Do Neural Language Models Show Preferences for Syntactic Formalisms?

Artur Kulmizev
Uppsala University
artur.kulmizev@lingfil.uu.se

Vinit Ravishankar
University of Oslo
vinitr@ifi.uio.no

Mostafa Abdou
University of Copenhagen
abdou@di.ku.dk

Joakim Nivre
Uppsala University
joakim.nivre@lingfil.uu.se

Abstract

Recent work on the interpretability of deep neural language models has concluded that many properties of natural language syntax are encoded in their representational spaces. However, such studies often suffer from limited scope by focusing on a single language and a single linguistic formalism. In this study, we aim to investigate the extent to which the semblance of syntactic structure captured by language models adheres to a surface-syntactic or deep syntactic style of analysis, and whether the patterns are consistent across different languages. We apply a probe for extracting directed dependency trees to BERT and ELMo models trained on 13 different languages, probing for two different syntactic annotation styles: Universal Dependencies (UD), prioritizing deep syntactic relations, and Surface-Syntactic Universal Dependencies (SUD), focusing on surface structure. We find that both models exhibit a preference for UD over SUD — with interesting variations across languages and layers — and that the strength of this preference is correlated with differences in tree shape.

1 Introduction

Recent work on interpretability in NLP has led to the consensus that deep neural language models trained on large, unannotated datasets manage to encode various aspects of syntax as a byproduct of the training objective. Probing approaches applied to models like ELMo (Peters et al., 2018a) and BERT (Devlin et al., 2019) have demonstrated that one can decode various linguistic properties such as part-of-speech categories, dependency relations, and named-entity types directly from the internal hidden states of a pretrained model (Tenney et al., 2019b,b; Peters et al., 2018b). Another line of work has tried to tie cognitive measurements or theories of human linguistic processing to the machinations of language models, often establishing strong parallels between the two (Prasad et al., 2019; Abnar et al., 2019; Gauthier and Levy, 2019).

As is the case for NLP in general, English has served as the de facto testing ground for much of this work, with other languages often appearing as an afterthought. However, despite its ubiquity in the NLP literature, English is generally considered to be atypical across many typological dimensions. Furthermore, the tendency of interpreting NLP models with respect to existing, canonical datasets often comes with the danger of conflating the theory-driven annotation therein with scientific fact. One can observe this to an extent with the Universal Dependencies (UD) project (Nivre et al., 2016), which aims to collect syntactic annotation for a large number of languages. Many interpretability studies have taken UD as a basis for training and evaluating probes, but often fail to mention that UD, like all annotation schemes, is built upon specific theoretical assumptions, which may not be universally accepted.

Our research questions start from these concerns. When probing language models for syntactic dependency structure, is UD — with its emphasis on syntactic relations between content words — really the best fit? Or is the representational structure of such models better explained by a scheme that is more oriented towards surface structure, such as the recently proposed Surface-Syntactic Universal Dependencies (SUD) (Gerdes et al., 2018)? And are these patterns consistent across typologically different languages? To explore these questions, we fit the structural probe of Hewitt and Manning (2019) on pretrained BERT and ELMo representations, supervised by UD/SUD treebanks for 13 languages, and extract directed dependency trees. We then conduct an extensive error analysis of the resulting probed parses, in an attempt to qualify our findings. Our main contributions are the following:

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 4077–4091

1. A simple algorithm for deriving directed trees from the disjoint distance and depth probes introduced by Hewitt and Manning (2019).
2. A multilingual analysis of the probe's performance across 13 different treebanks.
3. An analysis showing that the syntactic information encoded by BERT and ELMo fit UD better than SUD for most languages.

2 Related Work

There has been a considerable amount of recent work attempting to understand what aspects of natural language pre-trained encoders learn. The classic formulation of these probing experiments is in the form of diagnostic classification (Ettinger et al., 2016; Belinkov et al., 2017; Hupkes et al., 2018; Conneau et al., 2018), which attempts to unearth underlying linguistic properties by fitting relatively underparameterised linear models over representations generated by an encoder. These methods have also faced recent critique, for example, concerning the lack of transparency in the classifers' ability to *extract* meaningful information, as opposed to *learning* it. Alternative paradigms for interpretability have therefore been proposed, such as correlation-based methods (Raghu et al., 2017; Saphra and Lopez, 2018; Kornblith et al., 2019; Chrupała and Alishahi, 2019). However, this critique does not invalidate diagnostic classification: indeed, more recent work (Hewitt and Liang, 2019) describes methods to show the empirical validity of certain probes, via control tasks.

Among probing studies specifically pertinent to our paper, Blevins et al. (2018) demonstrate that deep RNNs are capable of encoding syntax given a variety of pre-training tasks, including language modeling. Peters et al. (2018b) demonstrate that, regardless of encoder (recurrent, convolutional, or self-attentive), biLM-based pre-training results in similar high-quality representations that implicitly encode a variety of linguistic phenomena, layer by layer. Similarly, Tenney et al. (2019a) employ the 'edge probing' approach of Tenney et al. (2019b) to demonstrate that BERT implicitly learns the 'classical NLP pipeline', with lower-level linguistic tasks encoded in lower layers and more complex phenomena in higher layers, and dependency syntax in layer 5–6. Finally, Hewitt and Manning (2019) describe a syntactic probe for extracting aspects of dependency syntax from pre-trained representations, which we describe in Section 4.

3 Aspects of Syntax

Syntax studies how natural language encodes meaning using expressive devices such as word order, case marking and agreement. Some approaches emphasize the formal side and primarily try to account for the distribution of linguistic forms. Other frameworks focus on the functional side to capture the interface to semantics. And some theories use multiple representations to account for both perspectives, such as c-structure and f-structure in LFG (Kaplan and Bresnan, 1982; Bresnan, 2000) or surface-syntactic and deep syntactic representations in Meaning-Text Theory (Mel'čuk, 1988).

When asking whether neural language models learn syntax, it is therefore relevant to ask which aspects of syntax we are concerned with. This is especially important if we probe the models by trying to extract syntactic representations, since these representations may be based on different theoretical perspectives. As a first step in this direction, we explore two different dependency-based syntactic representations, for which annotations are available in multiple languages. The first is Universal Dependencies (UD) (Nivre et al., 2016), a framework for cross-linguistically consistent morpho-syntactic annotation, which prioritizes direct grammatical relations between content words. These relations tend to be more parallel across languages that use different surface features to encode the relations. The second is Surface-Syntactic Universal Dependencies (SUD) (Gerdes et al., 2018), a recently proposed alternative to UD, which gives more prominence to function words in order to capture variations in surface structure across languages.

Figure 1 contrasts the two frameworks by showing how they annotate an English sentence. While the two annotations agree on most syntactic relations (in black), including the analysis of core grammatical relations like subject (nsubj[1]) and object (obj), they differ in the analysis of auxiliaries and prepositional phrases. The UD annotation (in blue) treats the main verb *chased* as the root of the clause, while the SUD annotation (in red) assigns this role to the auxiliary *has*. The UD annotation has a direct oblique relation between *chased* and *room*, treating the preposition *from* as a case marker, while the SUD annotation has an oblique relation between *chased* and *from*, analyzing *room* as the object of *from*. The purpose of the UD style of

[1]UD uses the *nsubj* relation, for *nominal* subject, while SUD uses a more general *subj* relation.

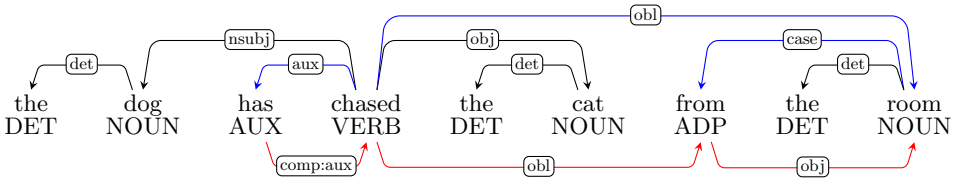

Figure 1: Simplified UD and SUD annotation for an English sentence.

annotation is to increase the probability of the root and oblique relations being parallel in other languages that use morphology (or nothing at all) to encode the information expressed by auxiliaries and adpositions. SUD is instead designed to bring out differences in surface structure in such cases.

The different treatment of function words affects not only adpositions (prepositions and postpositions) and auxiliaries (including copulas), but also subordinating conjunctions and infinitive markers. Because of these systematic differences, dependency trees in UD tend to have longer average dependency length and smaller height[2] than in SUD.

4 Probing Model

To conduct our experiments, we make use of the structural probe proposed by Hewitt and Manning (2019), which is made up of two complementary components — distance and depth. The former is an intuitive proxy for the notion of two words being connected by a dependency: any two words w_i, w_j in a tree T are neighbors if their respective distance in the tree amounts to $d_T(w_i, w_j) = 1$. This metric can theoretically be applied to the vector space of any pretrained neural language model sentence encoding, which ouputs a set of vectors $S = \mathbf{h}_1, ..., \mathbf{h}_n$ for a sentence. In practice, however, the distance between any two vectors $\{\mathbf{h}_i, \mathbf{h}_j\} \in S$ will not be directly comparable to their distance in a corresponding syntactic tree T, because the model does not encode syntax in isolation. To resolve this, Hewitt and Manning (2019) propose to learn a linear transformation matrix B, such that $d_B(\mathbf{h}_i, \mathbf{h}_j)$ extracts the distance between any two words w_i, w_j in a parse tree. For an annotated corpus of L sentences, the distance probe can be learned via gradient descent as follows:

$$\min_B \sum_{l=1}^{L} \frac{1}{|n^l|^2} \sum_{i,j} |d_{T^l}(w_i^l, w_j^l) - d_B(\mathbf{h}_i^l, \mathbf{h}_j^l)^2|$$

where $|n^l|$ is the length of sentence l, normalized by the number $|n^l|^2$ of word pairs, and $d_{T^l}(w_i^l, w_j^l)$ is the distance of words w_i^l and w_j^l in the gold tree.

While the distance probe can predict which words enter into dependencies with one another, it is insufficient for predicting which word is the head. To resolve this, Hewitt and Manning (2019) employ a separate probe for tree depth,[3] where they make a similar assumption as they do for distance: a given (square) vector L2 norm $||\mathbf{h}_i^2||$ is analogous to w_i's depth in a tree T. A linear transformation matrix B can therefore be learned in a similar way:

$$\min_B \sum_{l=1}^{L} \frac{1}{n_l} \sum_{i}^{n} (||w_i^l|| - ||B\mathbf{h}_i^l||^2)$$

where $||w_i^l||$ is the depth of a w_i^l in the gold tree.

To be able to score probed trees (against UD and SUD gold trees) using the standard metric of unlabeled attachment score (UAS), we need to derive a rooted directed dependency tree from the information provided by the distance and depth probes. Algorithm 1 outlines a simple method to retrieve a well-formed tree with the help of the Chu-Liu-Edmonds maximum spanning tree algorithm (Chu and Liu, 1965; McDonald et al., 2005). Essentially, in a sentence $S = w_1 \dots w_n$, for every pair of nodes (w_i, w_j) with an estimated distance of d between them, if w_i has smaller depth than w_j, we set the weight of the arc (w_i, w_j) to $-d$; otherwise, we set the weight to $-\infty$. This is effectively a mapping from distances to scores, with larger distances resulting in lower arc scores from the parent to the child, and infinitely low scores from the child to the parent. We also add a pseudo-root w_0 (essential for decoding), which has a single arc pointing to the shallowest node (weighted 0). We use the AllenNLP (Gardner et al., 2018) implementation of the Chu-Liu/Edmonds' algorithm.

5 Experimental Design

In order to evaluate the extent to which a given model's representational space fits either annota-

[2]The *height* of a tree is the length of the longest path from the root to a leaf (sometimes referred to as *depth*).

[3]The *depth* of a node is the length of the path from the root.

Language	Code	Treebank	# Sents	%ADP	%AUX	%ContRel		Dep Len		Height	
						UD	SUD	UD	SUD	UD	SUD
Arabic	arb	PADT	6075	15	1	37	24	4.17	3.92	7.20	9.82
Chinese	cmn	GSD	3997	5	3	37	30	3.72	3.74	4.30	6.56
English	eng	EWT	12543	8	6	20	12	3.13	2.94	3.48	5.11
Basque	eus	BDT	5396	2	13	34	25	2.99	2.90	3.49	4.18
Finnish	fin	TDT	12217	2	7	35	30	2.98	2.91	3.42	4.22
Hebrew	heb	HTB	5241	14	2	28	14	3.76	3.53	5.07	7.30
Hindi	hin	HDTB	13304	22	9	26	10	3.44	3.05	4.25	7.41
Italian	ita	ISDT	13121	14	5	21	8	3.30	3.12	4.21	6.28
Japanese	jap	GSD	7125	25	14	31	10	2.49	2.08	4.40	8.18
Korean	kor	GSD	4400	2	0	58	57	2.20	2.17	3.86	4.07
Russian	rus	SynTagRus	48814	10	1	31	22	3.28	3.13	4.21	5.24
Swedish	swe	Talbanken	4303	12	5	29	17	3.14	2.98	3.50	5.02
Turkish	tur	IMST	3664	3	2	33	30	2.21	2.12	3.01	3.37
Average	-	-	10784.62	12	5	32	22	3.14	3.00	4.20	5.91

Table 1: Treebank statistics: number of sentences (# Sents) and percentage of adpositions (ADP) and auxiliaries (AUX). Comparison of UD and SUD: percentage of direct relations involving only nouns and/or verbs (ContRel); average dependency length (DepLen) and average tree height (Height). Language codes are ISO 639-3.

Algorithm 1 Invoke CLE for sentence $S = w_{1,n}$ given distance matrix E and depth vector D

procedure INVOKECLE(E, D)
  $N \leftarrow |S| + 1$
  $M \leftarrow \text{INIT}(shape = (N, N), value = -\infty)$
  for $(w_i, w_j) \in E$ **do**
    if $D(w_i) < D(w_j)$ **then**
      $M(w_i, w_j) \leftarrow -E(w_i, w_j)$
  $root \leftarrow \mathbf{argmin}_i D(w_i)$
  $M(0, w_{root}) \leftarrow 0$
  return CLE(M)
end procedure

tion framework, we fit the structural probe on the model, layer by layer, using UD and SUD treebanks for supervision, and compute UAS over each treebank's test set as a proxy for a given layer's goodness-of-fit.

Language and Treebank Selection We reuse the sample of Kulmizev et al. (2019), which comprises 13 languages from different language families, with different morphological complexity, and with different scripts. We use treebanks from UD v2.4 (Nivre et al., 2019) and their conversions into SUD.[4] Table 1 shows background statistics for the treebanks, including the percentage of adpositions (ADP) and auxiliaries (AUX), two important function word categories that are treated differently by UD and SUD. A direct comparison of the UD and SUD representations shows that, as expected, UD has a higher percentage of relations directly connecting nouns and verbs (ContRel), higher average dependency length (DepLen) and lower average tree height (Height). However, the magnitude of the difference varies greatly across languages.[5]

Models We evaluate two pretrained language models: BERT (Devlin et al., 2019) and ELMo (Peters et al., 2018a). For BERT, we use the pretrained `multilingual-bert-cased` model provided by Google.[6] The model is trained on the concatenation of WikiDumps for the top 104 languages with the largest Wikipedias and features a 12-layer Transformer with 768 hidden units and 12 self-attention heads. For ELMo, we make use of the pretrained monolingual models made available by Che et al. (2018). These models are trained on 20 million words randomly sampled from the concatenation of WikiDump and CommonCrawl datasets for 44 different languages, including our 13 languages. Each model features a character-based word embedding layer, as well as 2 bi-LSTM layers, each of which is 1024-dimensions wide.

Though we fit the probe on all layers of each model separately, we also learn a weighted average over each full model:

$$\mathbf{model}_i = \sum_{j=0}^{L} s_j \mathbf{h}_{i,j}$$

where s_j is a learned parameter, $\mathbf{h}_{i,j}$ is the encoding of word i at layer j, and L is the number of

[4] https://surfacesyntacticud.github.io/data/

[5] For Chinese, UD actually has slightly lower average dependency length than SUD.

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

layers. We surmise that, in addition to visualizing the probes' fit across layers, this approach will give us a more general notion of how well either model aligns with the respective frameworks. We refer to this representation as the 13th BERT layer and the 3rd ELMo layer. When determining the dimensionality of the transformation matrix (i.e. probe rank), we defer to each respective encoder's hidden layer sizes. However, preliminary experiments indicated that probing accuracy was stable across ranks of decreasing sizes.

It is important to note that by *probe* we henceforth refer to the algorithm that combines both distance and depth probes to return a valid tree. One could argue that, per recent insights in the interpretability literature (e.g. (Hewitt and Liang, 2019)), this model is too expressive in that it combines supervision from two different sources. We do not consider this a problem, as the two probes are trained separately and offer views into two different abstract properties of the dependency tree. As such, we do not optimize for UAS directly.

6 Results and Discussion

Figure 2 displays the UAS after fitting the structural probes on BERT and ELMo, per language and layer. What is perhaps most noticeable is that, while BERT can achieve accuracies upwards of 79 UAS on some languages, ELMo fares consistently worse, maxing out at 65 for Hindi at layer 2. The most likely explanation for this is that the ELMo models are smaller than the multilingual BERT's 12-layer Transformer-based architecture, which was trained on orders of magnitude more data (albeit multilingually).

In general, we find that the probing performance is stable across languages, where layers 7–8 fare the best for BERT and layer 2 for ELMo.[7] This contrasts with prior observations (Tenney et al., 2019a), as the syntactic 'center of gravity' is placed higher in each model's hierarchy. However, computing a weighted average over layers tends to produce the best overall performance for each model, indicating that the probe can benefit from information encoded across various layers.

Once we compare the averaged results across syntactic representations, a preference for UD emerges, starting in layer 3 in BERT and layer 2 in ELMo. We observe the max difference in favor of UD in layer 7 for BERT, where the probe performs 3 UAS points better than SUD, and in the weighted average (layer 13), with 4 UAS points. The difference for the 13th BERT and 3rd ELMo layers is statistically significant at $p \leqslant 0.05$ (Wilcoxon signed ranks test). A further look at differences across languages reveals that, while most languages tend to overwhelmingly prefer UD, there are some that do not: Basque, Turkish, and, to a lesser extent, Finnish. Furthermore, the preference towards SUD in these languages tends to be most pronounced in the first four and last two layers of BERT. However, in the layers where we tend to observe the higher UAS overall (7–8), this is minimized for Basque/Turkish and almost eliminated for Finnish. Indeed, we see the strongest preferences for UD in these layers overall, where Italian and Japanese are overwhelmingly pro-UD, to the order of 10+ UAS points.

6.1 Controlling for Treebank Size

Overall, we note that some languages consistently achieve higher accuracy, like Russian with 71/69 UAS for UD/SUD for BERT, while others fare poorly, like Turkish (52/43) and Chinese (51/46). In the case of these languages, one can observe an obvious relation to the size of our reference treebanks, where Russian is by far the largest and Turkish and Chinese are the smallest. To test the extent to which training set size affects probing accuracy, we trained our probe on the same treebanks, truncated to the size of the smallest one — Turkish, with 3664 sentences. Though we did observe a decline in accuracy in the largest treebanks (e.g. Russian, Finnish, and English) for some layers, the difference in aggregate was minimal. Furthermore, the magnitude of the difference in UD and SUD probing accuracy was almost identical to that of the probes trained on full treebanks, speaking to the validity of our findings. We refer the reader to Appendix A for these results.

6.2 Connection to Supervised Parsing

Given that our findings seem to generally favor UD, another question we might ask is: are SUD treebanks simply harder to parse? This may seem like a straight-forward hypothesis, given SUD's tendency to produce higher trees in aggregate, which may affect parsing accuracy — even in the fully supervised case. To test this, we trained UD and SUD parsers using the UDify model (Kondratyuk

[7] It is important to note that layer 0 for ELMo is the non-recurrent embedding layer which contains no contextual information.

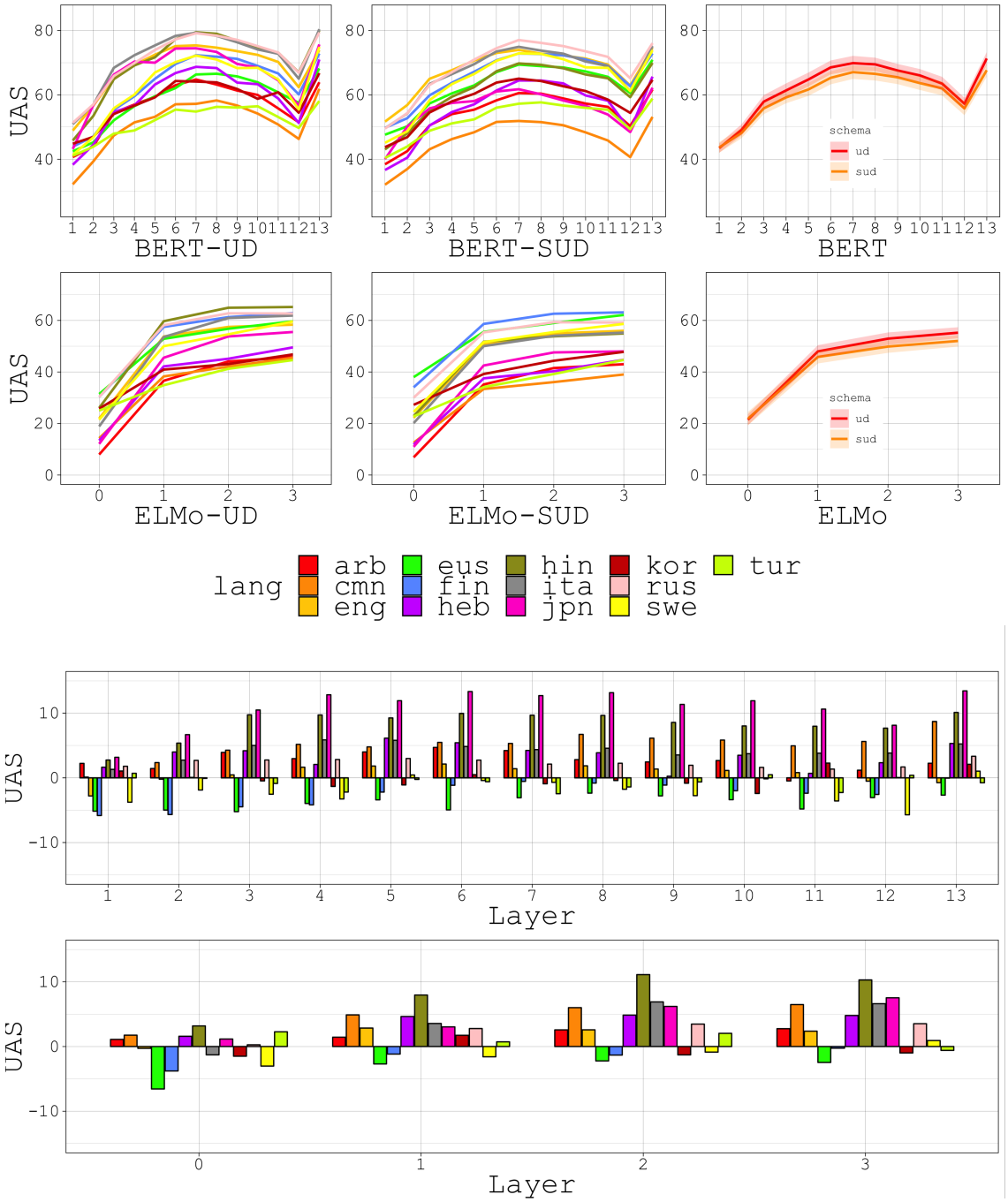

Figure 2: Probe results per model, layer, and language. First two rows depict UAS per layer and language for BERT and ELMo, with average performance and error over UD/SUD in 3rd column. Bottom two rows depict the difference in UAS across UD (+) and SUD (−) per model.

and Straka, 2019), which employs a biaffine attention decoder (Dozat and Manning, 2016) after fine-tuning BERT representations (similar to our 13th layer). The results showed a slightly higher average UAS for UD (89.9 vs. 89.6) and a slightly higher LAS for SUD (86.8 vs. 86.5). Neither difference is statistically significant (Wilcoxon signed ranks test), which seems to rule out an alternative explanation in terms of learnability. We include the full range of results in Appendix B.

In addition to this, we tested how well each framework's probing accuracy related to supervised UAS across languages. We computed this measure by taking the Pearson correlation of each BERT probe's layer accuracy (per-language) with its respective framework accuracy. All correlations proved to be significant at $p \leqslant 0.05$, with the exception of UD and SUD at layer 1. Figure 3 displays

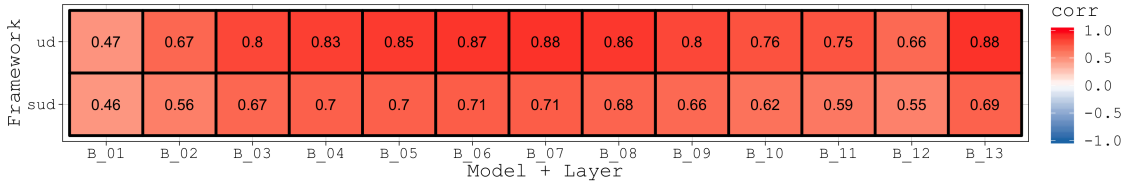

Figure 3: Pearson correlation between UD/SUD probing accuracy and supervised UAS, per layer.

these results. Here, we observe that probing accuracies correlate more strongly with supervised UAS for UD than for SUD. We can interpret this to mean that the rate at which trees are decoded by the UD probe is more indicative of how well they can be parsed given a full view of their structure, rather than vice-versa. Although correlation is an indirect measure here, we can still accept it to be in support of our general findings.

6.3 Parts of Speech

In order to gain a better understanding of these probing patterns, we move on to an error analysis over the dev sets of each treebank, as fit by the averaged models. Figure 4 shows probe accuracy for different models (BERT/ELMo) and syntactic representations (UD/SUD) when attaching words of specific part-of-speech categories to their heads. The general pattern is that we observe higher accuracy for UD for both models on all categories, the only exceptions being a slightly higher accuracy for both models on PRON and for ELMo on VERB and X.[8] However, the differences are generally greater for function words, in particular ADP, AUX, SCONJ, PART and DET. In some respects, this is completely expected given the different treatment of these words in UD and SUD, and we can use the case of adpositions (ADP) to illustrate this. In UD, the preposition *from* in a phrase like *from the room* is simply attached to the noun *room*, which is in general a short relation that is easy to identify. In SUD, the relation between the preposition and the noun is reversed, and the preposition now has to be attached to whatever the entire phrase modifies, which often means that difficult attachment ambiguities need to be resolved. However, exactly the same ambiguities need to be resolved for nominal words (NOUN, PRON, PROPN) in the UD representation, but there is no corresponding drop in accuracy for these classes in UD (except very marginally for PRON). Similar remarks can be made for other function word categories, in particular AUX, SCONJ and PART. It thus seems that the UD strategy of always connecting content words directly to other content words, instead of sometimes having these relations mediated by function words, results in higher accuracy overall when applying the probe to the representations learned by BERT and ELMo.

The behavior of different part-of-speech classes can also explain some of the differences observed across languages. In particular, as can be seen in Table 1, most of the languages that show a clear preference for UD — Chinese, Hebrew, Hindi, Italian and Japanese — are all characterized by a high proportion of adpositions. Conversely, the three languages that exhibit the opposite trend — Basque, Finnish and Turkish — have a very low proportion of adpositions. The only language that does not fit this pattern is Chinese, which has a low percentage of adpositions but nevertheless shows a clear preference for UD. Finally, it is worth noting that Korean shows no clear preference for either representation despite having a very low proportion of adpositions (as well as other function words), but this is due to the more coarse-grained word segmentation of the Korean treebank, which partly incorporates function words into content word chunks.[9]

6.4 Sentence and Tree Properties

Figure 5 depicts probing accuracy across different sentence lengths, dependency lengths, and distances to root. It is apparent that, despite the absolute differences between models, the relative differences between representations are strikingly consistent in favor of UD. For example, while the probe shows identical accuracy for the two representations for sentences of length 1–10, SUD decays more rapidly with increasing sentence length. Furthermore, while the SUD probe is slightly more accurate at detecting sentence roots and their immediate dependencies, we observe a consistent advantage for dependencies of length 2+, until drop-

[8] The X category is unspecified and extremely rare.

[9] This is reflected also in the exceptionally high proportion of direct content word relations; cf. Table 1.

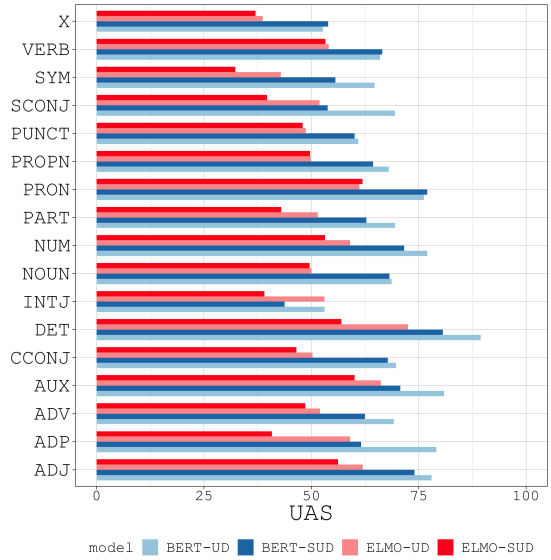

Figure 4: UAS accuracy for the average models (BERT 13, ELMo 3) on incoming dependencies of different part-of-speech categories.

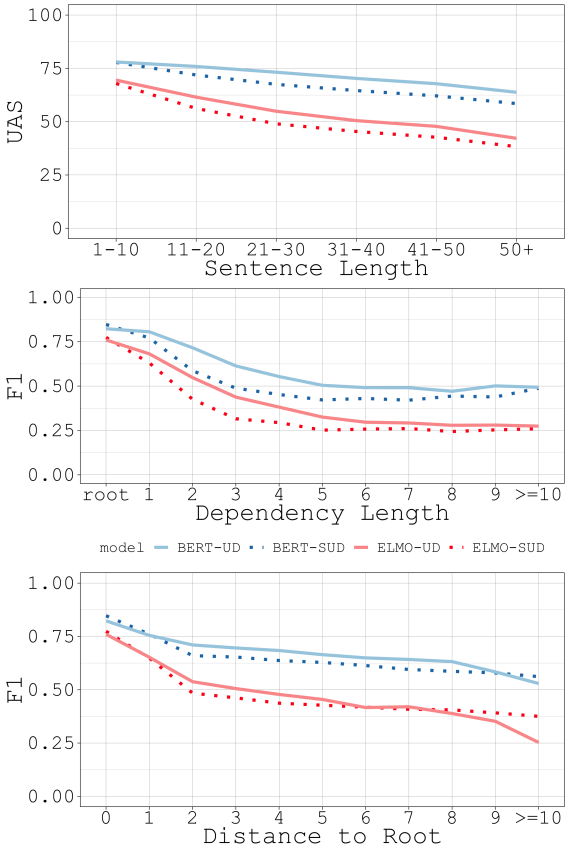

Figure 5: UAS across sentence length bins (top); F1 across varying dependency lengths (middle); F1 across varying distances to root (bottom)

ping off for the longest length bin of 10+. Though Table 1 indicates that UD dependencies are slightly longer than those of SUD, this factor does not appear to influence the probe, as there are no significant correlations between differences in average dependency length and differences in UAS.

We observe a similar curve for varying distances to root, where the SUD probe performs slightly better than UD at the shortest distance, but decays faster for nodes higher in the tree. In general, UD trees have lower height than SUD (see Table 1), which implies that tree height could be a major factor at play here. To verify this, we conducted a Pearson correlation test between the average increase in height from UD to SUD and the difference of the UD/SUD probe UAS per language. This test returned $\rho = 0.82, p < 0.001$, indicating that height is indeed crucial in accurately decoding trees across the two formalisms. In an attempt to visualize how this may play out across languages, we plotted the per-sentence difference in probing accuracy between UD/SUD as a function of the difference in height of the respective gold UD/SUD trees. Figure 6 depicts these results for BERT, where the x-axis indicates how many nodes higher a SUD tree is with respect to its reference UD tree.

It is apparent from Figure 6 that the preference for UD can be largely explained via its lower tree height. If we first examine Korean, the segmentation of which results in the smallest difference in height overall, we observe a distribution that is roughly centered around zero on both axes. If we instead refer to the UD-preferring languages (Chinese, Hebrew, Hindi, Italian, and Japanese), we notice a strong skew of distributions towards the top right of the plot. This indicates (i) that the trees in these samples are higher for SUD and (ii) that the corresponding sentences are easier to decode in UD. By contrast, for the SUD-preferring languages (Basque, Finnish, and Turkish), we observe narrow distributions centered around 0 (similar to that of Korean), indicating minimal variation in tree height between UD and SUD. What these language have in common is an agglutinative morphology, which means that they rely more on morphological inflection to indicate relationships between content words, rather than separate function words. Sentences in these languages are therefore less susceptible to variations in tree height, by mere virtue of being shorter and possessing fewer relations that are likely be a better fit for UD, like those concerning adpositions. We speculate that it is this inherent property that explains the layerwise preference for SUD (though a gen-

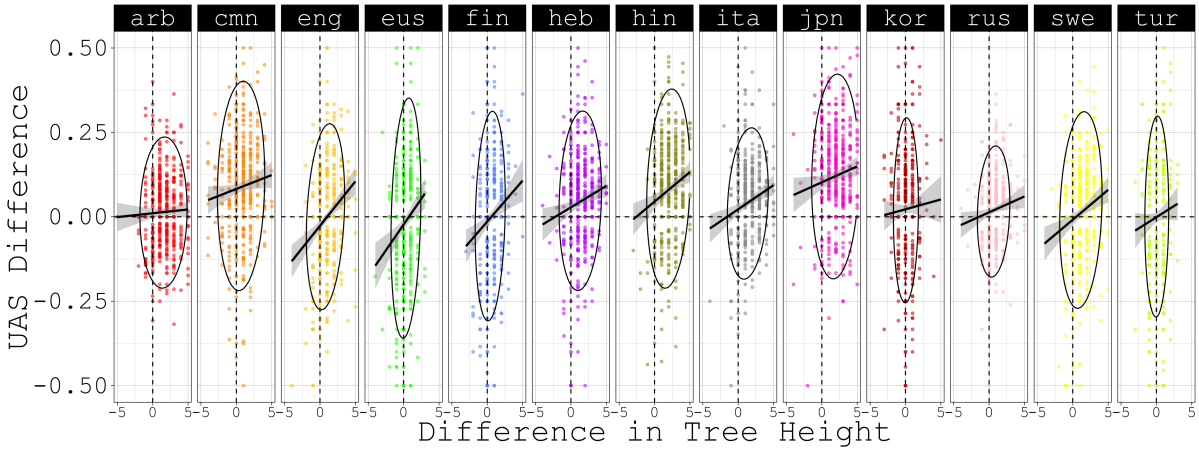

Figure 6: Differences in the BERT probe's UAS (UD +, SUD −) as a function of tree height per number of nodes (higher SUD tree +, higher UD tree −), with smoothed means and 95% confidence ellipses as implemented in `ggplot2`)

eral indifference in aggregate), allowing for some language-specific properties, like the crucial role of auxiliaries in Basque, to be easier to probe for in SUD. Conversely, with this in mind, it becomes easy to motivate the high preference for UD across some languages, given that they are not agglutinating and make heavy use of function words. If we take the probe to be a proper decoding of a model's representational space, the encoding of syntactic structure according to an SUD-style analysis then becomes inherently more difficult, as the model is required to attend to hierarchy between words higher in the tree. Interestingly, however, this does not seem to correspond to an increased difficulty in the case of supervised parsing, as observed earlier.

7 Conclusion and Future Work

We have investigated the extent to which the syntactic structure captured by neural language models aligns with different styles of analysis, using UD treebanks and their SUD conversions as proxies. We have extended the structural probe of Hewitt and Manning (2019) to extract directed, rooted trees and fit it on pretrained BERT and ELMo representations for 13 languages. Ultimately, we observed a better overall fit for the UD-style formalism across models, layers, and languages, with some notable exceptions. For example, while the Chinese, Hebrew, Hindi, Italian, and Japanese models proved to be overwhelmingly better-fit for UD, Basque aligned more with SUD, and Finnish, Korean and Turkish did not exhibit a clear preference. Furthermore, an error analysis revealed that, when attaching words of various part-of-speech tags to their heads, UD fared better across the vast majority of categories, most notably adpositions and determiners. Related to this, we found a strong correlation between differences in average tree height and the tendency to prefer one framework over the other. This suggested a tradeoff between morphological complexity — where differences in tree height between UD and SUD are minimal and probing accuracy similar — and a high proportion of function words — where SUD trees are significantly higher and probing accuracy favors UD.

For future work, besides seeking a deeper understanding of the interplay of linguistic factors and tree shape, we want to explore probes that combine the distance and depth assumptions into a single transformation, rather than learning separate probes and combining them post-hoc, as well as methods for alleviating treebank supervision altogether. Lastly, given recent criticisms of probing approaches in NLP, it will be vital to revisit the insights produced here within a non-probing framework, for example, using Representational Similarity Analysis (RSA) (Chrupała and Alishahi, 2019) over symbolic representations from treebanks and their encoded representations.

Acknowledgements

We want to thank Miryam De Lhoneux, Paola Merlo, Sara Stymne, and Dan Zeman and the ACL reviewers and area chairs for valuable feedback on preliminary versions of this paper. We acknowledge the computational resources provided by CSC in Helsinki and Sigma2 in Oslo through NeIC-NLPL (`www.nlpl.eu`).

References

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

Yonatan Belinkov, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James Glass. 2017. What do neural machine translation models learn about morphology? In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 861–872, Vancouver, Canada. Association for Computational Linguistics.

Terra Blevins, Omer Levy, and Luke Zettlemoyer. 2018. Deep RNNs encode soft hierarchical syntax. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 14–19, Melbourne, Australia. Association for Computational Linguistics.

Joan Bresnan. 2000. *Lexical-Functional Syntax*. Blackwell.

Wanxiang Che, Yijia Liu, Yuxuan Wang, Bo Zheng, and Ting Liu. 2018. Towards better UD parsing: Deep contextualized word embeddings, ensemble, and treebank concatenation. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 55–64.

Grzegorz Chrupała and Afra Alishahi. 2019. Correlating neural and symbolic representations of language. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2952–2962, Florence, Italy. Association for Computational Linguistics.

Y. J. Chu and T. H. Liu. 1965. On the shortest arborescence of a directed graph. *Science Sinica*, 14:1396–1400.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loc 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. Association for Computational Linguistics.

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

Timothy Dozat and Christopher D Manning. 2016. Deep biaffine attention for neural dependency parsing. *arXiv preprint arXiv:1611.01734*.

Allyson Ettinger, Ahmed Elgohary, and Philip Resnik. 2016. Probing for semantic evidence of composition by means of simple classification tasks. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 134–139, Berlin, Germany. Association for Computational Linguistics.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. 2018. Allennlp: A deep semantic natural language processing platform. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, pages 1–6.

Jon Gauthier and Roger Levy. 2019. Linking artificial and human neural representations of language. 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 529–539, Hong Kong, China. Association for Computational Linguistics.

Kim Gerdes, Bruno Guillaume, Sylvain Kahane, and Guy Perrier. 2018. SUD or surface-syntactic universal dependencies: An annotation scheme near-isomorphic to UD. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 66–74.

John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2733–2743, Hong Kong, China. Association for Computational Linguistics.

John Hewitt and Christopher D Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138.

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.

Ron Kaplan and Joan Bresnan. 1982. Lexical-Functional Grammar: A formal system for grammatical representation. In Joan Bresnan, editor, *The Mental Representation of Grammatical Relations*, pages 173–281. MIT Press.

Dan Kondratyuk and Milan Straka. 2019. 75 languages, 1 model: Parsing Universal Dependencies universally. 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 2779–2795, Hong Kong, China. Association for Computational Linguistics.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. 2019. Similarity of neural network representations revisited. *arXiv:1905.00414 [cs, q-bio, stat]*. ArXiv: 1905.00414.

Artur Kulmizev, Miryam de Lhoneux, Johannes Gontrum, Elena Fano, and Joakim Nivre. 2019. Deep contextualized word embeddings in transition-based and graph-based dependency parsing - a tale of two parsers revisited. 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 2755–2768, Hong Kong, China. Association for Computational Linguistics.

Ryan McDonald, Fernando Pereira, Kiril Ribarov, and Jan Hajič. 2005. Non-projective dependency parsing using spanning tree algorithms. pages 523–530.

Igor Mel'čuk. 1988. *Dependency Syntax: Theory and Practice*. State University of New York Press.

Joakim Nivre, Mitchell Abrams, Željko Agić, Lars Ahrenberg, Gabrielė Aleksandravičiūtė, Lene Antonsen, Katya Aplonova, Maria Jesus Aranzabe, Gashaw Arutie, Masayuki Asahara, Luma Ateyah, Mohammed Attia, Aitziber Atutxa, Liesbeth Augustinus, Elena Badmaeva, Miguel Ballesteros, Esha Banerjee, Sebastian Bank, Verginica Barbu Mititelu, Victoria Basmov, John Bauer, Sandra Bellato, Kepa Bengoetxea, Yevgeni Berzak, Irshad Ahmad Bhat, Riyaz Ahmad Bhat, Erica Biagetti, Eckhard Bick, Agnė Bielinskienė, Rogier Blokland, Victoria Bobicev, Loïc Boizou, Emanuel Borges Völker, Carl Börstell, Cristina Bosco, Gosse Bouma, Sam Bowman, Adriane Boyd, Kristina Brokaitė, Aljoscha Burchardt, Marie Candito, Bernard Caron, Gauthier Caron, Gülşen Cebiroğlu Eryiğit, Flavio Massimiliano Cecchini, Giuseppe G. A. Celano, Slavomír Čéplö, Savas Cetin, Fabricio Chalub, Jinho Choi, Yongseok Cho, Jayeol Chun, Silvie Cinková, Aurélie Collomb, Çağrı öltekin, Miriam Connor, Marine Courtin, Elizabeth Davidson, Marie-Catherine de Marneffe, Valeria de Paiva, Arantza Diaz de Ilarraza, Carly Dickerson, Bamba Dione, Peter Dirix, Kaja Dobrovoljc, Timothy Dozat, Kira Droganova, Puneet Dwivedi, Hanne Eckhoff, Marhaba Eli, Ali Elkahky, Binyam Ephrem, Tomaž Erjavec, Aline Etienne, Richárd Farkas, Hector Fernandez Alcalde, Jennifer Foster, Cláudia Freitas, Kazunori Fujita, Katarína Gajdošová, Daniel Galbraith, Marcos Garcia, Moa Gärdenfors, Sebastian Garza, Kim Gerdes, Filip Ginter, Iakes Goenaga, Koldo Gojenola, Memduh Gökırmak, Yoav Goldberg, Xavier Gómez Guinovart, Berta González Saavedra, Matias Grioni, Normunds Grūzītis, Bruno Guillaume, Céline Guillot-Barbance, Nizar Habash, Jan Hajič, Jan Hajič jr., Linh Hà Mỹ, Na-Rae Han, Kim Harris, Dag Haug, Johannes Heinecke, Felix Hennig, Barbora Hladká, Jaroslava Hlaváčová, Florinel Hociung, Petter Hohle, Jena Hwang, Takumi Ikeda, Radu Ion, Elena Irimia, Ọlájídé Ishola, Tomáš Jelínek, Anders Johannsen, Fredrik Jørgensen, Hüner Kaşıkara, Andre Kaasen, Sylvain Kahane, Hiroshi Kanayama, Jenna Kanerva, Boris Katz, Tolga Kayadelen, Jessica Kenney, Václava Kettnerová, Jesse Kirchner, Arne Köhn, Kamil Kopacewicz, Natalia Kotsyba, Jolanta Kovalevskaitė, Simon Krek, Sookyoung Kwak, Veronika Laippala, Lorenzo Lambertino, Lucia Lam, Tatiana Lando, Septina Dian Larasati, Alexei Lavrentiev, John Lee, Phng Lê H`ông, Alessandro Lenci, Saran Lertpradit, Herman Leung, Cheuk Ying Li, Josie Li, Keying Li, Kyung-Tae Lim, Yuan Li, Nikola Ljubešić, Olga Loginova, Olga Lyashevskaya, Teresa Lynn, Vivien Macketanz, Aibek Makazhanov, Michael Mandl, Christopher Manning, Ruli Manurung, Cătălina Mărănduc, David Mareček, Katrin Marheinecke, Héctor Martínez Alonso, André Martins, Jan Mašek, Yuji Matsumoto, Ryan McDonald, Sarah McGuinness, Gustavo Mendonça, Niko Miekka, Margarita Misirpashayeva, Anna Missilä, Cătălin Mititelu, Yusuke Miyao, Simonetta Montemagni, Amir More, Laura Moreno Romero, Keiko Sophie Mori, Tomohiko Morioka, Shinsuke Mori, Shigeki Moro, Bjartur Mortensen, Bohdan Moskalevskyi, Kadri Muischnek, Yugo Murawaki, Kaili Müürisep, Pinkey Nainwani, Juan Ignacio Navarro Horñiacek, Anna Nedoluzhko, Gunta Nešpore-Bērzkalne, Lng Nguy˜ên Thị, Huy`ên Nguy˜ên Thị Minh, Yoshihiro Nikaido, Vitaly Nikolaev, Rattima Nitisaroj, Hanna Nurmi, Stina Ojala, Adédayọ̀ Olúòkun, Mai Omura, Petya Osenova, Robert Östling, Lilja Øvrelid, Niko Partanen, Elena Pascual, Marco Passarotti, Agnieszka Patejuk, Guilherme Paulino-Passos, Angelika Peljak-Łapińska, Siyao Peng, Cenel-Augusto Perez, Guy Perrier, Daria Petrova, Slav Petrov, Jussi Piitulainen, Tommi A Pirinen, Emily Pitler, Barbara Plank, Thierry Poibeau, Martin Popel, Lauma Pretkalniņa, Sophie Prévost, Prokopis Prokopidis, Adam Przepiórkowski, Tiina Puolakainen, Sampo Pyysalo, Andriela Rääbis, Alexandre Rademaker, Loganathan Ramasamy, Taraka Rama, Carlos Ramisch, Vinit Ravishankar, Livy Real, Siva Reddy, Georg Rehm, Michael Rießler, Erika Rimkutė, Larissa Rinaldi, Laura Rituma, Luisa Rocha, Mykhailo Romanenko, Rudolf Rosa, Davide Rovati, Valentin Roca, Olga Rudina, Jack Rueter, Shoval Sadde, Benoît Sagot, Shadi Saleh, Alessio Salomoni, Tanja Samardžić, Stephanie Samson, Manuela Sanguinetti, Dage Särg, Baiba Saulīte, Yanin Sawanakunanon, Nathan Schneider, Sebastian Schuster, Djamé Seddah, Wolfgang Seeker, Mojgan Seraji, Mo Shen, Atsuko Shimada, Hiroyuki Shirasu, Muh Shohibussirri, Dmitry Sichinava, Natalia Silveira, Maria Simi, Radu Simionescu, Katalin Simkó,

Mária Šimková, Kiril Simov, Aaron Smith, Isabela Soares-Bastos, Carolyn Spadine, Antonio Stella, Milan Straka, Jana Strnadová, Alane Suhr, Umut Sulubacak, Shingo Suzuki, Zsolt Szántó, Dima Taji, Yuta Takahashi, Fabio Tamburini, Takaaki Tanaka, Isabelle Tellier, Guillaume Thomas, Liisi Torga, Trond Trosterud, Anna Trukhina, Reut Tsarfaty, Francis Tyers, Sumire Uematsu, Zdeňka Urešová, Larraitz Uria, Hans Uszkoreit, Sowmya Vajjala, Daniel van Niekerk, Gertjan van Noord, Viktor Varga, Eric Villemonte de la Clergerie, Veronika Vincze, Lars Wallin, Abigail Walsh, Jing Xian Wang, Jonathan North Washington, Maximilan Wendt, Seyi Williams, Mats Wirén, Christian Wittern, Tsegay Woldemariam, Tak-sum Wong, Alina Wróblewska, Mary Yako, Naoki Yamazaki, Chunxiao Yan, Koichi Yasuoka, Marat M. Yavrumyan, Zhuoran Yu, Zdeněk Žabokrtský, Amir Zeldes, Daniel Zeman, Manying Zhang, and Hanzhi Zhu. 2019. Universal Dependencies 2.4. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal Dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666, Portorož, Slovenia. European Language Resources Association (ELRA).

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018a. Deep contextualized word representations. 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 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Matthew Peters, Mark Neumann, Luke Zettlemoyer, and Wen-tau Yih. 2018b. Dissecting contextual word embeddings: Architecture and representation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1499–1509, Brussels, Belgium. Association for Computational Linguistics.

Grusha Prasad, Marten van Schijndel, and Tal Linzen. 2019. Using priming to uncover the organization of syntactic representations in neural language models. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 66–76, Hong Kong, China. Association for Computational Linguistics.

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. 2017. SVCCA: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In *Advances in Neural Information Processing Systems*, pages 6076–6085.

Naomi Saphra and Adam Lopez. 2018. Understanding learning dynamics of language models with SVCCA. *arXiv preprint arXiv:1811.00225*.

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*, pages 4593–4601.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R. Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R. Bowman, Dipanjan Das, and Ellie Pavlick. 2019b. What do you learn from context? probing for sentence structure in contextualized word representations. *arXiv:1905.06316 [cs]*. ArXiv: 1905.06316.

A Controlling for Treebank Size

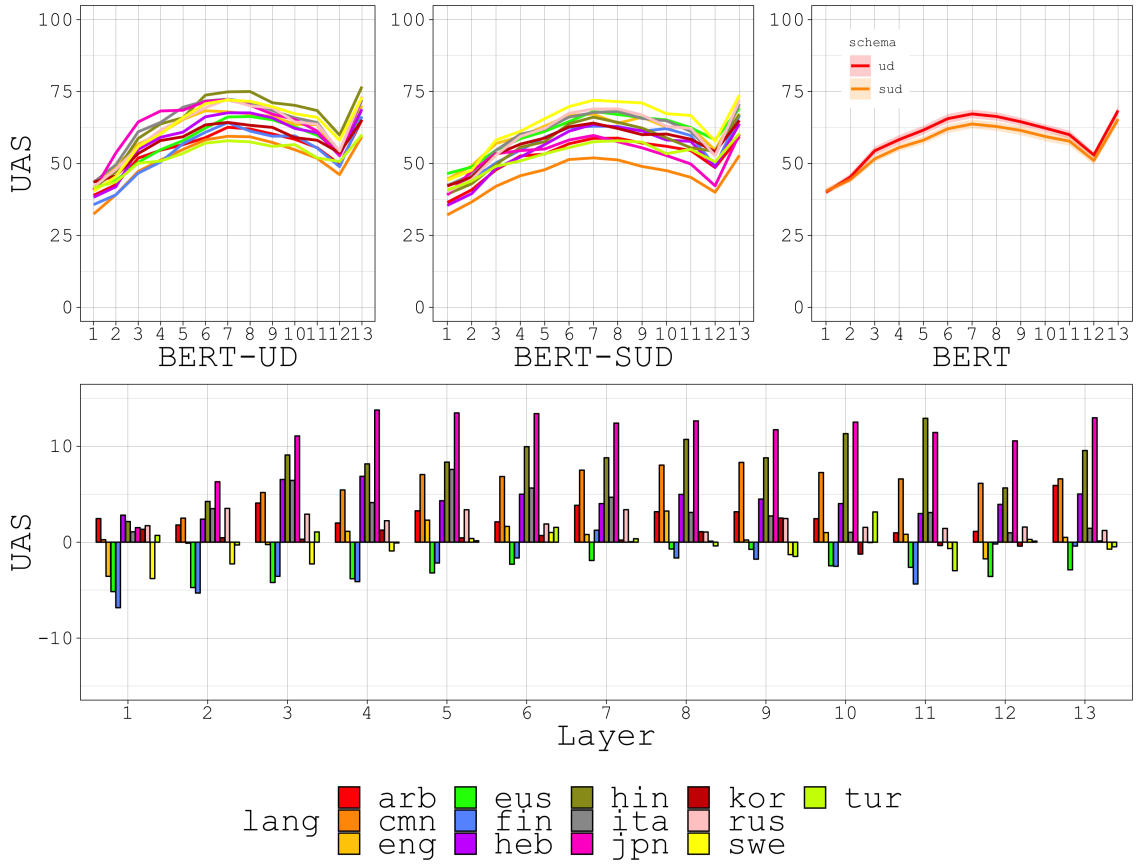

Figure 7: Probe results per framework, layer, and language, when trained on 3664 sentences. First row depicts UAS per layer and language for BERT, with average performance and error over UD/SUD in 3rd column. Bottom two row depicts the difference in UAS across UD (+) and SUD (−).

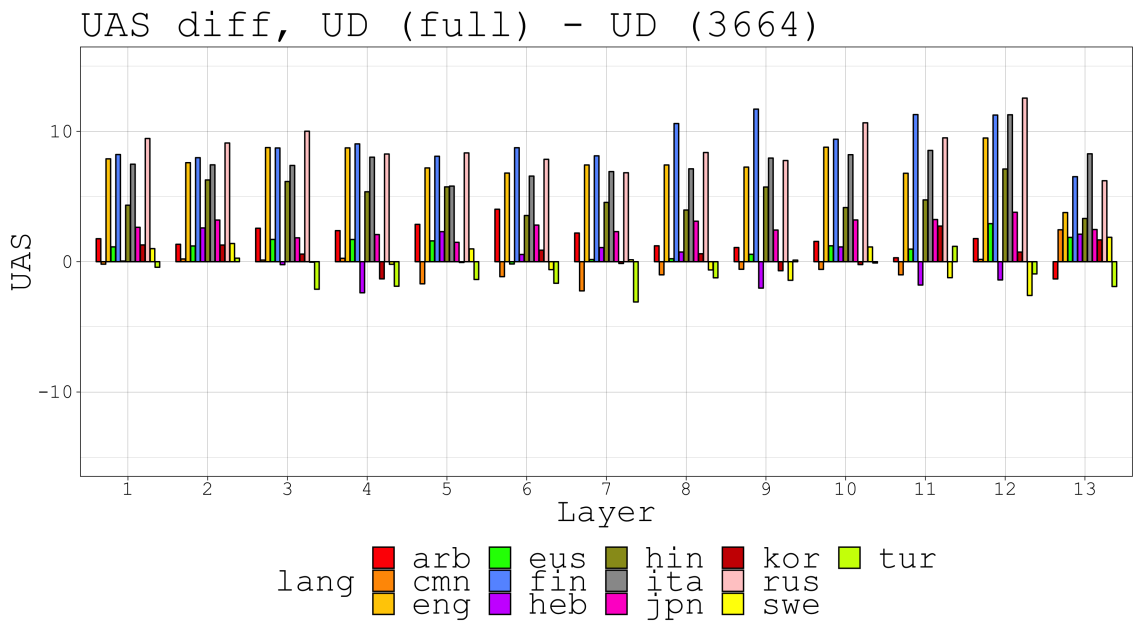

Figure 8: Difference in UAS across the UD probes trained on full data (+) and 3664 sentences (−).

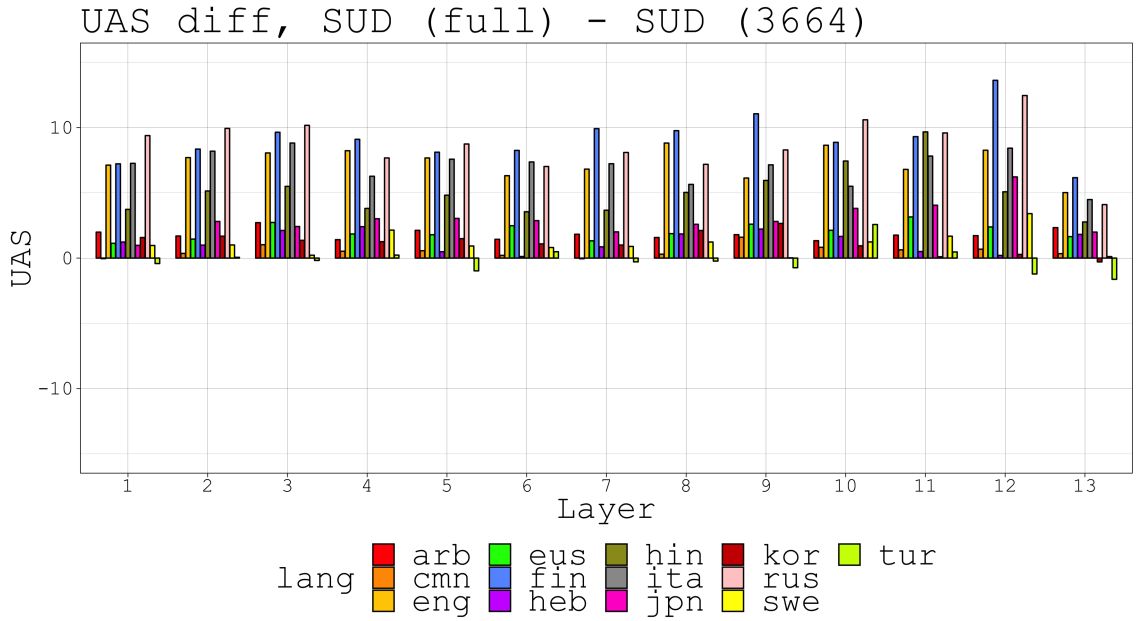

Figure 9: Difference in UAS across the SUD probes trained on full data (+) and 3664 sentences (−).

B Connection to Supervised Parsing

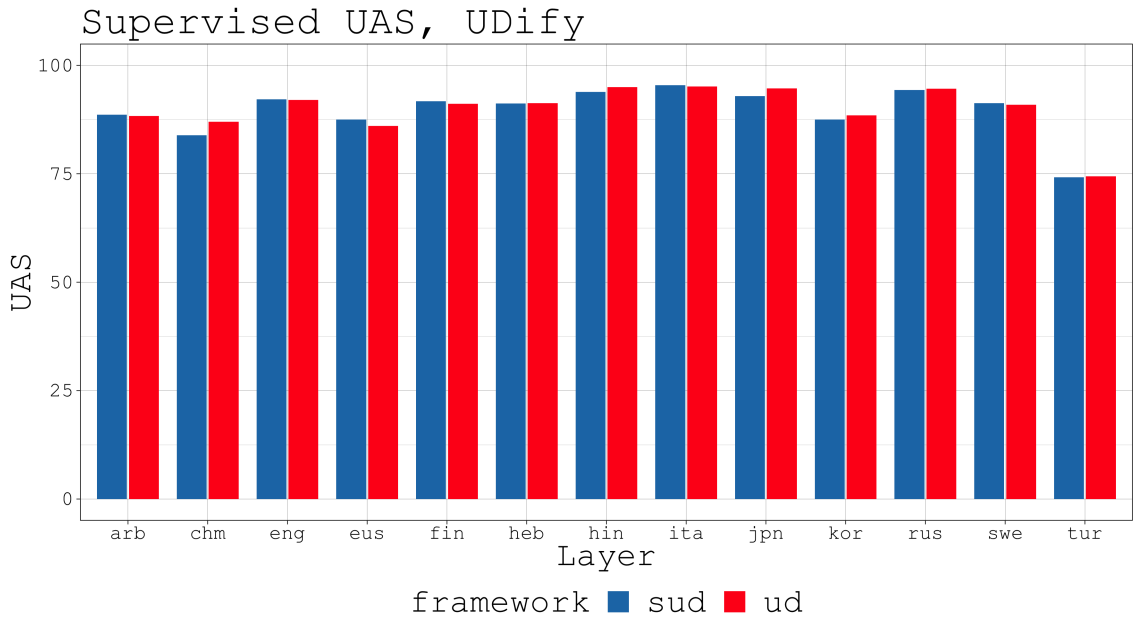

Figure 10: Supervised UDify UAS, UD and SUD, for all languages.

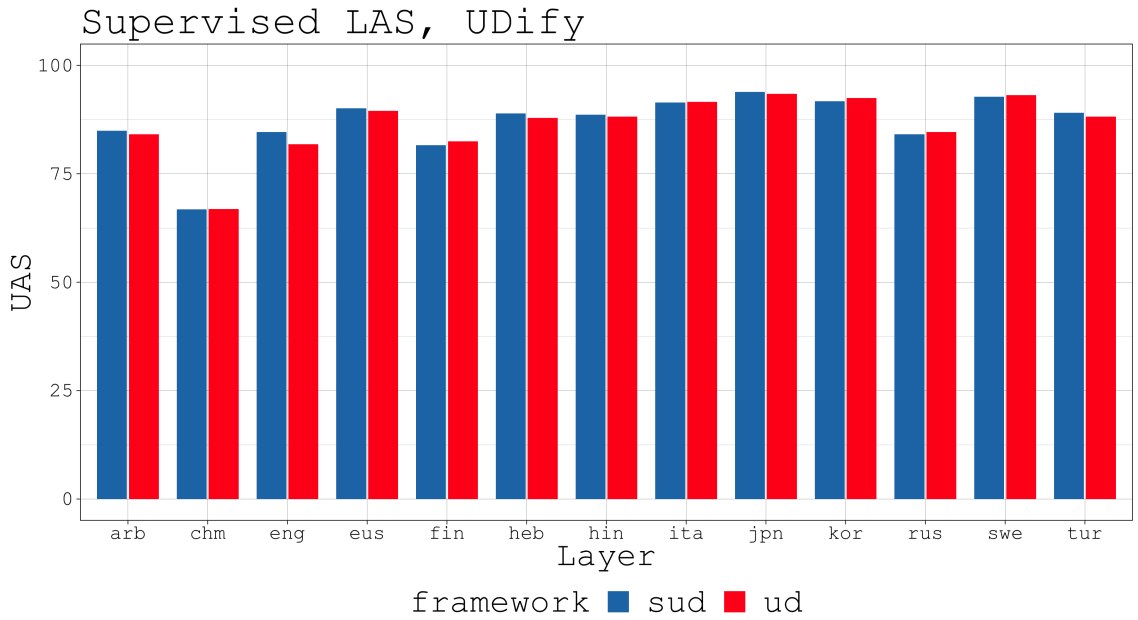

Figure 11: Supervised Udify LAS, UD and SUD, for all languages.