):

$$\hat{\mathbf{v}}_k(i) = \Big(d(\mathbf{v}_k(i), \mathbf{v}_n(i))\Big)_{n\in[1,\dots,N]}$$

Then, $\hat{\mathbf{v}}_k(\cdot) \in \mathbb{R}^N$ while $\mathbf{v}_k(\cdot) \in \mathbb{R}^d$. These second-order vectors can be used in Eq. (1) to replace the original vectors $\mathbf{v}_k$.

[5] http://northeuralex.org/

[6] http://dev.panlex.org/db/panlex_swadesh.zip

Armenian-Latin all belong to different subgroups which are directly connected to the root node of the Indo-European tree. We evaluate our inferred tree by calculating the number of butterfly quartets which deviate from the reference tree normalized by the number of all butterfly quartets in the reference tree.

It is also possible to subject the distance matrix to other forms of clustering or dimensionality reduction techniques such as k-nearest neighbor, PCA, or t-sne (Maaten and Hinton, 2008). However, such flat clustering methods do not induce a tree structure and are not directly comparable to the reference trees of language families available in an online repository such as Glottolog (Hammarström et al., 2020).

4 Experiments

In the following, we detail a series of probing experiments with both mBERT and FastText. To extract representations from mBERT, we feed a *single word* $v_k(i)$ from a language i corresponding to concept k into mBERT and extract the corresponding representations $\mathbf{v}_k^{(r)}(i)$ in all layers $r = 0, \ldots, 12$. We are fully aware that using mBERT in a context-independent way ignores main benefits of the model. We do so in order to leverage concept lists at large scale for a majority of the 100 languages available in mBERT. Otherwise, we would have to experiment with sentence-aligned data, which is available only for much smaller subsets (< 30) of our languages. Nevertheless, we believe that a good contextual model should also be equipped with good context-independent token representations.

4.1 Cross-lingual Semantics

Monolingual language models are commonly evaluated by their ability to model semantic similarity and their performance on downstream tasks. For multilingual models, a suitable evaluation task for lexical semantics is bilingual lexicon induction (BLI). The goal is to take an input word in the source language and retrieve its translation-equivalent in the target language. In a decontextualized setting, multiple targets can be considered to be a correct translation due to the polysemy of words. As the word lists only account for a single correct solution, we cast bilingual lexicon induction as a ranking task. We rank all target words in the concept list based on their representational distance to the source word in the model and evaluate this ranking using the mean reciprocal rank (MRR) as proposed in Glavaš et al. (2019). The MRR ranges from 0 to 1; a value of 1 indicates that the target is correctly ranked on 1, a value of $1/n$ indicates that the target is on averaged ranked on n.

For the PanLex list of 207 words, the MRR obtained by a random baseline would be 0.03. The result of mBERT for the language pair (`bos`,`hrv`) is almost perfect with an MRR of 0.98, but for pairs of more distant languages the BLI quality is considerably lower. Overall, the average performance for mBERT (0.16) is five times better than random guessing, but consistently lower than the performance for mFastText (0.46 on average).[7] Overall, this shows that mBERT does not properly capture multilingual semantics, a finding that is echoed in some other recent works (Cao et al., 2020; Zhao et al., 2020b). The apparent reason lies in its naive training process, which does not exploit cross-lingual signals but merely trains on the concatenation of all languages. Nonetheless, the model performs surprisingly well in some downstream cross-lingual tasks (Pires et al., 2019). In the following experiments, we examine whether the model instead relies on typological properties of languages.

4.2 Phylogenetic signal

We perform tree inference using both Neighbor-Joining (NJ) and UPGMA on the concept lists from PanLex (99 languages) and NorthEuraLex (54 languages). We infer trees for all the layers of mBERT and evaluate the quality of the inferred trees as described in Section 3.4. Table 2 shows that especially the initial-middle and the final layers of mBERT yield a small distance to the gold standard trees.

Overall, however, the strength of the phylogenetic signal varies with respect to the selected concept list and the tree inference algorithm. Interestingly, the results for the PanLex word list are better although this setup covers more languages. UPGMA yields lower distances to the gold tree for both concept lists. In comparison, the results for mFastText are considerably worse when using UPGMA (>0.5), but

[7] Only for 12 language pairs, the MRR is higher for mBERT than for mFastText

comparable when using NJ (around 0.32). It should be noted though that mFastText covers only 44 languages.

Method	Word List	0	1	2	3	4	5	6	7	8	9	10	11	12	Avg.
UPGMA	PanLex	.34	.30	**.17**	.18	.21	.26	.28	.20	.23	.21	.22	.23	.21	.20
	NorthEuralex	.43	.29	**.26**	.28	.30	.31	.31	.35	.34	.32	.37	.34	.32	.31
NJ	PanLex	.38	.31	.30	.30	.26	.31	**.25**	.32	.32	.32	.35	.34	.30	.30
	NorthEuralex	.41	.36	.35	.32	.32	**.31**	.31	.32	.32	.32	.32	.32	.40	.37

Table 2: Distances between the Glottolog reference tree and the phylogenetic tree inferred from mBERT representations from the 12 different layers and the average of all layers. Generalized quarted distances range between 0 and 1, lower distances are better.

Visual exploration In order to qualitatively analyze representational distances, we subject the distance matrix from the second layer (which yields the best scores according to UPGMA) to the t-sne algorithm as in previous work (Libovický et al., 2019; Kudugunta et al., 2019). The visualization in Figure 1 clearly shows a mix of phylogenetic and other clusters. Instances of phylogenetic clusters include separation of Germanic languages (excluding English which is placed apart) and Romance languages in the lower left. In contrast, the three Dravidian languages (Tamil, Malayalam,Telugu) are placed on the right most part of the plot together with Hindi and Bengali (Indo-European languages), illustrating more of a geographical similarity. The Slavic languages show up in two clusters: 1) Western Slavic Languages (Polish, Czech, and Slovak in the lower half) 2) Eastern Slavic languages such as Russian and Ukranian together with Turkic languages such as Azeri and Kazakh written in Cyrllic script. At the same time, the different word lists of Azeri (written in different scripts) are placed together, suggesting that mBERT representations are also script-agnostic. Uralic languages such as Finnish and Estonian are closer to the other Baltic languages which are not clustered together with the other Slavic languages. These clusters cannot be sufficiently explained solely by phylogenetic properties.

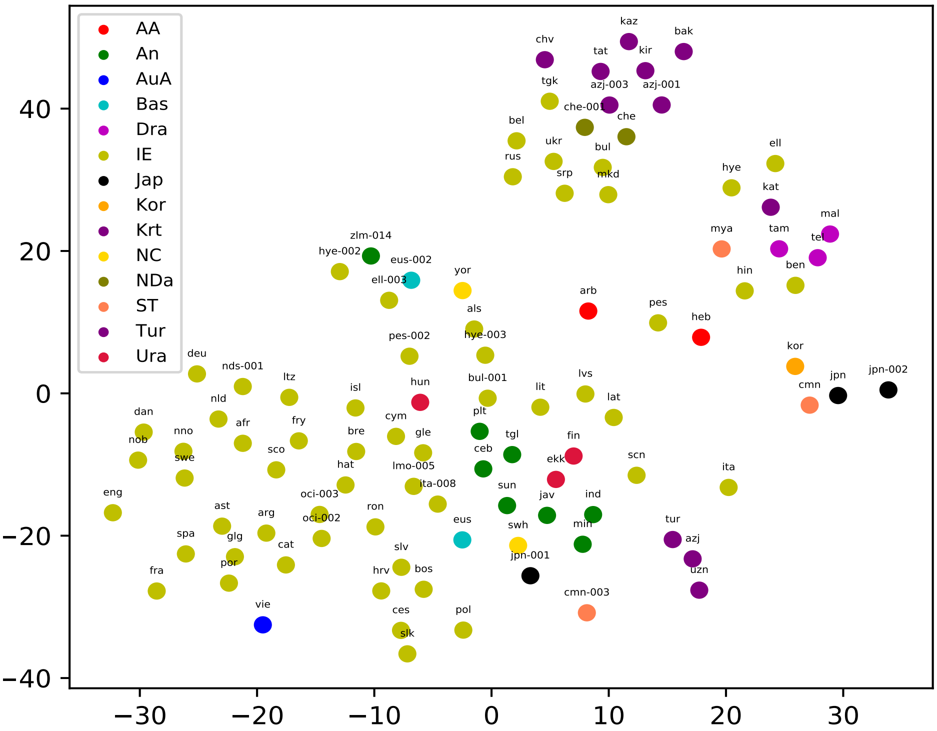


Figure 1: The t-sne plot for the Swadesh list distances from layer 2. The family codes are from the ASJP database (Wichmann et al., 2020) and are explained in Table 5 in appendix.

4.3 Other typological signals

In order to better disentangle the typological signal, we examine additional categories established by Littell et al. (2017). We determine the explainable predictors for the representational distances between languages using matrix regression (Legendre et al., 1994). We regress $d_{ij} = \text{dist}(i, j)$ computed based on Eq. (1) on the following language distances:

- Phylogenetic distance (gen_{ij}) between two languages computed from Glottolog reference trees as the ratio between the number of non-shared branches divided by the number of branches from root to the tip.
- Geographical distance (geo_{ij}) between two points on Earth approximated through great circle distance (Department, 1997).
- Structural distance (struc_{ij}): Cosine distance computed over averaged syntactic features from WALS[8], SSWL[9], and mini-grammars parsed from Ethnologue (Lewis, 2009).
- Phonological distance (phon_{ij}): Cosine distance computed over averaged phonological features available in WALS and Ethnologue.
- Phoneme Inventory distance (inv_{ij}): Cosine distance computed between binary feature vectors as given in the PHOIBLE database (Moran and McCloy, 2019) which consist of features such as presence or absence of retroflex sounds in a language.

For detailed description of the distances, see Littell et al. (2017).[10] In our experiments, we use the precomputed distance matrices provided along with the `lang2vec` Python package[11]. We estimate the coefficients as follows:

$$d_{ij} = c + \alpha \cdot \text{gen}_{ij} + \beta \cdot \text{geo}_{ij} + \gamma \cdot \text{struc}_{ij} + \eta \cdot \text{phon}_{ij} + \lambda \cdot \text{inv}_{ij} \tag{2}$$

Since the entries in the distance matrices are non-independent, we use matrix regression analysis (Legendre et al., 1994) as implemented in the R package `ecodist` (Goslee et al., 2007) for computing the regression coefficients. The significance of the regression coefficients is also tested using a Mantel test where the matrix columns are permuted 10^5 times.

The significant regression coefficients ($p < 0.001$) and their sizes are shown in Figure 2. The phylogenetic signal in mBERT is stronger than the geographical signal and this is especially true for PanLex where the geographical signal is never significant. The genetic signal is more prominent in the initial layers. It then decreases over layers, but re-emerges in the final layer. As we use isolated concepts in our setup, we did not expect structural features to be a significant predictor at all and are surprised about the PanLex results. We further hypothesized that phonological distances might be a weak but significant predictor due to related words (both cognates and borrowings) and shared scripts (also showing up in the t-sne plot in Figure 1) which is not supported in our experiments.

Overall, we find that the R^2 values (see Table 3) from the regression analyses are significant across all the layers for both the datasets, but they are not very large. This suggests that there exist other factors explaining the representational distances that our equation does not account for or that the linear model is not fully appropriate. In the case of the mFastText model, the R^2 value is at about 0.24 but none of the regression coefficients are significant.

4.4 Meaning Stability

We calculate cross-lingual variability in mBERT by considering the pairwise cosine similarities between representations in languages i and j for all concepts k in PanLex:

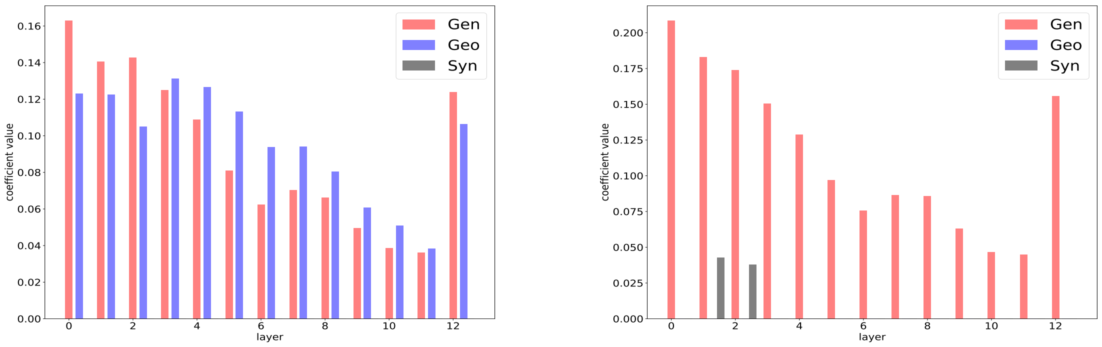
$$c_{ij}(k) = \text{cossim}(\mathbf{v}_k(i), \mathbf{v}_k(j)) \tag{3}$$

[8] https://wals.info/

[9] http://test.terraling.com/groups/7

[10] http://www.cs.cmu.edu/~dmortens/uriel.html

[11] https://github.com/antonisa/lang2vec



(a) NorthEuralex

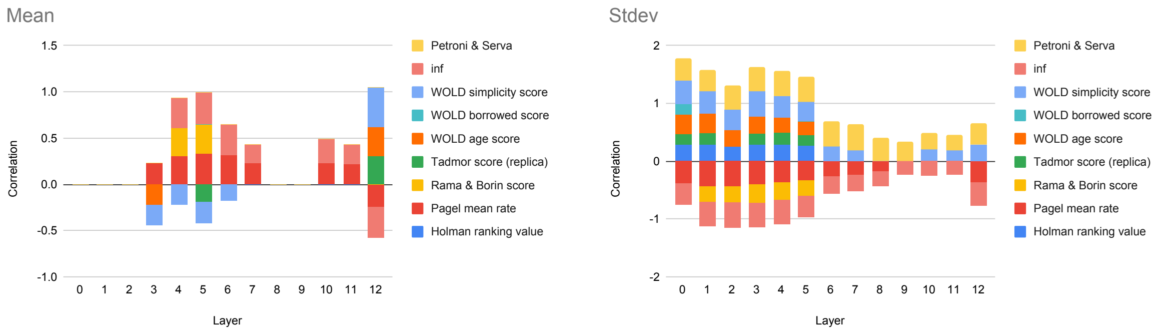
(b) PanLex

Figure 2: The coefficient values (α, β, γ) from Eq. (2) for each mBERT layer.

	0	1	2	3	4	5	6	7	8	9	10	11	12	Avg.
PanLex	.39	.39	**.41**	.37	.34	.28	.27	.29	.29	.25	.23	.25	.36	.37
NorthEuralex	.39	.45	.48	**.55**	.54	.50	.46	.43	.40	.37	.36	.30	.40	.48

Table 3: R^2 values from the regression analyses for each layer.

and then analyzing how these cosine similarities vary. We conjecture that variability of mBERT representations of a concept k across languages is indicative of its diachronic stability, where diachronic meaning stability measures the resistance of a concept to lexical replacement.



(a) Mean

(b) Standard Deviation

Figure 3: Correlation between variability of cross-lingual representations from mBERT and diachronic stability lists (significant at $p < .01$).

To capture cross-lingual variability, we calculate statistics s on the $c_{ij}(k)$ values. As statistics, we consider the mean and standard deviation values of $c_{ij}(k)$ for each fixed concept k. Thus, each concept k receives a 'variability score' $s(k)$ given by s applied to the $c_{ij}(k)$ values: $s(k) = s\left(c_{12}(k), c_{13}(k), \dots, c_{1M}(k), \dots\right)$, where M is the number of languages involved. Note that the standard deviation statistic (intuitively) captures the notion of cross-lingual variability while the mean statistic measures the average degree of similarity of the representations of words for a target concept across languages.

We finally correlate the statistics with diachronic stability scores for concepts k extracted from the following lists: (i) ranked list of Holman et al. (2008); (ii) WOLD's (World Loanword Database; Haspelmath and Tadmor (2009)) meaning scores for age, simplicity, and borrowing; (iii) 100-item Leipzig-Jakarta list (Tadmor, 2009) and its replication (LJ-replica) based on later versions of the WOLD scores

(Dellert and Buch, 2018); (iv) Swadesh list ranked by word replacement rates computed from a phylogenetic analysis (Pagel et al., 2007); (v) n-gram entropy based stability measure (Rama and Borin, 2014); (vi) *inf*, a information-theoretic weighted string similarity (Dellert and Buch, 2018); (vii) and a Levenshtein distance based measure (Petroni and Serva, 2011). All the ranked lists are drawn from Dellert and Buch (2018).

Figure 3a shows the correlation between $s(k)$ and diachronic stability when the statistic is the mean. It can be seen that Layers 3–7 and 10–11 correlate positively with Pagel et al. (2007) and *inf*. This suggests that the mean statistic measures susceptibility to replacement as opposed to stability since in both the lists, the higher the score, the lower is a concept's stability. Layer 12 correlates with 5 of the 9 lists and shows positive correlation with both WOLD's age and simplicity scores and LJ-replica. In contrast, layer 12 correlates negatively with both Pagel et al. (2007) and *inf*, suggesting that the measure is inconsistent in the layer. The mean statistic is also not consistent across the layers and shows inverse correlations with *inf* and Pagel et al. (2007) in layer 12 compared to the other layers.

The correlations from standard deviation statistic is given in Figure 3b. Here, layers 0–5 show correlations with 8 of the 9 ranked lists. In layers 0–5, the upper half of the figure has positive correlations with word lists ranked by decreasing order of stability (LJ-replica, Petroni and Serva (2011) and Holman et al. (2008)) whereas the lower half of the figure correlates negatively with the rankings of Pagel et al. (2007), *inf* and Rama and Borin (2014) where a lower score for a meaning indicates higher stability. Layers 0–7 & 10–12 correlate positively with WOLD indices such as age and simplicity. There is a negative correlation with *inf* and a positive correlation with the measure of Petroni and Serva (2011) across all the layers. The standard deviation statistic is consistent across all the layers in terms of correlations against the 9 lists.

We conclude from this experiment that cross-lingual variability of representations in mBERT, as measured by standard deviation,[12] indeed correlates with diachronic (cross-temporal) stability as given by proposed historical linguistic indicators.

5 Concluding remarks

We applied a series of tests for probing mBERT for typological signals. While the language trees inferred from mBERT representations are sometimes close to the reference trees, they may confound multiple factors. A more-fined grained investigation of t-sne plots followed by matrix regression analyses suggests that representational distances correlate most with phylogenetic and geographical distances between languages. Further, the rankings from cross-lingual stability scores correlate significantly with meaning lists for items supposed to be resistant to cross-temporal lexical replacement.

Our results contribute to the recent discourses on interpretability and introspection of black-box NLP representations (Conneau et al., 2018; Kudugunta et al., 2019; Jacovi and Goldberg, 2020). In our case, we asked how mBERT perceives of the similarity of two languages and related this to phylogenetic, geographic and structural factors. In future work, we aim to use our inferred similarities to predict transfer behavior in downstream tasks between specific language pairs.

Finally, we strongly caution against using our conclusions as support for hypotheses relating semantics and language phylogeny (e.g., the Sapir-Whorf hypothesis). Our results for bilingual lexicon induction indicate that mBERT representations are only mildly semantic cross-lingually which corroborates similar findings in related work.

Acknowledgements

The last author has been funded by the HMWK (Hessisches Ministerium für Wissenschaft und Kunst) as part of structural location promotion for TU Darmstadt in the context of the Hessian excellence cluster initiative "Content Analytics for the Social Good" (CA-SG).

[12] Like the mean, other statistics, such as minimum and maximum, did also not exhibit significant correlations. We found significant correlations only for the standard deviation.

References

Samira Abnar, Lisa Beinborn, Rochelle Choenni, and Willem Zuidema. 2019. Blackbox meets blackbox: Representational similarity and stability analysis of neural language models and brains. In *Proceedings of the ACL-Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 191–203.

Afra Alishahi, Grzegorz Chrupała, and Tal Linzen. 2019. Analyzing and interpreting neural networks for NLP: A report on the first BlackboxNLP workshop. *Natural Language Engineering*, 25(4):543–557.

Mikel Artetxe and Holger Schwenk. 2018. Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond. *arXiv e-prints*, page arXiv:1812.10464.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2017. Learning bilingual word embeddings with (almost) no bilingual data. In *ACL*, pages 451–462.

Lisa Beinborn and Rochelle Choenni. 2019. Semantic drift in multilingual representations. *arXiv preprint arXiv:1904.10820*.

Yonatan Belinkov and James Glass. 2019. Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 7:49–72.

Johannes Bjerva and Isabelle Augenstein. 2018. 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.

Johannes Bjerva, Robert Östling, Maria Han Veiga, Jörg Tiedemann, and Isabelle Augenstein. 2019. What do language representations really represent? *Computational Linguistics*, pages 381–389.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *TACL*, 5:135–146.

Steven Cao, Nikita Kitaev, and Dan Klein. 2020. Multilingual alignment of contextual word representations. In *International Conference on Learning Representations*.

Rochelle Choenni and Ekaterina Shutova. 2020. What does it mean to be language-agnostic? probing multilingual sentence encoders for typological properties.

Grzegorz Chrupała, Lieke Gelderloos, and Afra Alishahi. 2017. Representations of language in a model of visually grounded speech signal. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 613–622, Vancouver, Canada, July. Association for Computational Linguistics.

Alexis Conneau and Douwe Kiela. 2018. SentEval: An evaluation toolkit for universal sentence representations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680, Copenhagen, Denmark, September. Association for Computational Linguistics.

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

Johannes Dellert and Armin Buch. 2018. A new approach to concept basicness and stability as a window to the robustness of concept list rankings. *Language Dynamics and Change*, 8(2):157–181.

Johannes Dellert, Thora Daneyko, Alla Münch, Alina Ladygina, Armin Buch, Natalie Clarius, Ilja Grigorjew, Mohamed Balabel, Hizniye Isabella Boga, Zalina Baysarova, Roland Mühlenbernd, Johannes Wahle, and Gerhard Jäger. 2020. NorthEuraLex: a wide-coverage lexical database of Northern Eurasia. *Lang. Resour. Evaluation*, 54(1):273–301.

Great Britain. Navy Department. 1997. *Admiralty Manual of Navigation: BR 45(1)*. Number Bd. 1 in BR Series. Stationery Office.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, pages 4171–4186, June.

Prajit Dhar and Arianna Bisazza. 2018. Does syntactic knowledge in multilingual language models transfer across languages? In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 374–377, Brussels, Belgium, November. Association for Computational Linguistics.

Steffen Eger and Alexander Mehler. 2016. On the linearity of semantic change: Investigating meaning variation via dynamic graph models. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 52–58, Berlin, Germany, August. Association for Computational Linguistics.

Steffen Eger, Armin Hoenen, and Alexander Mehler. 2016. Language classification from bilingual word embedding graphs. In *COLING*, pages 3507–3518.

Steffen Eger, Johannes Daxenberger, and Iryna Gurevych. 2020. How to probe sentence embeddings in low-resource languages: On structural design choices for probing task evaluation. In *CONLL*.

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

Benjamin F. Fortson, IV. 2004. *Indo-European language and culture: an introduction*, volume 19 of *Blackwell Textbooks in Linguistics*. Blackwell, Oxford.

Mario Giulianelli, Jack Harding, Florian Mohnert, Dieuwke Hupkes, and Willem Zuidema. 2018. Under the hood: Using diagnostic classifiers to investigate and improve how language models track agreement information. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 240–248, Brussels, Belgium, November. Association for Computational Linguistics.

Goran Glavaš, Robert Litschko, Sebastian Ruder, and Ivan Vulić. 2019. How to (properly) evaluate cross-lingual word embeddings: On strong baselines, comparative analyses, and some misconceptions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 710–721, Florence, Italy, July. Association for Computational Linguistics.

Sarah C Goslee, Dean L Urban, et al. 2007. The ecodist package for dissimilarity-based analysis of ecological data. *Journal of Statistical Software*, 22(7):1–19.

Simon J Greenhill, Quentin D Atkinson, Andrew Meade, and Russell D Gray. 2010. The shape and tempo of language evolution. *Proceedings of the Royal Society B: Biological Sciences*, 277(1693):2443–2450.

Harald Hammarström, Robert Forkel, Martin Haspelmath, and Sebastian Bank. 2020. Glottolog 4.2.1. Max Planck Institute for the Science of Human History.

Martin Haspelmath and Uri Tadmor, editors. 2009. *WOLD*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Eric W Holman, Søren Wichmann, Cecil H Brown, Viveka Velupillai, André Müller, and Dik Bakker. 2008. Explorations in automated language classification. *Folia Linguistica*, 42(3-4):331–354.

Dieuwke Hupkes, Sara Veldhoen, and Willem Zuidema. 2018. Visualisation and 'diagnostic classifiers' reveal how recurrent and recursive neural networks process hierarchical structure. *Journal of Artificial Intelligence Research*, 61:907–926.

Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable nlp systems: How should we define and evaluate faithfulness?

Gerhard Jäger. 2014. Phylogenetic inference from word lists using weighted alignment with empirically determined weights. In *Language Dynamics and Change*, pages 155–204. Brill.

Armand Joulin, Piotr Bojanowski, Tomas Mikolov, Hervé Jégou, and Edouard Grave. 2018. Loss in translation: Learning bilingual word mapping with a retrieval criterion. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

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

David Kamholz, Jonathan Pool, and Susan M Colowick. 2014. Panlex: Building a resource for panlingual lexical translation. In *LREC*, pages 3145–3150.

Alexandre Klementiev, Ivan Titov, and Binod Bhattarai. 2012. Inducing crosslingual distributed representations of words. In *Proceedings of COLING 2012*, pages 1459–1474, Mumbai, India, December. The COLING 2012 Organizing Committee.

Katarzyna Krasnowska-Kieraś and Alina Wróblewska. 2019. Empirical linguistic study of sentence embeddings. In *ACL*, pages 5729–5739.

Sneha Kudugunta, Ankur Bapna, Isaac Caswell, and Orhan Firat. 2019. Investigating multilingual NMT representations at scale. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1565–1575, Hong Kong, China, November. Association for Computational Linguistics.

Pierre Legendre, François-Joseph Lapointe, and Philippe Casgrain. 1994. Modeling brain evolution from behavior: a permutational regression approach. *Evolution*, 48(5):1487–1499.

M. Paul Lewis, editor. 2009. *Ethnologue: Languages of the World*. SIL International, Dallas, TX, USA, sixteenth edition.

Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, Xiaodong Fan, Ruofei Zhang, Rahul Agrawal, Edward Cui, Sining Wei, Taroon Bharti, Ying Qiao, Jiun-Hung Chen, Winnie Wu, Shuguang Liu, Fan Yang, Daniel Campos, Rangan Majumder, and Ming Zhou. 2020. Xglue: A new benchmark dataset for cross-lingual pre-training, understanding and generation. *arXiv*, abs/2004.01401.

Jindřich Libovickỳ, Rudolf Rosa, and Alexander Fraser. 2019. How language-neutral is multilingual bert? *arXiv preprint arXiv:1911.03310.*

Patrick Littell, David R Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. 2017. Uriel and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 8–14.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605.

Chaitanya Malaviya, 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.

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

Mark Pagel, Quentin D Atkinson, and Andrew Meade. 2007. Frequency of word-use predicts rates of lexical evolution throughout indo-european history. *Nature*, 449(7163):717–720.

Filippo Petroni and Maurizio Serva. 2011. Automated word stability and language phylogeny. *Journal of Quantitative Linguistics*, 18(1):53–62.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy, July. Association for Computational Linguistics.

Simone Pompei, Vittorio Loreto, and Francesca Tria. 2011. On the accuracy of language trees. *PloS one*, 6(6).

Ella Rabinovich, Noam Ordan, and Shuly Wintner. 2017. Found in translation: Reconstructing phylogenetic language trees from translations. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 530–540.

Taraka Rama and Lars Borin. 2014. N-gram approaches to the historical dynamics of basic vocabulary. *Journal of Quantitative Linguistics*, 21(1):50–64.

Taraka Rama and Lars Borin. 2015. Comparative evaluation of string similarity measures for automatic language classification. In Ján Mačutek and George K. Mikros, editors, *Sequences in Language and Text*, pages 203–231. Walter de Gruyter.

Naruya Saitou and Masatoshi Nei. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular biology and evolution*, 4(4):406–425.

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

R. R. Sokal and C. D. Michener. 1958. A statistical method for evaluating systematic relationships. *University of Kansas Science Bulletin*, 38:1409–1438.

Uri Tadmor. 2009. Loanwords in the world's languages. findings and results. In Martin Haspelmath and Uri Tadmor, editors, *Loanwords in the world's languages. A comparative handbook*, pages 55–75. de Gruyter, Berlin and New York.

Zirui Wang, Jiateng Xie, Ruochen Xu, Yiming Yang, Graham Neubig, and Jaime G. Carbonell. 2020. Cross-lingual alignment vs joint training: A comparative study and a simple unified framework. In *International Conference on Learning Representations*.

Leila Wehbe, Ashish Vaswani, Kevin Knight, and Tom Mitchell. 2014. Aligning context-based statistical models of language with brain activity during reading. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 233–243, Doha, Qatar, October. Association for Computational Linguistics.

Søren Wichmann and Arpiar Saunders. 2007. How to use typological databases in historical linguistic research. *Diachronica*, 24(2):373–404.

Søren Wichmann, Eric W Holman, and Cecil H Brown. 2020. The ASJP database (version 19). *Jena: Max Planck Institute for the Science of Human History*.

Kelly Zhang and Samuel Bowman. 2018. Language modeling teaches you more than translation does: Lessons learned through auxiliary syntactic task analysis. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 359–361.

Wei Zhao, Steffen Eger, Johannes Bjerva, and Isabelle Augenstein. 2020a. Inducing language-agnostic multilingual representations.

Wei Zhao, Goran Glavaš, Maxime Peyrard, Yang Gao, Robert West, and Steffen Eger. 2020b. On the limitations of cross-lingual encoders as exposed by reference-free machine translation evaluation. In *ACL*.

A Example list

	English	German	French	···
I	I	ich	je	···
YOU	you	du	tu	···
HE	he	er	il	···
MAN	man	Mann	homme	···
⋮				

Table 4: Schematic illustration of concept lists.

B Family codes list

Family Code	Family Name
AA	Afro-Asiatic
An	Austronesian
AuA	Austro-Asiatic
Bas	Basque
Dra	Dravidian
IE	Indo-European
Jap	Japonic
Kor	Koreanic
Krt	Kartvelian
NC	Niger-Congo
NDa	Nakh-Dagestan
ST	Sino-Tibetan
Tur	Turkic
Ura	Uralic

Table 5: Family codes list