

The Rosetta Probe: Cross-Lingual Syntactic Transfer in Monolingual English BERT

Stanford CS224N Custom Project

Ananya Navale

Department of Computer Science
Stanford University
anavale@stanford.edu

Abstract

If a language model learns the grammatical rules of one language, does it implicitly learn something universal about language itself? We investigate this question by asking what English BERT, trained exclusively on English text, latently encodes about the syntax of other languages. Understanding how BERT encodes syntactic structure can reveal whether language models capture deeper linguistic relationships beyond surface-level context. Prior work has demonstrated that English BERT embeddings encode syntactic parse tree structure in a linear geometric space, recoverable by a trained structural probe. However, this finding was limited to English, leaving open the question of whether this geometric encoding generalizes to languages that BERT was never exposed to during training. In this work, we apply the structural probe framework to eight languages—English, French, Spanish, Italian, German, Dutch, Czech, and Finnish—using a single frozen English BERT model across all languages. We find that English BERT’s syntactic geometry transfers cross-lingually to a surprising degree, with probe performance (and consequently syntax recoverability) degrading as typological distance from English increases. Furthermore, we observe a consistent pattern in which Germanic languages outperform Romance languages on root identification accuracy, suggesting that the geometric encoding reflects language family structure in addition to typological proximity.

1 Key Information to include

- Mentor: Julie Kallini
- External Collaborators (if you have any): None
- Sharing project: No
- Late Days (if you are using any):

Team member	Late days remaining
Ananya Navale	2
Total pooled	2
Team size	1
Late days applied for the report	2

2 Introduction

BERT is a historically acclaimed language model whose high-quality word embeddings have driven strong performance in a wide range of NLP tasks. Understanding the extent of linguistic structure implicitly encoded in these embeddings is an active area of research. Prior work (Hewitt and Manning, 2019; Chi et al., 2020) has established two relevant findings: that English BERT implicitly

encodes English syntactic structure in a recoverable linear geometric space, and that multilingual BERT (mBERT) encodes syntax across many languages simultaneously (Devlin et al., 2019). Yet an unexplored middle ground remains: can a monolingual English BERT model, never exposed to other languages, nonetheless encode syntactic structure recoverable from multilingual treebanks? By applying Hewitt and Manning’s structural probe framework to seven non-English languages, we find clear evidence that English BERT’s syntactic geometry transfers cross-lingually, with transfer quality varying systematically across language families and typological distance from English.

Identifying any underlying cross-linguistic syntax capabilities of English BERT embeddings would suggest that the model may be more language-agnostic than previously understood. This could also reveal that BERT captures some syntactic similarities and word order patterns across language families, indicating that there may exist a more universal syntax that is sufficiently encodable with exposure to only a single language such as English. Such findings could additionally indicate that current approaches to multilingual NLP may not be effectively utilizing these underlying cross-lingual ties encoded in monolingual models.

There is no existing evidence to indicate the degree to which these ties would appear in the encoding of cross-lingual syntax. Even if syntactic transfer is possible, it is unclear whether transfer will pattern consistently within language families or whether cross-family patterns might emerge that typological distance alone cannot predict. Additionally, the probe utilized in this study is a single linear measure of syntactic structure—this may not be the only or most accurate method to detect syntax in BERT embeddings.

Hewitt and Manning (2019) created a foundation with their original English-trained structural probe run on English BERT embeddings. Chi et al. (2020) expanded on this work to explore a similar relationship between a SINGLETRAN probe trained on 11 languages and mBERT embeddings based on data from 104 languages. The original probe tested both distance (the number of edges between word pairs on a parse tree) and depth (the number of edges from the root of a dependency tree to a given word), while the multilingual probe tested only distance; both learn a linear transformation matrix $\mathbf{B}$ that maps word representations into a geometric space where syntactic structure is recoverable. This leaves two clear intermediate avenues for further inquiry: (1) how well does an English-trained probe recover English syntax from a multilingual model, and (2) how well does a non-English language-specific probe recover that language’s syntax from a frozen English model—the latter is what we chose to analyze.

Using the original Hewitt and Manning structural probe framework, this time incorporating both the distance and depth probe in contrast to Chi et al., we train the probe separately on eight languages (French, German, Spanish, Czech, Italian, Dutch, Finnish, and including English as an in-language baseline) using Universal Dependencies treebanks (Nivre et al., 2016) and run each probe individually using English BERT embeddings of the corresponding language’s sentences. The results show an intuitive decrease in extraction proportional to the typological distance of a given language from English; however, we also find that sentence root identification breaks this trend by producing stronger Root% scores in Germanic languages as opposed to Romance languages. These results are subject to a few limitations: all languages were restricted to 11K training sentences, and the linear probe framework tests only one possible method of syntactic encoding in a single layer of English BERT, leaving open questions about other layers and probing approaches.

3 Related Work

In the late 2010s, probing classifiers emerged as a standard methodology for analyzing the linguistic properties encoded in neural network representations. Conneau et al. (2018) established a foundational probing methodology by training simple classifiers on top of frozen sentence representations to test what linguistic properties they encode. Using a suite of 10 probing tasks covering surface, syntactic, and semantic properties, they demonstrate that different encoder architectures vary significantly in the linguistic information they capture. Belinkov and Glass (2019) survey neural NLP model analysis and interpretation methods, including probing classifiers as introduced by Conneau et al. While noting that a probe’s success indicates that a property is encoded in representations rather than necessarily used by the model, probing remains a widely adopted diagnostic tool for understanding

neural representations. This leads to Hewitt and Manning (2019)’s structural probe and the basis for this study. Using separate distance and depth probes on monolingual English BERT embeddings, they find that syntactic information is linearly encoded and recoverable through a learned transformation.

However, this work was limited to English, leaving open the question of cross-lingual syntactic encoding. Pires et al. (2019) establish that multilingual representations derived from mBERT can transfer cross-lingually to a certain degree, notably that transfer carries over more effectively for typologically similar languages yet is less effective for languages with very different structures. These findings are supplemented by the results discovered by Wu and Dredze (2019) through task analysis that indicate that mBERT performs effectively in cross-lingual zero-shot transfer. Additionally, they find that the percentage of subword sharing is a key component of transfer performance; this observation raises interesting questions about whether lexical overlap between languages might influence syntactic transfer in monolingual models. Taking multilingual model analysis one step further, Chi et al. (2020) find that mBERT embeddings encode recoverable cross-lingual syntactic structure in a linear geometric space, extending Hewitt and Manning’s structural probe framework to a multilingual setting through a single transformation trained across 11 languages simultaneously. This creates a second bookend that complements Hewitt and Manning’s findings, leaving an intermediate gap for investigation of multilingual syntax encoding in monolingual model embeddings.

In efforts to locate specific layers within these models that encode individual linguistic features, Tenney et al. (2019) find the mirroring between the BERT encoding process and the traditional NLP pipeline: lower level linguistic features are encoded first, higher level features later. Jawahar et al. (2019) provide additional specificity to these results, identifying that intermediate layers encode the richest syntactic information. Notably, they also find that BERT’s comprehension of linguistic information mimics a tree-like structure, aligning with research techniques involving parse or dependency tree comparisons.

Our work targets the middle ground between Hewitt and Manning and Chi et al., examining the cross-lingual syntactic structure encoded by monolingual English BERT and utilizing the middle layers of BERT-base (particularly layer 7) for access to the strongest syntactically grounded embeddings.

4 Approach

4.1 Probe

The Hewitt and Manning structural probe is a learned linear transformation $\mathbf{B}$ that maps input BERT embeddings into a lower-dimensional syntactic subspace via a matrix of shape $k \times m$, where $k = 64$ is the rank of the projected syntactic subspace and $m = 768$ is the BERT hidden dimension. Critically, $\mathbf{B}$ is the only learned component of the probe—BERT’s weights are frozen throughout training, meaning any syntactic structure recovered by the probe was already latently encoded in BERT’s representations. The distance and depth probes operate with slight variation in their calculations. For the distance probe, the difference between every word pair in a sentence is computed, transformed by $\mathbf{B}$, and squared to find the squared L2 distance using the equation:

$$d^2(\mathbf{h}_i, \mathbf{h}_j) = (\mathbf{B}(\mathbf{h}_i - \mathbf{h}_j))^T (\mathbf{B}(\mathbf{h}_i - \mathbf{h}_j)) \quad (1)$$

where $\mathbf{h}_i \in \mathbb{R}^m$ is the BERT embedding for word i , $\mathbf{h}_j \in \mathbb{R}^m$ is the BERT embedding for word j , and $\mathbf{B} \in \mathbb{R}^{k \times m}$ is the learned probe transformation matrix. The probe compares each resulting scalar against the gold parse tree distance between the corresponding word pair in the dependency treebank, calculating L1 loss for a sentence of length n averaged over all word pairs:

$$\mathcal{L}_{\text{dist}} = \frac{1}{n^2} \sum_{i,j} |d_{\text{pred}}^2(\mathbf{h}_i, \mathbf{h}_j) - d_{\text{gold}}^2(i, j)| \quad (2)$$

The depth probe operates similarly, computing the squared L2 norm of each word’s embedding:

$$\|\mathbf{h}_i\|^2 = (\mathbf{B}\mathbf{h}_i)^T (\mathbf{B}\mathbf{h}_i) \quad (3)$$

where $\mathbf{h}_i$ is as defined above. The word with the minimum predicted norm is taken as the predicted root, which is used to evaluate Root% accuracy. All word depths are compared against gold depths, with L1 loss averaged across all words in a sentence:

$$\mathcal{L}_{\text{depth}} = \frac{1}{n} \sum_i \left| \|\mathbf{h}_i\|_{\text{pred}}^2 - \text{depth}_{\text{gold}}(i) \right| \quad (4)$$

In both cases, the Adam optimizer computes gradients and updates **B**’s weights to minimize loss, while BERT remains frozen.

4.2 Model

We use the monolingual English BERT (specifically bert-base-cased) model to extract embeddings on which the probe is run; this particular model is trained exclusively on English data, making any cross-lingual syntactic transfer a non-trivial finding. While BERT-base contains 12 layers, we performed all probing on layer 7, which Hewitt and Manning (2019) found to encode the richest syntactic structure.

4.3 Pipeline

The experimental pipeline spans four sequential steps, from raw CoNLL-U treebank files to final probe evaluation scores.¹ First, the CoNLL-U files cloned from the individual treebank codebases are converted to raw text (.txt) files for easy BERT tokenization. CoNLL-U files may contain multi-word tokens with hyphenated indices (e.g. '1-2') representing surface contractions, as well as empty nodes with decimal indices (e.g. '1.1'); these are excluded during conversion as they lack independent syntactic annotations. Monolingual English BERT processes these text files to produce contextual word embeddings, stored in HDF5 format. Prior to probe training, since BERT’s wordpiece tokenizer may split a single word into multiple subword tokens, the embeddings of all subword tokens belonging to the same word are averaged into a single word-level representation, ensuring alignment between BERT’s output and the treebank’s word-level syntactic annotations. Finally, the probe is trained and tested on the resulting embeddings to optimize the **B** transformation matrix, generate and compare predictions against gold parse trees, and output resulting evaluation scores.

4.4 LINEAR Baseline

We use a single LINEAR baseline to provide a point of comparison against the learning capabilities of the probe in order to assess whether positional information alone is sufficient to recover syntactic structure, without any learned representations. This baseline predicts parse distances between words simply as the absolute difference in their positional indices $|i - j|$, based on the assumption that words situated nearby within the context of a sentence are likely to be closer on a parse tree as well. An MST is then constructed from these distances to create a predicted parse tree. The LINEAR baseline serves as a lower bound—the simplest possible predictor of parse structure. From the original study, the LINEAR baseline parse trees were defined as left-to-right chains, while our cross-lingual context more accurately defines the LINEAR as a positional distance baseline.

5 Experiments

Using the Hewitt and Manning structural probe, we evaluate whether monolingual English BERT embeddings encode non-English syntactic information. The probe is trained and tested on eight languages (English, French, German, Spanish, Czech, Italian, Dutch, and Finnish), assessing distance and depth separately for each. The results of each experiment are compared with the corresponding lower bound LINEAR baseline outputs.

5.1 Data

Universal Dependencies treebanks provide a consistent cross-lingual framework for grammatical annotation (parts of speech, morphological features, and syntactic dependencies). Our probe is trained and evaluated on eight Universal Dependencies treebanks; each is a widely regarded gold-standard resource for monolingual syntactic analysis. These languages and treebanks were selected based primarily on their typological distance from English, usage of Latin script, and dataset size. Languages were additionally grouped by family (Germanic, Romance, Slavic, and Uralic) to facilitate analysis of family-level transfer patterns. All training sets were set to an upper limit of 11K sentences for consistency and to avoid dataset size as a confounding variable in cross-lingual performance differences. Pre-processing across the board involved truncating training sets larger than 11K sentences

¹Code available at <https://github.com/AnanyaNavale/cs224n-rosetta-probe>.

and removal of sentences exceeding BERT’s maximum input length of 512 tokens. Additionally, the Czech Treebank originally included a variety of treebank training sets, separated based on subject domain including newspaper articles, parliamentary proceedings, and academic texts; to prevent bias based on success with a specific domain, we sampled proportionally across each training set to build a new compiled set with equal relative domain variation.

Table 1: Dataset statistics. Train/Dev/Test columns indicate number of sentences.

Language	Treebank	Full Name	Family	Train	Dev	Test
English	UD_English-EWT	English Web Treebank	Germanic	11,000	2,001	2,077
German	UD_German-GSD	German Google Stanford Dependencies	Germanic	11,000	799	977
Dutch	UD_Dutch-Alpino	Alpino Treebank	Germanic	11,000	718	596
French	UD_French-GSD	French Google Stanford Dependencies	Romance	10,999	1,476	416
Spanish	UD_Spanish-AnCor	Spanish AnCor Corpus	Romance	11,000	1,654	1,721
Italian	UD_Italian-ISDT	Italian Stanford Dependency Treebank	Romance	10,998	564	482
Czech	UD_Czech-PDTC	Prague Dependency Treebank - Consolidated	Slavic	11,000	2,000	2,000
Finnish	UD_Finnish-TDT	Turku Dependency Treebank	Uralic	11,000	1,363	1,555

5.2 Evaluation method

Probe performance is measured with distinct metrics based on the feature being evaluated. The distance probe is assessed through two metrics: the Undirected Unlabeled Attachment Score (UUAS) and Distance Spearman Correlation (DSpr.). The depth probe is also assessed through two metrics: the Root% score and Norm Spearman Correlation (NSpr.), where ‘Norm’ refers to the L2 norm computed by this probe. The UUAS metric represents the percentage of undirected edges in a predicted parse tree that match the edges of a gold parse tree. Predicted parse trees are constructed using a minimum spanning tree (MST) algorithm over the predicted word-pair distances, and the resulting edges are compared against gold parse tree edges. The Root% metric is an evaluation method implemented by Hewitt and Manning (2019) that emerges naturally from the depth probe’s predictions—the word with the minimum predicted norm is taken as the predicted root; this score represents the percentage of test sentences from which the probe correctly identifies the root. Punctuation tokens (identified via the universal UPOS PUNCT tag rather than language-specific XPOS tags) are excluded from both the UUAS and Root% metrics for cross-lingual consistency. The DSpr. is computed by ranking predicted word-pair distances, while NSpr. is computed by ranking predicted single-word norms. The ranks for a single sentence are then compared to gold rankings using the Spearman correlation coefficient:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \quad (5)$$

where d_i is the difference between the predicted and gold ranks for observation i , and n is the number of observations in the sentence. Finally, correlation scores are averaged across sentence lengths between 5 and 50 to minimize the influence of outlier sentences on aggregate correlation scores.

5.3 Experimental details

The probe is trained to convergence independently with its own instance for each language. It utilizes a learning rate of 0.001 in batch sizes of 20 for a maximum of 40 epochs, or until dev L1 loss does not improve by more than 0.0001 for 5 consecutive epochs. Final reported results are returned from the epoch with the lowest dev loss, not necessarily the last epoch. The linear transformation matrix $\mathbf{B}$ weights are updated using an Adam optimizer. A ReduceLROnPlateau scheduler reduces the learning rate by a factor of 0.1 when dev loss fails to improve. The probe is run at rank $k = 64$, the dimensionality of the projected syntactic subspace, which Hewitt and Manning (2019) found to be an effective upper bound on parsing performance gains.

5.4 Results

While UUAS, DSpr., and NSpr. scores follow an intuitive decline from Romance to Germanic to Slavic to Uralic languages compared to English, Root% scores display an unexpected reversal, with Germanic languages outperforming Romance languages. This was better than expected, considering the monolingual English BERT model was unaware of the existence of other languages during

training. These results suggest that English BERT’s syntactic geometry is partially language-agnostic, with transfer quality modulated by typological proximity and language family membership.

Table 2: Structural probe results across eight languages, grouped by language family.

Language	Family	UUAS	DSpr.	Root%	NSpr.
English	Germanic	78.5	0.81	85.4%	0.84
German	Germanic	62.3	0.73	61.1%	0.73
Dutch	Germanic	58.5	0.72	58.5%	0.71
French	Romance	72.9	0.79	57.5%	0.78
Italian	Romance	71.5	0.78	50.7%	0.75
Spanish	Romance	66.9	0.77	56.2%	0.76
Czech	Slavic	57.6	0.70	53.3%	0.67
Finnish	Uralic	56.6	0.69	48.5%	0.62

Table 3: LINEAR baseline results across eight languages, grouped by language family.

Language	Family	UUAS	DSpr.
English	Germanic	43.1	0.66
German	Germanic	42.7	0.59
Dutch	Germanic	41.0	0.62
French	Romance	44.4	0.65
Italian	Romance	45.5	0.68
Spanish	Romance	44.9	0.67
Czech	Slavic	47.3	0.66
Finnish	Uralic	50.8	0.66

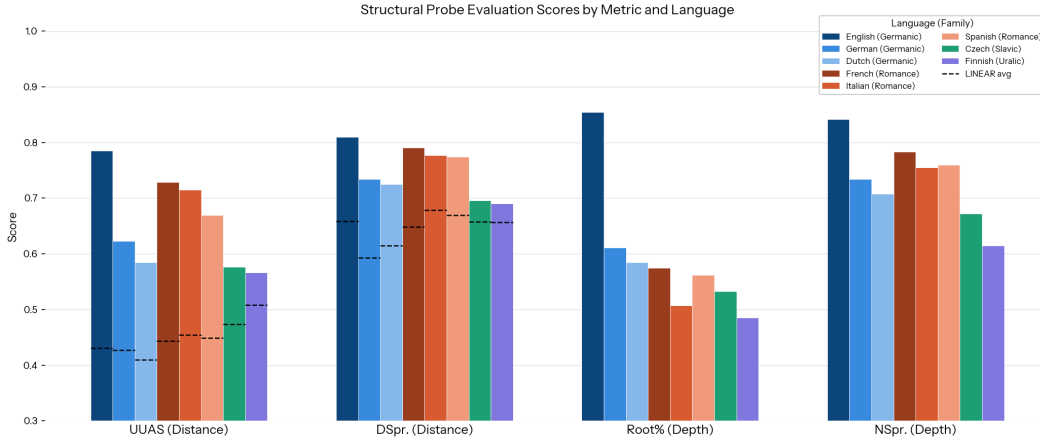


Figure 1: Structural probe evaluation scores across eight languages grouped by language family. Germanic languages are shown in blue, Romance in orange, Czech (Slavic) in teal, and Finnish (Uralic) in purple. Dashed lines on UUAS and DSpr. indicate the corresponding LINEAR baseline score for each language.

6 Analysis

The distance and depth probes reveal divergent patterns of cross-lingual syntactic transfer, suggesting that English BERT’s geometric encoding captures distinct aspects of syntax that generalize differently across language families.

Across the UUAS and DSpr. metrics of the distance probe, Romance languages (French, Italian, Spanish) consistently outperform Germanic non-English languages (German, Dutch), despite Germanic

languages being typologically closer to English. This may be attributed to the shared SVO word order between English and Romance languages (Dryer and Haspelmath, 2013) —French, Spanish, and Italian all follow subject-verb-object structure similar to English, resulting in more consistent syntactic distances between word pairs. Germanic languages like German and Dutch, by contrast, employ verb-second (V2) and subordinate clause SOV ordering, which may disrupt the geometric distance predictions of an English-trained probe.

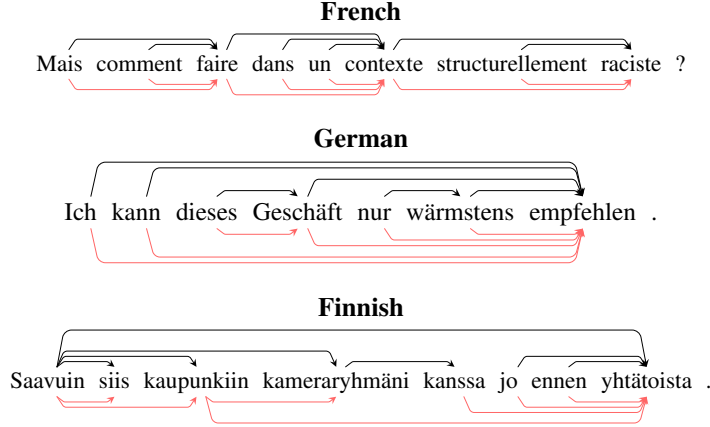


Figure 2: Parse tree examples across three languages (French UUAS = 1.00, German UUAS = 0.83, Finnish UUAS = 0.57). Black edges (above) = gold; red edges (below) = predicted.

In contrast, while the NSpr. metrics of the depth probe show similar trends, the pattern reverses with the Root% metric—Germanic languages outperform Romance languages on root identification and depth ranking. This may reflect shared clause-level structural properties between English and Germanic languages, where the position of the root verb relative to its dependents more closely mirrors English syntax than in Romance languages. Romance languages exhibit pro-drop and clitic constructions (Dryer and Haspelmath, 2013) that may obscure root identification, while Germanic V2 word order, despite disrupting distance predictions, may preserve structural properties relevant to depth encoding that are more similar to English.

An additional hypothesis for the Romance language advantage in distance probe performance concerns lexical overlap between English and Romance languages. English has historically borrowed heavily from French and Latin, meaning many Romance language words share surface forms with English vocabulary (e.g. 'nation', 'administration', 'communication'). Since BERT's wordpiece vocabulary was built exclusively from English text, Romance language words that share orthographic forms with English words may receive similar tokenizations and thus similar embeddings to their English counterparts. This shared subword representation could partially explain why Romance languages transfer more effectively than Germanic languages on distance metrics, and conversely, why Finnish — with essentially no lexical overlap with English — shows the weakest transfer of all tested languages.

The Finnish results present the smallest gap between probe UUAS (56.6%) and LINEAR UUAS (50.8%) of all eight tested languages—the probe barely outperforms its LINEAR baseline. This might be a strong indication that English BERT adds very little syntactic information for Finnish beyond what word order alone predicts; this is consistent with Finnish being the most typologically distant language in our study: agglutinative, non-Indo-European, and with syntactic properties that differ fundamentally from English. The probe is essentially learning a random transformation that happens to do slightly better than chance.

Czech, a Slavic language that is the most typologically distant Indo-European language in our study, sits between the Romance/Germanic languages and Finnish. This may signal a shared underlying syntactic structure across Indo-European languages not shared with languages beyond these borders. However, due to Czech's highly flexible, relatively "free" word order enabled by its complex declension/conjugation system, where cases indicate grammatical roles rather than position, it falls closer to Finnish than its closest Indo-European neighbors in both distance and depth.

Across all languages, the probe consistently and substantially outperforms the LINEAR baseline through the UUAS metric (except Finnish, where the gap is minimal). However, the DSpr. gap is smaller; LINEAR DSpr. is surprisingly high for Romance languages (0.65-0.68), suggesting word order alone explains a great deal of distance structure for this language family. Overall, this confirms that English BERT is encoding genuine syntactic information beyond positional information, but the degree varies by language.

7 Conclusion

In this study, we applied the Hewitt and Manning structural probe to determine whether monolingual English BERT embeddings encode recoverable syntactic information across seven non-English languages spanning Germanic, Romance, Slavic, and Uralic families that use a Latin script. The distance probe finds a consistent trend across UUAS and DSpr. scores, with Romance languages scoring highest and closest to English, followed by Germanic languages, and lastly Slavic and Uralic. The depth probe replicates this trend for NSpr.; however, Root% scores reveal an unexpected reversal, with Germanic languages outperforming Romance languages on root identification.

These results are based purely on a single linear probe extraction technique, while further exploration into non-linear syntax recovery techniques may reveal that linear encoding is not the most expressive method for cross-lingual syntactic extraction. Additionally, we train and run all probes on embeddings taken from BERT-base layer 7 based on the results from Hewitt and Manning’s original study; however, it may be possible that multilingual syntax encoding is higher on a different layer of the model. All training datasets were capped at 11K sentences for consistency; future work could explore whether larger training sets improve cross-lingual transfer. Finally, this study evaluated only languages using Latin script to avoid excessive character-wise BERT tokenization; non-Latin scripts may produce lower quality BERT tokenizations, as bert-base-cased was trained exclusively on English text.

Several promising avenues remain for future research, including probing more Latin script and non-Latin script languages using the same pipeline, as well as running the probe on other monolingual BERT models with their respective language treebanks (e.g. French GSD treebank with CamemBERT) to determine whether the transfer pattern generalizes beyond English BERT. More critically, we encourage a deeper linguistic analysis of the Root% anomaly found in Germanic languages to identify additional methods for testing syntax recoverability.

References

- Yonatan Belinkov and James Glass. Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 2019.
- Ethan A. Chi, John Hewitt, and Christopher D. Manning. Finding universal grammatical relations in multilingual BERT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020.
- Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 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*, 2018.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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*, 2019.
- Matthew S. Dryer and Martin Haspelmath. WALS online, 2013. URL <https://wals.info/>.
- John Hewitt and Christopher D. Manning. 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*, 2019.
- Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019.

- Joakim Nivre et al. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, 2016.
- Telmo Pires, Eva Schlinger, and Dan Garrette. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019.
- Ian Tenney, Dipanjan Das, and Ellie Pavlick. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019.
- Shijie Wu and Mark Dredze. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, 2019.

Appendix

Individual Metric Score Charts

The following figures present probe and LINEAR baseline scores broken down by individual metric.

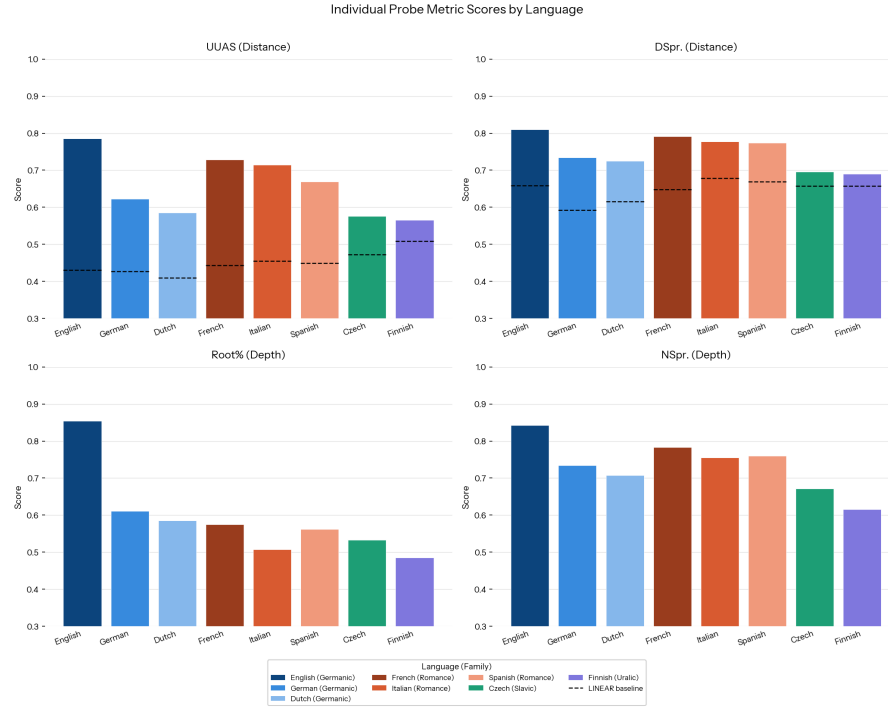


Figure 3: Individual probe metric scores by language. Dashed lines on UUAS and DSpr. indicate the LINEAR baseline score for each language. No LINEAR baseline was computed for Root% and NSpr.

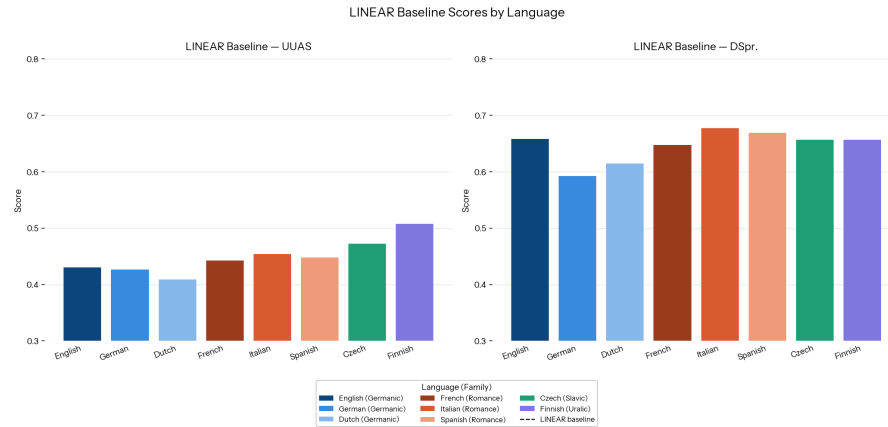


Figure 4: LINEAR baseline scores by language for UUAS and DSpr.

Parse Tree Examples

The following figure presents one representative distance probe parse tree example per language, illustrating prediction accuracy across the typological spectrum from English to Finnish.

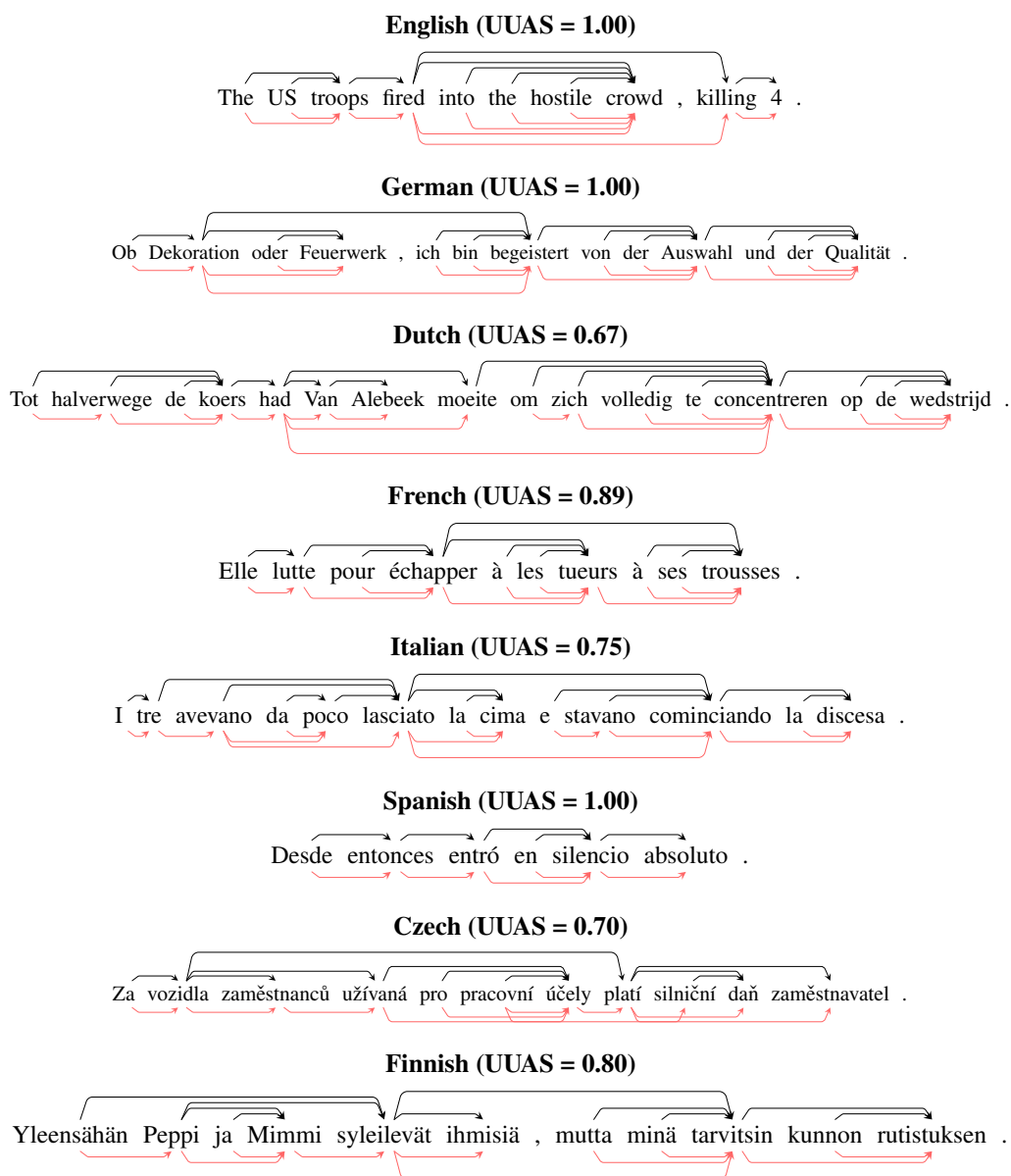


Figure 5: Parse tree examples for all eight languages. Black edges (above) = gold; red edges (below) = predicted. Examples were selected for sentence length and readability; overall performance statistics are reported in Tables 2 and 3.

Distance and Depth Probe Visualizations

The following figure presents gold and predicted distance heatmaps alongside depth prediction graphs for four representative languages, illustrating the varying degrees of cross-lingual syntactic transfer.

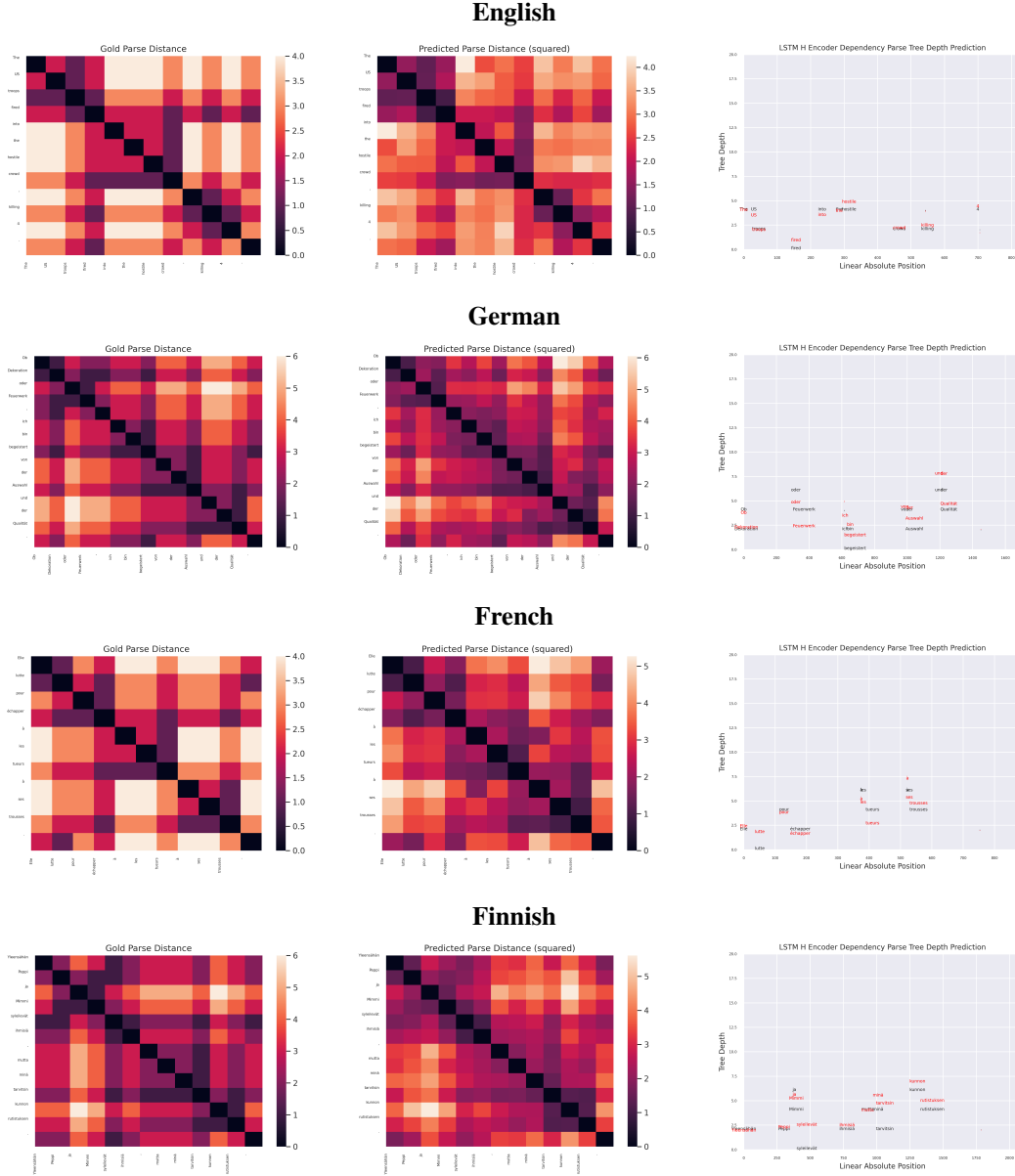


Figure 6: Distance probe heatmaps (gold and predicted) and depth probe graphs for English, German, French, and Finnish sentences (English sentence 14, German sentence 8, French sentence 6, Finnish sentence 6). In distance heatmaps, darker colors indicate larger pairwise distances between words. In depth graphs, black text shows gold depths and red text shows predicted depths.