Finding Universal Grammatical Relations in Multilingual BERT

Ethan A. Chi, John Hewitt, and **Christopher D. Manning**
Department of Computer Science
Stanford University
{ethanchi,johnhew,manning}@cs.stanford.edu

Abstract

Recent work has found evidence that Multilingual BERT (mBERT), a transformer-based multilingual masked language model, is capable of zero-shot cross-lingual transfer, suggesting that some aspects of its representations are shared cross-lingually. To better understand this overlap, we extend recent work on finding syntactic trees in neural networks' internal representations to the multilingual setting. We show that subspaces of mBERT representations recover syntactic tree distances in languages other than English, and that these subspaces are approximately shared across languages. Motivated by these results, we present an unsupervised analysis method that provides evidence mBERT learns representations of syntactic dependency labels, in the form of clusters which largely agree with the Universal Dependencies taxonomy. This evidence suggests that even without explicit supervision, multilingual masked language models learn certain linguistic universals.

1 Introduction

Past work (Liu et al., 2019; Tenney et al., 2019a,b) has found that masked language models such as BERT (Devlin et al., 2019) learn a surprising amount of linguistic structure, despite a lack of direct linguistic supervision. Recently, large multilingual masked language models such as Multilingual BERT (mBERT) and XLM (Conneau and Lample, 2019; Conneau et al., 2019) have shown strong *cross-lingual* performance on tasks like XNLI (Lample and Conneau, 2019; Williams et al., 2018) and dependency parsing (Wu and Dredze, 2019). Much previous analysis has been motivated by a desire to explain why BERT-like models perform so well on downstream applications in the monolingual setting, which begs the question: what properties of these models make them so cross-lingually effective?

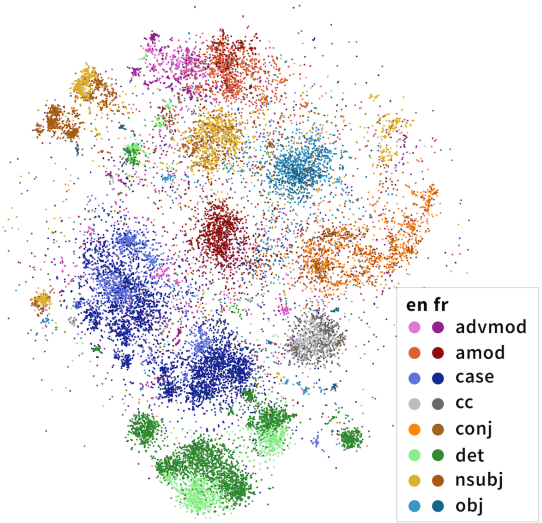


Figure 1: t-SNE visualization of head-dependent dependency pairs belonging to selected dependencies in English and French, projected into a syntactic subspace of Multilingual BERT, as learned on English syntax trees. Colors correspond to gold UD dependency type labels. Although neither mBERT nor our probe was ever trained on UD dependency labels, English and French dependencies exhibit cross-lingual clustering that largely agrees with UD dependency labels.

In this paper, we examine the extent to which Multilingual BERT learns a cross-lingual representation of syntactic structure. We extend probing methodology, in which a simple supervised model is used to predict linguistic properties from a model's representations. In a key departure from past work, we not only evaluate a probe's performance (on recreating dependency tree structure), but also use the probe as a window into understanding aspects of the representation that the probe was not trained on (i.e. dependency labels; Figure 1). In particular, we use the *structural probing* method of Hewitt and Manning (2019), which probes for syntactic trees by finding a linear transformation under which two words' distance in their dependency parse is approximated by the squared

distance between their model representation vectors under a linear transformation. After evaluating whether such transformations recover syntactic tree distances across languages in mBERT, we turn to analyzing the transformed vector representations themselves.

We interpret the linear transformation of the structural probe as defining a *syntactic subspace* (Figure 2), which intuitively may focus on syntactic aspects of the mBERT representations. Since the subspace is optimized to recreate syntactic tree distances, it has no supervision about edge labels (such as *adjectival modifier* or *noun subject*). This allows us to unsupervisedly analyze how representations of head-dependent pairs in syntactic trees cluster and qualitatively discuss how these clusters relate to linguistic notions of grammatical relations.

We make the following contributions:

- We find that structural probes extract considerably more syntax from mBERT than baselines in 10 languages, extending the structural probe result to a multilingual setting.
- We demonstrate that mBERT represents some syntactic features in syntactic subspaces that overlap between languages. We find that structural probes trained on one language can recover syntax in other languages (zero-shot), demonstrating that the syntactic subspace found for each language picks up on features that BERT uses across languages.
- Representing a dependency by the difference of the head and dependent vectors in the syntactic space, we show that mBERT represents dependency clusters that largely overlap with the dependency taxonomy of Universal Dependencies (UD) (Nivre et al., 2020); see Figure 1. Our method allows for fine-grained analysis of the distinctions made by mBERT that disagree with UD, one way of moving past probing's limitation of detecting only linguistic properties we have training data for rather than properties inherent to the model.

Our analysis sheds light on the cross-lingual properties of Multilingual BERT, through both zero-shot cross-lingual structural probe experiments and novel unsupervised dependency label discovery experiments which treat the probe's syntactic subspace as an object of study. We find evidence that mBERT induces universal grammatical relations without any explicit supervision, which largely agree with the dependency labels of Universal Dependencies.[1]

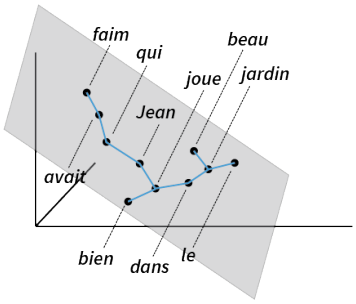


Figure 2: The structural probe recovers syntax by finding a syntactic subspace in which all syntactic trees' distances are approximately encoded as squared L_2 distance (Hewitt and Manning, 2019).

2 Methodology

We present a brief overview of Hewitt and Manning (2019)'s structural probe, closely following their derivation. The method represents each dependency tree T as a distance metric where the distance between two words $d_T(w_i, w_j)$ is the number of edges in the path between them in T. It attempts to find a single linear transformation of the model's word representation vector space under which squared distance recreates tree distance in any sentence. Formally, let $\mathbf{h}^\ell_{1:n}$ be a sequence of n representations produced by a model from a sequence of n words $w^\ell_{1:n}$ composing sentence ℓ. Given a matrix $B \in \mathbb{R}^{k \times m}$ which specifies the probe parameters, we define a squared distance metric d_B as the squared L_2 distance after transformation by B:

$$d_B(\mathbf{h}^\ell_i, \mathbf{h}^\ell_j) = ||B\mathbf{h}^\ell_i - B\mathbf{h}^\ell_j||^2_2$$

We optimize to find a B that recreates the tree distance d_{T^ℓ} between all pairs of words (w^ℓ_i, w^ℓ_j) in all sentences s^ℓ in the training set of a parsed corpus. Specifically, we optimize by gradient descent:

$$\arg\min_B \sum_\ell \frac{1}{|s^\ell|^2} \sum_{i,j} |d_{T^\ell}(w^\ell_i, w^\ell_j) - d_B(\mathbf{h}^\ell_i, \mathbf{h}^\ell_j)|$$

For more details, see Hewitt and Manning (2019).

Departing from prior work, we view the probe-transformed word vectors $B\mathbf{h}$ themselves—not just the distances between them—as objects of study.

[1]Code for reproducing our experiments is available here: https://github.com/ethanachi/multilingual-probing-visualization

The rows of B are a basis that defines a subspace of $\mathbb{R}^m$, which we call the *syntactic subspace*, and may focus only on parts of the original BERT representations. A vector $B\mathbf{h}$ corresponds to a point in that space; the value of each dimension equals the dot product of $\mathbf{h}$ with one of the basis vectors.[2]

2.1 Experimental Settings

These settings apply to all experiments using the structural probe throughout this paper.

Data Multilingual BERT is pretrained on corpora in 104 languages; however, we probe the performance of the model in 11 languages (Arabic, Chinese, Czech, English, Farsi, Finnish, French, German, Indonesian, Latvian, and Spanish).[3,4] Specifically, we probe the model on trees encoded in the Universal Dependencies v2 formalism (Nivre et al., 2020).

Model In all our experiments, we investigate the 110M-parameter pre-trained weights of the BERT-Base, Multilingual Cased model.[5]

Baselines We use the following baselines:[6]

- **MBERTRAND**: A model with the same parametrization as mBERT but no training. Specifically, all of the contextual attention layers are reinitialized from a normal distribution with the same mean and variance as the original parameters. However, the subword embeddings and positional encoding layers remain unchanged. As randomly initialized ELMo layers are a surprisingly competitive baseline for syntactic parsing (Conneau et al., 2018), we also expect this to be the case for BERT. In our experiments, we find that this baseline performs approximately equally across layers, so we draw always from Layer 7.

- **LINEAR**: All sentences are given an exclusively left-to-right chain dependency analysis.

EVALUATION To evaluate transfer accuracy, we use both of the evaluation metrics of Hewitt and Manning (2019). That is, we report the Spearman correlation between predicted and true word pair distances (DSpr.).[7] We also construct an undirected minimum spanning tree from said distances, and evaluate this tree on undirected, unlabeled attachment score (UUAS), the percentage of undirected edges placed correctly when compared to the gold tree.

3 Does mBERT Build a Syntactic Subspace for Each Language?

We first investigate whether mBERT builds syntactic subspaces, potentially private to each language, for a subset of the languages it was trained on; this is a prerequisite for the existence of a *shared*, cross-lingual syntactic subspace.

Specifically, we train the structural probe to recover tree distances in each of our eleven languages. We experiment with training syntactic probes of various ranks, as well as on embeddings from all 12 layers of mBERT.

3.1 Results

We find that the syntactic probe recovers syntactic trees across all the languages we investigate, achieving on average an improvement of 22 points UUAS and 0.175 DSpr. over both baselines (Table 1, section IN-LANGUAGE).[8]

Additionally, the probe achieves significantly higher UUAS (on average, 9.3 points better on absolute performance and 6.7 points better on improvement over baseline) on Western European languages.[9]. Such languages have been shown to have better performance on recent shared task results on multilingual parsing (e.g. Zeman et al., 2018). However, we do not find a large improvement when evaluated on DSpr. (0.041 DSpr. absolute, -0.013 relative).

We find that across all languages we examine, the structural probe most effectively recovers tree structure from the 7th or 8th mBERT layer (Figure 4). Furthermore, increasing the probe maximum rank beyond approximately 64 or 128 gives

[2] For ease of notation, we will discuss vectors $B\mathbf{h}$ as being in the syntactic subspace, despite being in $\mathbb{R}^k$.

[3] When we refer to *all languages*, we refer to all languages in this set, not all languages that mBERT trains on.

[4] This list is not typologically representative of all human languages. However, we are constrained by the languages for which both large UD datasets and mBERT's pretraining are available. Nevertheless, we try to achieve a reasonable spread over language families, while also having some pairs of close languages for comparison.

[5] https://github.com/google-research/bert

[6] We omit a baseline that uses uncontextualized word embeddings because Hewitt and Manning (2019) found it to be a weak baseline compared to the two we use.

[7] Following Hewitt and Manning (2019), we evaluate only sentences of lengths 5 to 50, first average correlations for word pairs in sentences of a specific length, and then average across sentence lengths.

[8] Throughout this paper, we report improvement over the stronger of our two baselines per-language.

[9] Here, we define Western European as Czech, English, French, German, and Spanish.

Structural Probe Results: Undirected Unlabeled Attachment Score (UUAS)

	Arabic	Czech	German	English	Spanish	Farsi	Finnish	French	Indonesian	Latvian	Chinese	Average
LINEAR	57.1	45.4	42.8	41.5	44.6	52.6	50.1	46.4	55.2	47.0	44.2	47.9
MBERTRAND	49.8	57.3	55.2	57.4	55.3	43.2	54.9	61.2	53.2	53.0	41.1	52.9
IN-LANG	72.8	83.7	83.4	80.1	79.4	70.7	76.3	81.3	74.4	77.1	66.3	76.8
Δ_{BASELINE}	15.7	26.4	28.1	22.6	24.1	18.0	21.4	20.1	19.1	24.1	22.1	22.0
SINGLETRAN	68.6	74.7	70.8	65.4	75.8	61.3	69.8	74.3	69.0	73.2	51.1	68.5
Δ_{BASELINE}	11.5	17.4	15.6	8.0	20.4	8.7	14.9	13.1	13.8	20.2	6.9	13.7
HOLDOUT	70.4	77.8	75.1	68.9	75.5	63.3	70.7	76.4	70.8	73.7	51.3	70.4
Δ_{BASELINE}	13.3	20.5	19.8	11.5	20.1	10.7	15.8	15.2	15.6	20.7	7.1	15.5
ALLLANGS	72.0	82.5	79.6	75.9	77.6	68.2	73.0	80.3	73.1	75.1	57.8	74.1
Δ_{BASELINE}	14.9	25.2	24.4	18.5	22.2	15.6	18.1	19.1	17.9	22.1	13.7	19.2
Structural Probe Results: Distance Spearman Correlation (DSpr.)												
LINEAR	.573	.570	.533	.567	.589	.489	.564	.598	.578	.543	.493	.554
MBERTRAND	.657	.658	.672	.659	.693	.611	.621	.710	.656	.608	.590	.649
IN-LANG	.822	.845	.846	.817	.859	.813	.812	.864	.807	.798	.777	.824
Δ_{BASELINE}	.165	.187	.174	.158	.166	.202	.191	.154	.151	.190	.187	.175
SINGLETRAN	.774	.801	.807	.773	.838	.732	.787	.836	.772	.771	.655	.777
Δ_{BASELINE}	.117	.143	.135	.114	.145	.121	.166	.126	.117	.163	.064	.128
HOLDOUT	.779	.821	.824	.788	.838	.744	.792	.840	.776	.775	.664	.786
Δ_{BASELINE}	.122	.163	.152	.129	.146	.133	.171	.130	.121	.166	.074	.137
ALLLANGS	.795	.839	.836	.806	.848	.777	.802	.853	.789	.783	.717	.804
Δ_{BASELINE}	.138	.181	.165	.147	.155	.165	.181	.143	.134	.174	.127	.156

Table 1: Performance (in UUAS and DSpr.) of the structural probe trained on the following cross-lingual sources of data: the evaluation language (IN-LANG); the single other best language (SINGLETRAN); all other languages (HOLDOUT); and all languages, including the evaluation language (ALLLANGS). Note that all improvements over baseline (Δ_{BASELINE}) are reported against the stronger of our two baselines per-language.

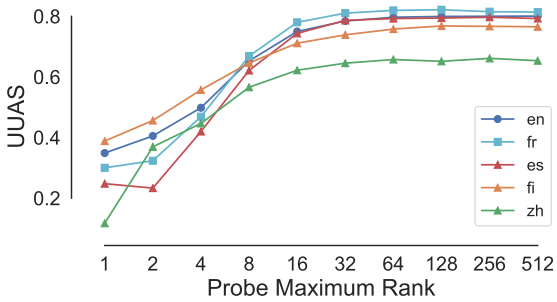


Figure 3: Parse distance tree reconstruction accuracy (UUAS) for selected languages at layer 7 when the linear transformation is constrained to varying maximum dimensionality.

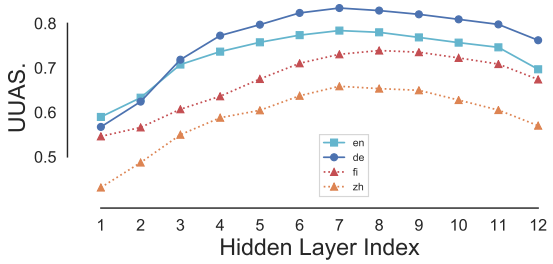


Figure 4: Parse distance tree reconstruction accuracy (UUAS) on layers 1–12 for selected languages, with probe maximum rank 128.

no further gains, implying that the syntactic subspace is a small part of the overall mBERT representation, which has dimension 768 (Figure 3).

These results closely correspond to the results found by Hewitt and Manning (2019) for an equivalently sized monolingual English model trained and evaluated on the Penn Treebank (Marcus et al., 1993), suggesting that mBERT behaves similarly to monolingual BERT in representing syntax.

4 Cross-Lingual Probing

4.1 Transfer Experiments

We now evaluate the extent to which Multilingual BERT's syntactic subspaces are similar across languages. To do this, we evaluate the performance of a structural probe when evaluated on a language unseen at training time. If a probe trained to predict syntax from representations in language i also predicts syntax in language j, this is evidence that mBERT's syntactic subspace for language i also encodes syntax in language j, and thus that syntax

is encoded similarly between the two languages.

Specifically, we evaluate the performance of the structural probe in the following contexts:

- **Direct transfer**, where we train on language i and evaluate on language j.
- **Hold-one-out transfer**, where we train on all languages other than j and evaluate on language j.

4.2 Joint Syntactic Subspace

Building off these cross-lingual transfer experiments, we investigate whether there exists a single joint syntactic subspace that encodes syntax in all languages, and if so, the degree to which it does so. To do so, we train a probe on the concatenation of data from all languages, evaluating it on the concatenation of validation data from all languages.

4.3 Results

We find that mBERT's syntactic subspaces are transferable across all of the languages we examine. Specifically, transfer from the best source language (chosen *post hoc* per-language) achieves on average an improvement of 14 points UUAS and 0.128 DSpr. over the best baseline (Table 1, section SINGLETRAN).[10] Additionally, our results demonstrate the existence of a cross-lingual syntactic subspace; on average, a holdout subspace trained on all languages but the evaluation language achieves an improvement of 16 points UUAS and 0.137 DSpr. over baseline, while a joint ALLLANGS subspace trained on a concatenation of data from all source languages achieves an improvement of 19 points UUAS and 0.156 DSpr. (Table 1, section HOLDOUT, ALLLANGS).

Furthermore, for most languages, syntactic information embedded in the *post hoc* best cross-lingual subspace accounts for 62.3% of the total possible improvement in UUAS (73.1% DSpr.) in recovering syntactic trees over the baseline (as represented by in-language supervision). Holdout transfer represents on average 70.5% of improvement in UUAS (79% DSpr.) over the best baseline, while evaluating on a joint syntactic subspace accounts for 88% of improvement in UUAS (89% DSpr.). These results demonstrate the degree to which the cross-lingual syntactic space represents syntax cross-lingually.

[10] For full results, consult Appendix Table 1.

4.4 Subspace Similarity

Our experiments attempt to evaluate syntactic overlap through zero-shot evaluation of structural probes. In an effort to measure more directly the degree to which the syntactic subspaces of mBERT overlap, we calculate the average principal angle[11] between the subspaces parametrized by each language we evaluate, to test the hypothesis that syntactic subspaces which are closer in angle have closer syntactic properties (Table 4).

We evaluate this hypothesis by asking whether closer subspaces (as measured by lower average principal angle) correlate with better cross-lingual transfer performance. For each language i, we first compute an ordering of all other languages j by increasing probing transfer performance trained on j and evaluated on i. We then compute the Spearman correlation between this ordering and the ordering given by decreasing subspace angle. Averaged across all languages, the Spearman correlation is 0.78 with UUAS, and 0.82 with DSpr., showing that transfer probe performance is substantially correlated with subspace similarity.

4.5 Extrapolation Testing

To get a finer-grained understanding of how syntax is shared cross-lingually, we aim to understand whether less common syntactic features are embedded in the same cross-lingual space as syntactic features common to all languages. To this end, we examine two syntactic relations—prenominal and postnominal adjectives—which appear in some of our languages but not others. We train syntactic probes to learn a subspace on languages that primarily only use one ordering (i.e. majority class is greater than 95% of all adjectives), then evaluate their UUAS score solely on adjectives of the other ordering. Specifically, we evaluate on French, which has a mix (69.8% prenominal) of both orderings, in the hope that evaluating both orderings in the same language may help correct for biases in pairwise language similarity. Since the evaluation ordering is out-of-domain for the probe, predicting evaluation-order dependencies successfully suggests that the learned subspace is capable of generalizing between both kinds of adjectives.

We find that for both categories of languages, accuracy does not differ significantly on either prenominal or postnominal adjectives. Specifi-

[11] https://docs.scipy.org/doc/scipy/reference/generated/scipy.linalg.subspace_angles.html

Language	Prenom.	Postnom.	% data prenom.
de	0.932	0.900	100.0%
zh	0.801	0.826	100.0%
lv	0.752	0.811	99.7%
en	0.906	0.898	99.1%
fi	0.834	0.840	98.5%
cz	0.830	0.894	95.4%
fa	0.873	0.882	9.6%
id	0.891	0.893	4.9%
ar	0.834	0.870	0.1%
Average pre:	0.843	0.862	
Average post:	0.866	0.881	

Table 2: Performance of syntactic spaces trained on various languages on recovering prenominal and postnominal French noun–adjective edges.

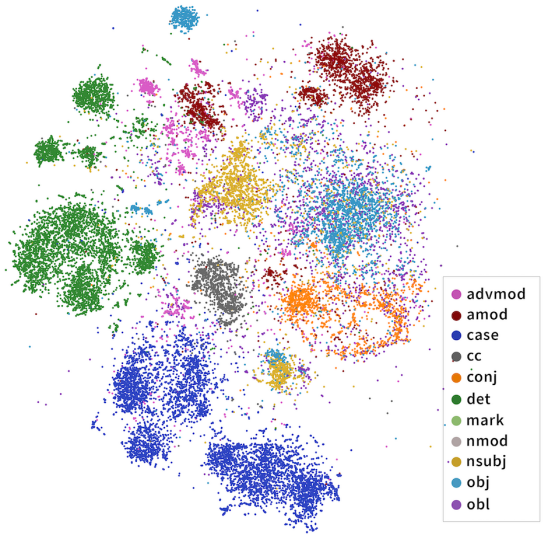


Figure 5: t-SNE visualization of syntactic differences in Spanish projected into a holdout subspace (learned by a probe trained to recover syntax trees in languages other than Spanish). Despite never seeing a Spanish sentence during probe training, the subspace captures a surprisingly fine-grained view of Spanish dependencies.

cally, for both primarily-prenominal and primarily-postnominal training languages, postnominal adjectives score on average approximately 2 points better than prenominal adjectives (Table 2).

5 mBERT Dependency Clusters Capture Universal Grammatical Relations

5.1 Methodology

Given the previous evidence that mBERT shares syntactic representations cross-lingually, we aim to more qualitatively examine the nature of syntactic dependencies in syntactic subspaces. Let $\mathcal{D}$ be a dataset of parsed sentences, and the linear transformation $B \in \mathbb{R}^{k \times m}$ define a k-dimensional syntactic subspace. For every non-root word and hence syntactic dependency in $\mathcal{D}$ (since every word is a dependent of some other word or an added ROOT symbol), we calculate the k-dimensional *head-dependent vector* between the head and the dependent after projection by B. Specifically, for all head-dependent pairs $(w_{\text{head}}, w_{\text{dep}})$, we compute $v_{\text{diff}} = B(\mathbf{h}_{\text{head}} - \mathbf{h}_{\text{dep}})$. We then visualize all differences over all sentences in two dimensions using t-SNE (van der Maaten and Hinton, 2008).

5.2 Experiments

As with multilingual probing, one can visualize head-dependent vectors in several ways; we present the following experiments:

- dependencies from one language, projected into a different language's space (Figure 1)
- dependencies from one language, projected into a holdout syntactic space trained on all other languages (Figure 5)
- dependencies from all languages, projected into a joint syntactic space trained on all languages (Figure 6)

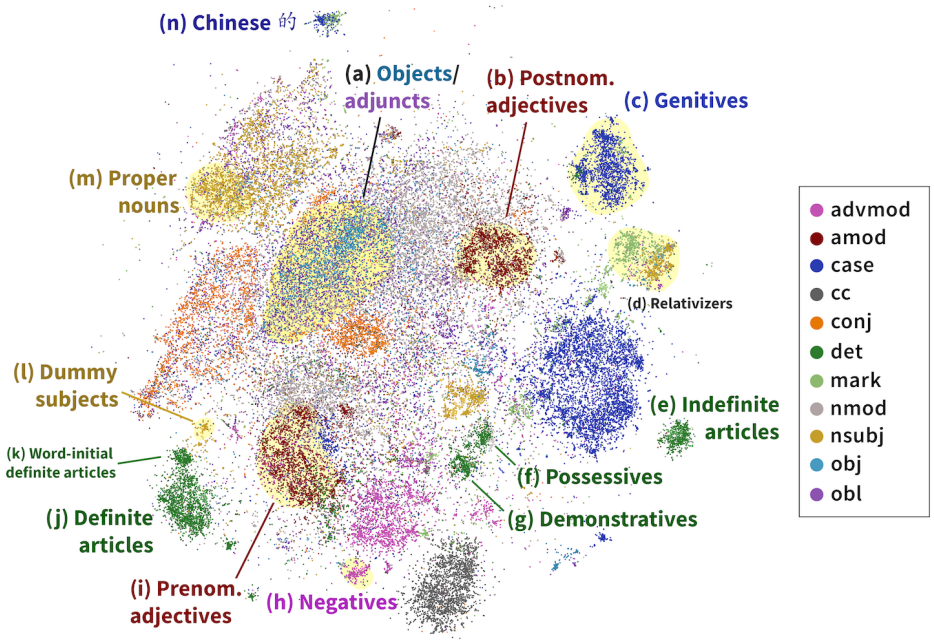
For all these experiments, we project into 32-dimensional syntactic spaces.[12] Additionally, we expose a web interface for visualization in our GitHub repository.[13]

5.3 Results

When projected into a syntactic subspace determined by a structural probe, we find that difference vectors separate into clusters reflecting linguistic characteristics of the dependencies. The cluster identities largely overlap with (but do not exactly agree with) dependency labels as defined by Universal Dependencies (Figure 6). Additionally, the clusters found by mBERT are highly multilingual. When dependencies from several languages are projected into the same syntactic subspace, whether trained monolingually or cross-lingually, we find that dependencies of the same label share the same cluster (e.g. Figure 1, which presents both English

[12] We reduce the dimensionality of the subspaces here as compared to our previous experiments to match t-SNE suggestions and more aggressively filter non-syntactic information.

[13] https://github.com/ethanachi/multilingual-probing-visualization/blob/master/visualization.md



Example sentences (trimmed for clarity). Heads in **bold**; dependents in ***bold italic***.

(b) Postnominal adjectives	fr	Le gaz développe ses **applications** ***domestiques***.
	id	**Film** ***lain*** yang menerima penghargaan istimewa.
	fa	صمماتعین ع اوپک در تنظم **قمت خام**
(c) Genitives	en	The assortment ***of*** **customers** adds entertainment.
	es	Con la recuperación ***de*** la **democracia** y las libertades
	lv	Svešiniece piecēlās, atvadījās ***no*** vecā **vīra**
(j) Definite articles	en	The value of ***the*** highest **bid**
	fr	Merak est une ville d'Indonésie sur ***la*** côte **occidentale**.
	de	Selbst mitten in ***der*** **Woche** war das Lokal gut besucht.

Figure 6: t-SNE visualization of 100,000 syntactic difference vectors projected into the cross-lingual syntactic subspace of Multilingual BERT. We exclude **punct** and visualize the top 11 dependencies remaining, which are collectively responsible for 79.36% of the dependencies in our dataset. Clusters of interest highlighted in yellow; linguistically interesting clusters labeled.

and French syntactic difference vectors projected into an English subspace).

5.4 Finer-Grained Analysis

Visualizing syntactic differences in the syntactic space provides a surprisingly nuanced view of the native distinctions made by mBERT. In Figure 6, these differences are colored by gold UD dependency labels. A brief summary is as follows:

Adjectives Universal Dependencies categorizes all adjectival noun modifiers under the **amod** relation. However, we find that mBERT splits adjectives into two groups: prenominal adjectives in cluster (b) (e.g., Chinese **獨特的**地理) and postnominal adjectives in cluster (u) (e.g., French *applications* ***domestiques***).

Nominal arguments mBERT maintains the UD distinction between subject (**nsubj**) and object (**obj**). Indirect objects (**iobj**) cluster with direct objects. Interestingly, mBERT generally groups adjunct arguments (**obl**) with **nsubj** if near the beginning of a sentence and **obj** otherwise.

Relative clauses In the languages in our dataset, there are two major ways of forming relative clauses. Relative pronouns (e.g., English *the man* ***who*** *is* hungry are classed by Universal Dependencies as being an **nsubj** dependent, while subordinating markers (e.g., English *I know* ***that*** *she saw me*) are classed as the dependent of a **mark** relation. However, mBERT groups both of these relations together, clustering them distinctly from most **nsubj** and **mark** relations.

Negatives Negative adverbial modifiers (English *not*, Farsi غیر, Chinese 不) are not clustered with other adverbial syntactic relations (**advmod**), but form their own group (h).[14]

Determiners The linguistic category of determiners (**det**) is split into definite articles (i), indefinite articles (e), possessives (f), and demonstratives (g). Sentence-initial definite articles (k) cluster separately from other definite articles (j).

Expletive subjects Just as in UD, with the separate relation **expl**, expletive subjects, or third-person pronouns with no syntactic meaning (e.g. English ***It*** *is cold*, French ***Il*** *faudrait*, Indonesian ***Yang*** *menjadi masalah kemudian*), cluster separately (k) from other **nsubj** relations (small cluster in the bottom left).

Overall, mBERT draws slightly different distinctions from Universal Dependencies. Although some are more fine-grained than UD, others appear to be more influenced by word order, separating relations that most linguists would group together. Still others are valid linguistic distinctions not distinguished by the UD standard.

5.5 Discussion

Previous work has found that it is possible to recover dependency labels from mBERT embeddings, in the form of very high accuracy on dependency label probes (Liu et al., 2019; Tenney et al., 2019b). However, although we know that dependency label probes are able to use supervision to map from mBERT's representations to UD dependency labels, this does not provide full insight into the nature of (or existence of) latent dependency label structure in mBERT. By contrast, in the structural probe, B is optimized such that $\|v_{\text{diff}}\|_2 \approx 1$, but no supervision as to dependency label is given. The contribution of our method is thus to provide a view into mBERT's "own" dependency label representation. In Appendix A, Figure 8, we provide a similar visualization as applied to MBERTRAND, finding much less cluster coherence.

5.6 Probing as a window into representations

Our head-dependent vector visualization uses a supervised probe, but its objects of study are properties of the representation *other* than those relating to the probe supervision signal. Because the probe never sees supervision on the task we visualize for, the visualized behavior cannot be the result of the probe memorizing the task, a problem in probing methodology (Hewitt and Liang, 2019). Instead, it is an example of using probe supervision to focus in on aspects that may be drowned out in the original representation. However, the probe's linear transformation may not pick up on aspects that are of causal influence to the model.

[14] Stanford Dependencies and Universal Dependencies v1 had a separate **neg** dependency, but it was eliminated in UDv2.

6 Related Work

Cross-lingual embedding alignment Lample et al. (2018) find that independently trained monolingual word embedding spaces in ELMo are isometric under rotation. Similarly, Schuster et al. (2019) and Wang et al. (2019) geometrically align contextualized word embeddings trained independently. Wu et al. (2019) find that cross-lingual transfer in mBERT is possible even without shared vocabulary tokens, which they attribute to this isometricity. In concurrent work, Cao et al. (2020) demonstrate that mBERT embeddings of similar words in similar sentences across languages are approximately aligned already, suggesting that mBERT also aligns semantics across languages. K et al. (2020) demonstrate that strong cross-lingual transfer is possible without any word piece overlap at all.

Analysis with the structural probe In a monolingual study, Reif et al. (2019) also use the structural probe of Hewitt and Manning (2019) as a tool for understanding the syntax of BERT. They plot the words of individual sentences in a 2-dimensional PCA projection of the structural probe distances, for a geometric visualization of individual syntax trees. Further, they find that distances in the mBERT space separate clusters of word senses for the same word type.

Understanding representations Pires et al. (2019) find that cross-lingual BERT representations share a common subspace representing useful linguistic information. Libovický et al. (2019) find that mBERT representations are composed of a language-specific component and a language-neutral component. Both Libovický et al. (2019) and Kudugunta et al. (2019) perform SVCCA on LM representations extracted from mBERT and a massively multilingual transformer-based NMT model, finding language family-like clusters.

Li and Eisner (2019) present a study in syntactically motivated dimensionality reduction; they find that after being passed through an information bottleneck and dimensionality reduction via t-SNE, ELMo representations cluster naturally by UD part of speech tags. Unlike our syntactic dimensionality reduction process, the information bottleneck is directly supervised on POS tags, whereas our process receives no linguistic supervision other than unlabeled tree structure. In addition, the reduction process, a feed-forward neural network, is more complex than our linear transformation.

Singh et al. (2019) evaluate the similarity of mBERT representations using Canonical Correlation Analysis (CCA), finding that overlap among subword tokens accounts for much of the representational similarity of mBERT. However, they analyze cross-lingual overlap across all components of the mBERT representation, whereas we evaluate solely the overlap of syntactic subspaces. Since syntactic subspaces are at most a small part of the total BERT space, these are not necessarily mutually contradictory with our results. In concurrent work, Michael et al. (2020) also extend probing methodology, extracting latent ontologies from contextual representations without direct supervision.

7 Discussion

Language models trained on large amounts of text have been shown to develop surprising emergent properties; of particular interest is the emergence of non-trivial, easily accessible linguistic properties seemingly far removed from the training objective. For example, it would be a reasonable strategy for mBERT to share little representation space between languages, effectively learning a private model for each language and avoiding destructive interference. Instead, our transfer experiments provide evidence that at a syntactic level, mBERT shares portions of its representation space between languages. Perhaps more surprisingly, we find evidence for fine-grained, cross-lingual syntactic distinctions in these representations. Even though our method for identifying these distinctions lacks dependency label supervision, we still identify that mBERT has a cross-linguistic clustering of grammatical relations that qualitatively overlaps considerably with the Universal Dependencies formalism.

The UUAS metric We note that the UUAS metric alone is insufficient for evaluating the accuracy of the structural probe. While the probe is optimized to directly recreate parse distances, (that is, $d_B(\mathbf{h}_i^\ell, \mathbf{h}_j^\ell) \approx d_T^\ell(w_i^\ell, w_j^\ell)$) a perfect UUAS score under the minimum spanning tree construction can be achieved by ensuring that $d_B(\mathbf{h}_i^\ell, \mathbf{h}_j^\ell)$ is small if there is an edge between w_i^ℓ and w_j^ℓ, and large otherwise, instead of accurately recreating distances between words connected by longer paths. By evaluating Spearman correlation between all pairs of words, one directly evaluates the extent to which the ordering of words j by distance to each word i is correctly predicted, a key notion of the geometric interpretation of the structural probe. See Maudslay et al. (2020) for further discussion.

Limitations Our methods are unable to tease apart, for all pairs of languages, whether transfer performance is caused by subword overlap (Singh et al., 2019) or by a more fundamental sharing of parameters, though we do note that language pairs with minimal subword overlap do exhibit non-zero transfer, both in our experiments and in others (K et al., 2020). Moreover, while we quantitatively evaluate cross-lingual transfer in recovering dependency distances, we only conduct a qualitative study in the unsupervised emergence of dependency labels via t-SNE. Future work could extend this analysis to include quantitative results on the extent of agreement with UD. We acknowledge as well issues in interpreting t-SNE plots (Wattenberg et al., 2016), and include multiple plots with various hyperparameter settings to hedge against this confounder in Figure 11.

Future work should explore other multilingual models like XLM and XLM-RoBERTa (Lample and Conneau, 2019) and attempt to come to an understanding of the extent to which the properties we've discovered have causal implications for the decisions made by the model, a claim our methods cannot support.

8 Acknowledgements

We would like to thank Erik Jones, Sebastian Schuster, and Chris Donahue for helpful feedback and suggestions. We would also like to thank the anonymous reviewers and area chair Adina Williams for their helpful comments on our draft.

References

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

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

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*, pages 2126–2136.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 7059–7069. Curran Associates, Inc.

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*, pages 4171–4186.

John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. In *North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*.

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*.

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)*.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*.

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

Xiang Lisa Li and Jason Eisner. 2019. Specializing word embeddings (for parsing) by information bottleneck. 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)*.

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

Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. 2019. Linguistic knowledge and transferability of contextual representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Rowan Hall Maudslay, Josef Valvoda, Tiago Pimentel, Adina Williams, and Ryan Cotterell. 2020. A tale of a probe and a parser. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Julian Michael, Jan A. Botha, and Ian Tenney. 2020. Asking without telling: Exploring latent ontologies in contextual representations. *arXiv preprint arXiv:2004.14513*.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the Twelfth International Conference on Language Resources and Evaluation*.

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*.

Emily Reif, Ann Yuan, Martin Wattenberg, Fernanda B Viegas, Andy Coenen, Adam Pearce, and Been Kim. 2019. Visualizing and measuring the geometry of bert. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8592–8600.

Tal Schuster, Ori Ram, Regina Barzilay, and Amir Globerson. 2019. Cross-lingual alignment of contextual word embeddings, with applications to zero-shot dependency parsing. *arXiv preprint arXiv:1902.09492*.

Jasdeep Singh, Bryan McCann, Richard Socher, and Caiming Xiong. 2019. BERT is not an interlingua and the bias of tokenization. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019a. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

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

L.J.P. van der Maaten and G.E. Hinton. 2008. Visualizing high-dimensional data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605.

Yuxuan Wang, Wanxiang Che, Jiang Guo, Yijia Liu, and Ting Liu. 2019. Cross-lingual BERT transformation for zero-shot dependency parsing. 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)*.

Martin Wattenberg, Fernanda Viégas, and Ian Johnson. 2016. How to use t-sne effectively. *Distill*, 1(10):e2.

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

Shijie Wu, Alexis Conneau, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Emerging cross-lingual structure in pretrained language models. *arXiv preprint arXiv:1911.01464*.

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.

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

A Additional Syntactic Difference Visualizations

A.1 Visualization of All Relations

In our t-SNE visualization of syntactic difference vectors projected into the cross-lingual syntactic subspace of Multilingual BERT (Figure 6), we only visualize the top 11 relations, excluding **punct**. This represents 79.36% of the dependencies in our dataset. In Figure 7, we visualize all 36 relations in the dataset.

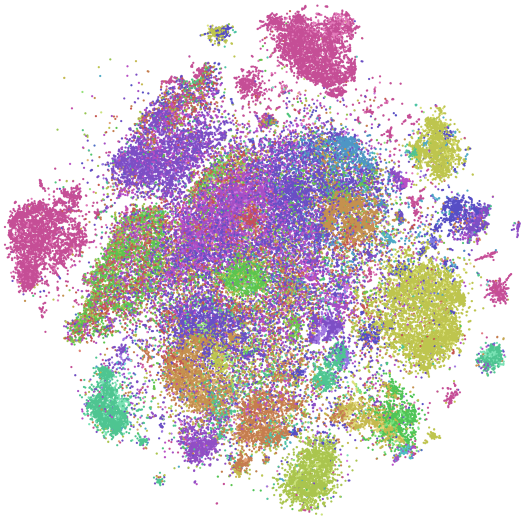


Figure 7: t-SNE visualization of dependency head-dependent pairs projected into the cross-lingual syntactic subspace of Multilingual BERT. Colors correspond to gold UD dependency type labels, which are unlabeled given that there are 43 in this visualization.

A.2 Visualization with Randomly-Initialized Baseline

In Figure 8, we present a visualization akin to Figure 1; however, both the head-dependency representations, as well as the syntactic subspace, are derived from MBERTRAND. Clusters around the edges of the figure are primarily type-based (e.g. one cluster for the word *for* and another for *pour*), and there is insignificant overlap between clusters with parallel syntactic functions from different languages.

B Alternative Dimensionality Reduction Strategies

In an effort to confirm the level of clarity of the clusters of dependency types which emerge from

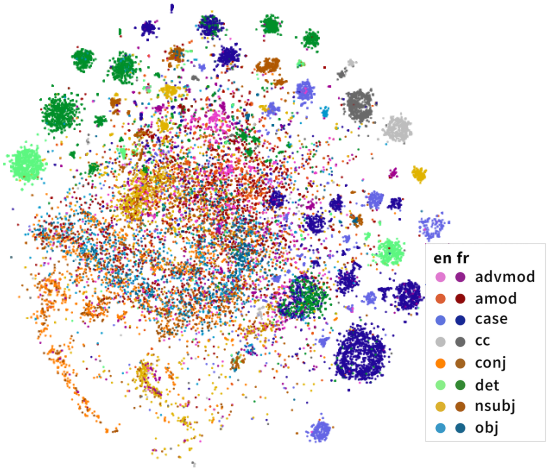


Figure 8: t-SNE visualization of head-dependent dependency pairs belonging to selected dependencies in English and French, projected into a syntactic subspace of MBERTRAND, as learned on English syntax trees. Colors correspond to gold UD dependency type labels.

syntactic difference vectors, we examine simpler strategies for dimensionality reduction.

B.1 PCA for Visualization Reduction

We project difference vectors as previously into a 32-dimensional syntactic subspace. However, we visualize in 2 dimensions using PCA instead of t-SNE. There are no significant trends evident.

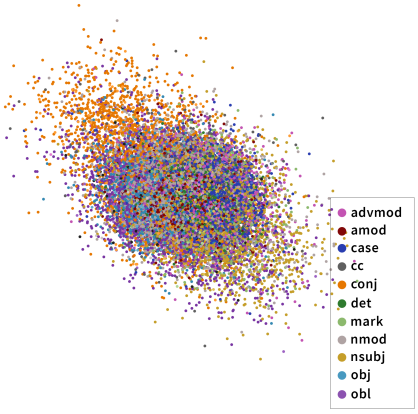


Figure 9: Syntactic difference vectors visualized after dimensionality reduction with PCA, instead of t-SNE, colored by UD dependency types. There are no significant trends evident.

B.2 PCA for Dimensionality Reduction

Instead of projecting difference vectors into our syntactic subspace, we first reduce them to a 32-dimensional representation using PCA,[15] then reduce to 2 dimensions using t-SNE as previously.

We find that projected under PCA, syntactic difference vectors still cluster into major groups, and major trends are still evident (Figure 10). In addition, many finer-grained distinctions are still apparent (e.g. the division between common nouns and pronouns). However, in some cases, the clusters are motivated less by syntax and more by semantics or language identities. For example:

- The **nsubj** and **obj** clusters overlap, unlike our syntactically-projected visualization, where there is clearer separation.
- Postnominal adjectives, which form a single coherent cluster under our original visualization scheme, are split into several different clusters, each primarily composed of words from one specific language.
- There are several small monolingual clusters without any common syntactic meaning, mainly composed of languages parsed more poorly by BERT (i.e. Chinese, Arabic, Farsi, Indonesian).

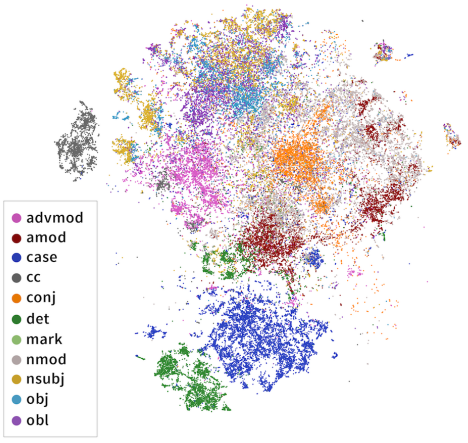


Figure 10: t-SNE visualization of syntactic differences in all languages we study, projected to 32 dimensions using PCA.

C Additional Experiment Settings

C.1 Pairwise Transfer

We present full pairwise transfer results in Table 3. Each experiment was run 3 times with different random seeds; experiment settings with range in

[15] This is of equal dimensionality to our syntactic subspace.

Structural Probe Results: Undirected Unlabeled Attachment Score (UUAS)

Tgt \Src	ar	cz	de	en	es	fa	fi	fr	id	lv	zh	linear	rand	holdout	all
ar	72.7	68.6	66.6	65.3	67.5	64.0	60.8	68.1	65.3	60.1	53.4	57.1	49.8	70.4	72.0
cz	57.5*	83.6	74.7	72.6	71.1	63.5	68.9	71.5	62.4	71.0	58.0	45.4	57.3	77.8	82.5
de	49.3	70.2	83.5	70.8	68.2	58.7	61.1	70.6	56.9*	62.0	52.0*	42.8	55.2	75.1	79.6
en	47.2	61.2	65.0	79.8	63.9	50.8	55.3	65.4	54.5	54.0	50.5	41.5	57.4	68.9	75.9
es	52.0	67.2	69.8	69.4	79.7	56.9	56.8	75.8	61.0	55.6	49.2	44.6	55.3	75.5	77.6
fa	51.7	61.3	60.3	57.0	57.8	70.8	53.7	59.7	56.5	53.1	49.7	52.6	43.2	63.3	68.2
fi	55.5	69.8	68.4	66.6	66.0	60.2	76.5	66.0	61.2	68.2	59.2	50.1	54.9	70.7	73.0
fr	50.8*	67.8	73.0	70.0	74.3	56.9	55.9	84.0	60.9	55.1	49.6	46.4	61.2	76.4	80.3
id	57.1	66.3	67.4	63.6	67.0	61.0	59.2	69.0	74.8	57.5	54.6	55.2	53.2	70.8	73.1
lv	56.9*	73.2	69.2	69.1	67.0	61.5	70.8	66.7	61.1	77.0	60.7	47.0	53.0	73.7	75.1
zh	41.2*	49.7	49.6	51.1	47.3	42.7*	48.1	47.9	44.5*	47.2	65.7	44.2	41.1	51.3	57.8

Structural Probe Results: Distance Spearman Correlation (DSpr.)

Tgt \Src	ar	cz	de	en	es	fa	fi	fr	id	lv	zh	linear	rand	holdout	all
ar	.822	.772	.746	.744	.774	.730	.723	.770	.750	.722	.640	.573	.657	.779	.795
cz	.730	.845	.799	.781	.801	.741	.782	.796	.745	.791	.656	.570	.658	.821	.839
de	.690	.807	.846	.792	.792	.736	.767	.796	.723	.765	.652*	.533	.672	.824	.836
en	.687	.765	.764	.817	.770	.696	.732	.773	.720	.725	.655	.567	.659	.788	.806
es	.745	.821	.812	.806	.859	.741	.775	.838	.777	.774	.669	.589	.693	.838	.848
fa	.661	.732	.724	.706	.705	.813	.683	.714	.686	.684	.629	.489	.611	.744	.777
fi	.682*	.787	.771	.756	.764	.712	.812	.762	.715	.781	.658	.564	.621	.792	.802
fr	.731*	.810	.816	.806	.836	.738	.767	.864	.776	.760	.674	.598	.710	.840	.853
id	.715	.757	.752	.739	.765	.718	.714	.772	.807	.704	.657	.578	.656	.776	.789
lv	.681	.771	.746	.737	.745	.699	.763	.740	.698	.798	.644	.543	.608	.775	.783
zh	.538*	.655	.644	.644	.633	.593*	.652	.638	.584*	.639	.777	.493	.590	.664	.717

Table 3: Performance (in UUAS and DSpr.) on transfer between all language pairs in our dataset. All runs were repeated 3 times; runs for which the range in performance exceeded 2 points (for UUAS) or 0.02 (for DSpr.) are marked with an asterisk (*).

	ar	cz	de	en	es	fa	fi	fr	id	lv	zh
ar	0.000	1.044	1.048	1.049	1.015	1.046	1.058	1.022	1.031	1.059	1.076
cz	1.044	0.000	0.982	1.017	0.970	1.064	1.021	1.007	1.053	1.011	1.083
de	1.048	0.982	0.000	1.005	0.973	1.044	1.017	0.971	1.022	1.029	1.065
en	1.049	1.017	1.005	0.000	0.983	1.051	1.033	0.994	1.035	1.040	1.060
es	1.015	0.970	0.973	0.983	0.000	1.038	1.023	0.936	1.010	1.024	1.065
fa	1.046	1.064	1.044	1.051	1.038	0.000	1.060	1.028	1.040	1.063	1.069
fi	1.058	1.021	1.017	1.033	1.023	1.060	0.000	1.020	1.042	1.011	1.058
fr	1.022	1.007	0.971	0.994	0.936	1.028	1.020	0.000	0.993	1.028	1.041
id	1.031	1.053	1.022	1.035	1.010	1.040	1.042	0.993	0.000	1.051	1.052
lv	1.059	1.011	1.029	1.040	1.024	1.063	1.011	1.028	1.051	0.000	1.068
zh	1.076	1.083	1.065	1.060	1.065	1.069	1.058	1.041	1.052	1.068	0.000

Table 4: Subspace angle overlap as evaluated by the pairwise mean principal angle between subspaces

UUAS greater than 2 points are labeled with an asterisk (*).

C.2 Subspace Overlap

Table 4 presents the average principal angle between the subspaces parametrized by each language we evaluate. Table 5 contains the per-language Spearman correlation between the ordering given by (negative) subspace angle and structural probe transfer accuracy, reported both on UUAS and DSpr.

D Data Sources

We use the following UD corpora in our experiments: Arabic-PADT, Chinese-GSD, Czech-PDT, English-EWT, Finnish-TDT, French-GSD, German-GSD, Indonesian-GSD, Latvian-LVTB, Persian-Seraji, and Spanish-Ancora.

E t-SNE reproducibility

Previous work (Wattenberg et al., 2016) has investigated issues in the interpretability of tSNE plots. Given the qualitative nature of our experiments, to avoid this confounder, we include multiple plots with various settings of the perplexity hyperparameter in Figure 11.

Language	ar	cz	de	en	es	fa	fi	fr	id	lv	zh
Spearman Correl. (UUAS)	0.88	0.85	0.87	0.91	0.91	0.48	0.85	0.89	0.71	0.90	0.41
Spearman Correl. (DSpr.)	0.95	0.96	0.95	0.96	0.97	0.50	0.90	0.93	0.72	0.94	0.23

Table 5: The Spearman correlation between two orderings of all languages for each language i. The first ordering of languages is given by (negative) subspace angle between the B matrix of language i and that of all languages. The second ordering is given by the structural probe transfer accuracy from all languages (including i) to i. This is repeated for each of the two structural probe evaluation metrics.

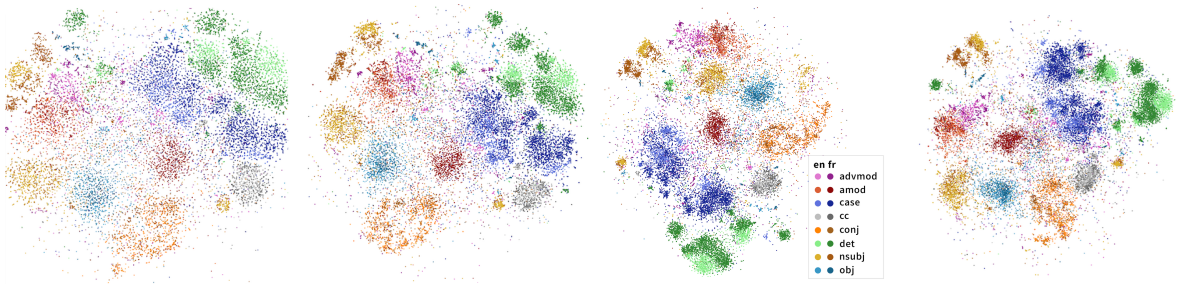


Figure 11: t-SNE visualization of head-dependent dependency pairs belonging to selected dependencies in English and French, projected into a syntactic subspace of Multilingual BERT, as learned on English syntax trees. Colors correspond to gold UD dependency type labels, as in Figure 1, varying the perplexity (PPL) t-SNE hyperparmeter. From left to right, figures correspond to PPL $5, 10, 30, 50$, spanning the range of PPL suggested by van der Maaten and Hinton (2008). Cross-lingual dependency label clusters are exhibited across all four figures.