Is German secretly a Slavic language? What BERT probing can tell us about language groups

Aleksandra Mysiak and **Jacek Cyranka**
Faculty of Mathematics, Informatics, and Mechanics
University of Warsaw
an.mysiak@student.uw.edu.pl, jcyranka@gmail.com

Abstract

In the light of recent developments in NLP, the problem of understanding and interpreting large language models has gained a lot of urgency. Methods developed to study this area are subject to considerable scrutiny. In this work, we take a closer look at one such method, the structural probe introduced by Hewitt and Manning (2019). We run a series of experiments involving multiple languages, focusing principally on the group of Slavic languages. We show that probing results can be seen as a reflection of linguistic classification, and conclude that multilingual BERT learns facts about languages and their groups.

1 Introduction

Transformers (Vaswani et al., 2017) have revolutionised the area of natural language processing. State-of-the-art solutions for virtually all NLP problems – including machine translation, text summarization and generation – are nowadays transformer-based. In recent years models such as BERT (Devlin et al., 2019) and Generative Pre-trained Transformers (Radford et al., 2018) have shifted the public view of artificial intelligence. This is also true for Slavic languages – for example, the Polish language understanding benchmark KLEJ (Rybak et al., 2020) is dominated by models such as HerBERT (Mroczkowski et al., 2021) or Polish RoBERTa (Dadas et al., 2020).

This success has led to a significant interest in studying the interpretability of such models. Multiple probing techniques have been developed to assess the extent of linguistic knowledge learned in masked language modelling, especially by models based on BERT. Those methods typically feature a set of secondary tasks that are learned by a smaller model (the *probe*), using BERT's embeddings as inputs.

Using probing with multiple tasks, Tenney et al. (2019) and Jawahar et al. (2019) have found a surprisingly regular structure encoded in BERT's layers. Their results are supported by Hewitt and Manning (2019), where the authors use the task of dependency tree prediction in a method they call the *structural probe*. They use it to find evidence of syntax learning, especially exhibited by BERT's middle layers. Going a step further, authors of Chi et al. (2020) apply structural probing to a multilingual version of BERT (Devlin et al., 2019), and find a degree of universality in how the syntactic relations are encoded in a single embedding space for multiple languages.

On the other hand, the interpretability of probing results is the subject of much discussion. Although authors typically use a baseline to quantify what the probe actually learned, those results are still called into question. A parameter-free method of probing is introduced by Wu et al. (2020), although the results prove to be much more conservative.

The problem of whether probes extract knowledge from embeddings or learn new tasks is discussed in depth by Hewitt and Liang (2019), where they are shown to be able to learn randomly generated control tasks. In Niu et al. (2022), the authors find a strong argument against interpreting accuracy as a measure of information contained. They show that performance drops when more layers become accessible to the probe, which theoretically should provide it with more information.

In this work, we aim to investigate the usability of probing techniques – specifically the structural probe of Hewitt and Manning (2019) – by relating them to real-life ideas developed by theoretical linguists, such as the classification of languages into families and word order types. We take a closer look at the group of Slavic languages and the claim that they constitute a separate word order class, as proposed by Haider and Szucsich (2022).

1.1 Main Contributions

Inspired by Chi et al. (2020), we investigate probing in a multilingual context, focusing our attention on relations between syntax encoding for a group of Slavic languages. We show that probing results can be related to pre-existing linguistic knowledge, which suggests that, in spite of interpretability problems, this methodology can be used to discover quantitative relations between languages.

To highlight the role of mBERT pre-training in recovering grammatical relations differentiating between language families, we contrast our findings with the results of a randomised baseline. In Table 2, we show that an identical architecture with random parameters does not uncover similar patterns. This suggests that the pre-training task of masked language modeling constructs the embedding space in a way that allows meaningful investigation of relations between languages.

2 Methodology

Our methodology is based on the structural probing method introduced in Hewitt and Manning (2019) and applied to a multilingual setting in Chi et al. (2020).

In this method, the most important data form is the dependency tree, which is a formal way of representing a sentence's syntax. Each word in a sentence is represented by a node, with (directed and labeled) edges indicating syntactical relations between words they connect.

The authors' idea is to find the structure of dependency trees in BERT's embedding space. To recover the structure of a tree, they aim to find a metric in the embedding space that approximates the distance between words in dependency trees (expressed as the number of edges). They search for an appropriate geometry in the family of linear transformations of the embeddings. Our loss function (L) thus becomes

$$L(B) = \sum_\ell \frac{1}{\left|s^\ell\right|^2} \sum_{i,j} \left| d_{T^\ell}\left(w_i^\ell, w_j^\ell\right) - d_B\left(\mathbf{h}_i^\ell, \mathbf{h}_j^\ell\right) \right|$$

where

- $\{s^\ell\}$ is the set of training sentences,
- $\left|s^\ell\right|$ is the sentence length,
- w_i^ℓ is the i-th word of s^ℓ,
- $d_{T^\ell}\left(w_i^\ell, w_j^\ell\right)$ is the number of edges between w_i^ℓ and w_j^ℓ in the sentence's dependency tree,
- $\mathbf{h}_i^\ell$ is the contextualized embedding of word w_i^ℓ in sentence s^ℓ, taken from a BERT layer with a fixed index,
- B is a real matrix of shape (probe rank, embedding dimension),
- d_B is the squared Euclidean distance between vectors transformed by B, that is

$$d_B\left(\mathbf{h}_i^\ell, \mathbf{h}_j^\ell\right) = \left\| B\mathbf{h}_i^\ell - B\mathbf{h}_j^\ell \right\|_2^2 .$$

We can thus see that the real probe here is the matrix B, which is found by minimizing the loss using gradient descent.

Evaluation We assess the probes based on their ability to predict the structures of unseen dependency trees. For that, we utilise two metrics defined in Hewitt and Manning (2019).

The first metric is Spearman's rank correlation coefficient between predicted and gold standard distances (originally named distance Spearman, or "DSpr."). The coefficient is designed to measure monotonicity of a relation between two variables. Here, it is calculated separately for each sentence, averaged across all sentences of a given length, and then over lengths between 5 and 50. The coefficient is expressed as

$$\rho(X, Y) = \frac{\operatorname{cov}(\mathrm{R}(X), \mathrm{R}(Y))}{\sigma_{\mathrm{R}(X)}\sigma_{\mathrm{R}(Y)}}$$

where R is a ranking function, cov is a standard covariance, and σ is standard deviation.

The second metric is the UUAS – undirected, unlabeled attachment score. It requires construction of predicted undirected trees, which is done in an iterative process, based on a ranking of predicted distances. In each step, two words for which the embeddings are predicted to be the closest are connected, unless that would violate the tree property (that is, only if a path between them does not yet exist). This procedure is conducted until a spanning tree of the sentence is constructed. It is then evaluated by calculating the percentage of correctly placed edges, which gives us a value from range $[0, 100]$.

To give a sense of scale here, in Hewitt and Manning (2019) a non-contextualised baseline reaches a score of 26.8, and a randomly contextualised one – 59.8, while the highest value reached on BERT is 82.5, indicating over 82% of correctly predicted edges.

We collected values of both UUAS and DSpr. Since we found that both metrics are highly correlated (Pearson's $r > 0.97$) and lead to identical qualitative conclusions, our reporting focuses on the UUAS, which is easily interpretable as a percentage of successses.

Datasets In our work, we selected two groups of languages: train and test languages, listed in Table 1. The test set is a subset of the group of Slavic languages, with some additional non-Slavic languages added in the train set. For each of the languages, we source our data – manually annotated dependency trees – from the Universal Dependencies project (Nivre et al., 2017).

Language	Size	Train	Test	Slavic
Belarusian	22852	✓	✓	✓
Chinese	3996	✓		
Croatian	6913	✓	✓	✓
Czech	68494	✓	✓	✓
English	12542	✓		
Finnish	12216	✓		
French	14448	✓		
German	13813	✓		
Indonesian	4481	✓		
Latvian	12520	✓		
Lithuanian	2340	✓		
Polish	17721	✓	✓	✓
Russian	69629	✓	✓	✓
Slovak	8482	✓	✓	✓
Slovene	10902	✓	✓	✓
Spanish	14286	✓		
Ukrainian	5495	✓	✓	✓

Table 1: All considered languages, with dataset sizes in number of sentences. Note that the set of test languages is a subset of the train set.

Experimental setup We conduct all experiments at layer 7 (out of 1 - 12) of mBERT base, with a fixed probe rank of 128. Since our goal is not to investigate the properties of mBERT itself, but the properties of probing methodology and relations between languages, we do not consider the whole set of hyperparameters used in Hewitt and Manning (2019). We choose hyperparameters that were found to be optimal in Chi et al. (2020).

To balance the differences in dataset sizes – see Table 1 – and investigate the impact of those differences, we introduce an additional hyperparameter of dataset size. We consider subsets of 100, 1k, 2.5k, 5k, 7.5k and 10k sentences (where available).

Baseline To differentiate between the impact of probe training and mBERT pre-training, we utilise the mBERTRand baseline as described in Chi et al. (2020). In this setup, we run experiments on an mBERT-like architecture with randomly initialized parameters and no pre-training. As such, this baseline should not carry any linguistic information, other than what is learned by the probe itself.

In our setup of the baseline, we only consider a single test language - Polish - since the results were deemed to prove satisfactorily that pretraining enhances linguistic knowlege – see Section 4. The list of train languages remains the same.

3 Experimental results

3.1 Dataset size study

In Figures 1 and 2, we present averaged UUAS scores for probes trained on several dataset sizes and languages, all tested on Polish. In both cases, we can see a saturation of the score for datasets of 10k sentences – the score curves flatten out.

We can also see that the ranking of languages stabilizes, with minor changes between size 7.5k and 10k. For both mBERT and the baseline, it becomes well established that the best train language for Polish is Polish – which is not the case for smaller sizes, especially for 1k sentences and less. In the case of Belarusian, the maximum considered size is necessary to separate it from non-Slavic languages.

Non-baseline results for other test languages were similar, so the plots were omitted here. All numerical results can be found in Table 2 and Appendix A.

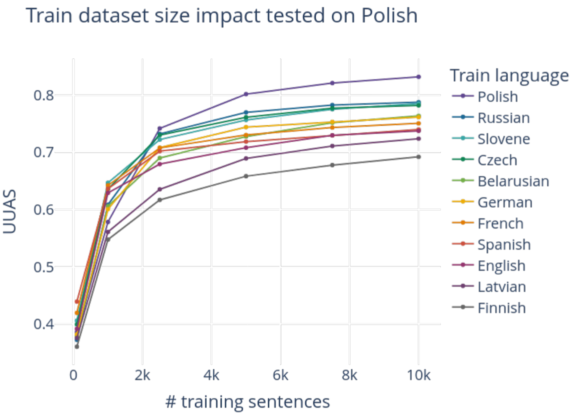

Figure 1: Plot of UUAS scores for probes trained on various languages and dataset sizes, tested on Polish, averaged across 3 independent runs. Higher values indicate better syntax recall.

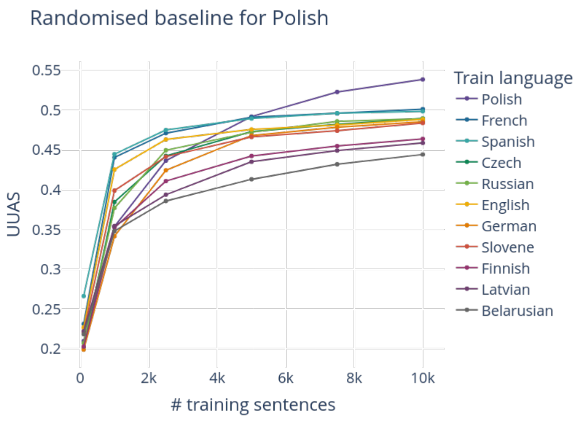

Figure 2: Plot of baseline UUAS scores for probes trained on various languages and dataset sizes, tested on Polish, averaged across 3 independent runs. Word embeddings here are randomly initialised, so the probe cannot access BERT's knowledge. Higher values indicate better syntax recall.

3.2 Relations between Slavic languages

Numerical results (averaged UUAS values) for training datasets of size 10k (the maximal considered) are shown in Table 2. The columns represent all test languages, with 2 additional columns for baseline results and an average across all test languages. The rows represent train languages, they are sorted by the Average column. Only the train languages with at least 10k sentences are shown. For additional languages with smaller sizes see Appendix A.

In non-baseline results, we can see a naturally emerging separation between Slavic and non-Slavic languages. There are significant (> 1 UUAS point in this context) score gaps in a couple of positions in the ranking: between Belarusian and other Slavic languages, between German and Belarusian, between German and other non-Slavic languages, and at the bottom of the ranking, between Finnish and other languages.

The baseline results are not statistically significantly correlated with non-baseline results tested on Polish, except for the visible dominance of Polish as the best train language. Excluding Polish from both rankings, we get $p = 0.38$, with $p = 0.04$ without the exclusion. We can see that the ranking here would be vastly different, with the top train languages being Polish, French, Spanish, and Czech. The bottom language is Belarusian, with a significantly worse result than any other language.

The experiments were executed using two RTX 2080 Ti GPU units (or equivalent). 2816 experiments were carried out in total, with an average experiment with 10k train sentences taking 16 minutes.

4 Discussion

The results for pre-trained mBERT described in the previous section and shown in Table 2 can be related to the following linguistic facts:

- For each test language, the set of top 5 train languages is exactly the same – it is the set of all Slavic languages present in train data for the given dataset size. The group of Slavic languages is recognised as inter-related.
- For each test language, the top-scoring non-Slavic train language is German. This can be related to a matter of discussion raised by Haider and Szucsich (2022) and referred to in Fuß (2022). Haider and Szucsich (2022) propose a new class of word order in languages, to which they postulate that all Slavic languages should belong. They also mention the fact that Germanic languages evolved from a grammar of the same type, which might explain the high scores of German as a predictor of Slavic languages' sentence structure.
- The Finnish language is the worst-scoring train language for all test languages. This can be related to the fact that it is the only language present in the train set that does not belong to the Indo-European family.

There is no such interpretation to be found for baseline results. As noted in the previous section, those results are not correlated with non-baseline results for Polish. In Figure 2 and Table 2, we can see Slavic languages mixed with non-Slavic languages, with no visible separation even for large dataset sizes. Except for the fact that Polish is the highest-scored train language, there is no clear relation between linguistic classification and the results of the baseline. We conclude that pre-training of mBERT plays a vital role in the ability of the probe to reproduce the well-known classification of Slavic languages.

Additionally, we can note that for main results, the scores achieved using the same train and test language differ between languages, ranging from 78.82 (Belarusian) to 83.19 (Polish). Although in

	Baseline	Slovene	Russian	Polish	Czech	Belarusian	Croatian	Slovak	Ukrainian	Average
Slavic languages										
Slovene	48.39	81.43	76.16	78.52	**77.10**	75.19	**77.60**	82.26	77.08	**78.17**
Russian	48.95	75.35	81.32	**78.77**	76.75	**76.88**	76.08	80.54	**79.41**	78.14
Polish	53.87	75.64	**76.84**	83.19	77.07	75.91	75.03	81.20	77.74	77.83
Czech	48.96	**76.02**	76.37	78.20	80.47	74.96	75.90	**83.24**	77.45	77.83
Belarusian	44.44	72.88	75.54	76.38	73.94	78.82	73.36	77.97	76.99	75.73
Non-Slavic languages										
German	48.56	**73.17**	**74.62**	**76.15**	**74.23**	**73.08**	**73.17**	**78.17**	**75.20**	**74.72**
English	48.86	70.34	73.08	73.75	71.42	70.40	72.03	75.79	73.36	72.52
French	50.14	70.20	72.22	75.07	71.10	70.93	71.84	74.57	73.01	72.37
Latvian	45.88	70.84	70.97	72.39	70.69	70.59	69.99	75.41	72.01	71.61
Spanish	49.86	69.64	71.12	73.99	70.20	69.70	70.57	72.55	71.66	71.18
Finnish	46.40	68.09	68.33	69.22	68.07	67.97	67.43	72.14	68.87	68.77

Table 2: Average UUAS scores for probes trained using 10k sentences. The test languages are in columns, and the train languages in rows. Higher values indicate better syntax recall and suggest syntactic similarity, with top results highlighted in each column. The results are averaged over three independent runs with different random seeds. Standard deviations of results are not reported, since values are below 1 UUAS point.

each case, the test language is also the best train language, the score values differ. This can be interpreted as a reflection of the fact that mBERT learns certain languages' representations more clearly, especially when coupled with results from Chi et al. (2020) and Alves et al. (2022). However, this could also be an artefact of dataset differences between languages – their quality, diversity and representativeness – which brings our attention back to the fact that interpretability remains an issue in probing.

5 Conclusions

Using Slavic languages as an example, we have shown that the method of structural probing can be used to achieve results that are clearly related to pre-existing linguistic knowledge. In spite of interpretability problems, we conclude that probing can be used to extract linguistic knowledge from transformer models. This can be used both to enhance our knowledge about language models, and about languages themselves. In this case, we show that mBERT implicitly learns facts about language groups during its simple pre-training tasks. We also conclude that the implication of Haider and Szucsich (2022) that German has a similar word order heritage as Slavic languages can be related to empirical data.

Limitations

The main limitation of this work is that it is concerned with a limited subset of languages. The only languages that have been investigated here are Slavic languages, and even then, some of them were omitted from experiments and results analysis – for example Slovak, Bulgarian or Ukrainian.

Another limitation explicitly stated in the work is the number of train sentences. In Subsection 3.1, we show that in order to draw meaningful conclusions, at least 5000 annotated sentences per language are needed. Coupled with the typical sizes of multilingual transfomer models, this leads to high computational and memory capacity being required to run experiments for multiple language groups.

References

Diego Alves, Marko Tadić, and Božo Bekavac. 2022. Multilingual comparative analysis of deep-learning dependency parsing results using parallel corpora. In *Proceedings of the BUCC Workshop within LREC 2022*, pages 33–42, Marseille, France. European Language Resources Association.

Ethan A. Chi, John Hewitt, and Christopher D. Manning. 2020. Finding universal grammatical relations in multilingual BERT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5564–5577, Online. Association for Computational Linguistics.

Slawomir Dadas, Michal Perelkiewicz, and Rafal Poswiata. 2020. Pre-training polish transformer-based language models at scale. *CoRR*, abs/2006.04229.

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.

Eric Fuß. 2022. Early german = slavic? *Theoretical Linguistics*, 48(1-2):57–71.

Hubert Haider and Luka Szucsich. 2022. Slavic languages – “svo” languages without svo qualities? *Theoretical Linguistics*, 48(1-2):1–39.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy. Association for Computational Linguistics.

Robert Mroczkowski, Piotr Rybak, Alina Wróblewska, and Ireneusz Gawlik. 2021. HerBERT: Efficiently pretrained transformer-based language model for Polish. In *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*, pages 1–10, Kiyv, Ukraine. Association for Computational Linguistics.

Jingcheng Niu, Wenjie Lu, and Gerald Penn. 2022. Does BERT rediscover a classical NLP pipeline? In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3143–3153, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Joakim Nivre, Daniel Zeman, Filip Ginter, and Francis Tyers. 2017. Universal Dependencies. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Tutorial Abstracts*, Valencia, Spain. Association for Computational Linguistics.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Piotr Rybak, Robert Mroczkowski, Janusz Tracz, and Ireneusz Gawlik. 2020. KLEJ: comprehensive benchmark for polish language understanding. *CoRR*, abs/2005.00630.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Zhiyong Wu, Yun Chen, Ben Kao, and Qun Liu. 2020. Perturbed masking: Parameter-free probing for analyzing and interpreting BERT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4166–4176, Online. Association for Computational Linguistics.

A Results for all dataset sizes

This appendix shows result tables, similar to Table 2, for decreasing dataset sizes. The tables feature additional train languages (rows) for which 10k sentences were not available, sorted by UUAS scores averaged across all test languages.

The division into Slavic and non-Slavic sections has been dropped in cases where the two groups are not separated – we can see that this is true for all sizes below 7.5k. We can also see that scores in general decrease as the dataset size decreases, which is visible especially when comparing Table 7 with other tables.

As concluded in Subsection 3.1 and Section 4, smaller dataset sizes seem to provide less meaningful results. There is however a visible tendency in Tables 6 and 7 for a single train language to dominate the scores for all Slavic test languages – this might be a reflection of quality (e.g. diversity or representativeness of average sentence structure) of the randomly sampled train subsets.

The fact that Chinese is the bottom language in Tables 5 and 6 is also noticeable, and might suggest an impact of a different writing systems on results. Unfortunately, the sample sizes are not enough to draw any conclusions.

	Baseline	Slovene	Russian	Polish	Czech	Belarusian	Croatian	Slovak	Ukrainian	Average
Slavic languages										
Russian	**48.61**	74.19	80.49	**78.27**	75.89	**76.37**	75.25	79.84	**78.55**	**77.36**
Slovene	47.42	80.08	75.06	77.54	**76.36**	74.07	**76.73**	81.01	76.20	77.13
Czech	48.24	**75.37**	75.47	77.72	79.69	74.06	74.87	**82.92**	76.64	77.09
Polish	52.30	74.39	**76.09**	82.11	75.97	75.30	74.33	79.90	76.76	76.86
Slovak	44.67	73.75	73.98	76.44	75.91	71.65	72.51	83.50	74.23	75.25
Belarusian	43.21	71.63	74.28	75.17	72.98	77.93	71.23	76.31	75.73	74.41
Non-Slavic languages										
German	47.87	**72.62**	**74.01**	**75.32**	**73.70**	**72.92**	**72.36**	**77.41**	**74.62**	**74.12**
French	49.60	69.51	71.51	74.34	70.22	70.54	71.12	73.18	72.49	71.61
English	48.14	69.15	72.43	72.98	70.59	69.81	71.53	73.87	72.41	71.60
Latvian	44.93	69.17	69.95	71.11	69.31	68.80	67.53	74.04	70.02	69.99
Spanish	**49.64**	68.25	69.72	72.96	68.66	68.50	69.30	71.13	70.35	69.86
Finnish	45.49	66.51	66.83	67.77	66.55	67.14	66.05	69.92	67.52	67.29

Table 3: Average UUAS scores for probes trained using 7.5k sentences. The test languages are in columns, and the train languages in rows. Higher values indicate better syntax recally. The results are averaged over three independent runs with different random seeds. Standard deviations of results are not reported, since values are below 1 UUAS point.

	Baseline	Slovene	Russian	Polish	Czech	Belarusian	Croatian	Slovak	Ukrainian	Average
Croatian	47.29	**75.71**	73.64	75.84	**75.30**	73.12	78.84	79.02	75.81	**75.91**
Ukrainian	46.14	72.90	**76.22**	**77.23**	74.06	**75.65**	73.89	78.37	78.68	75.88
Czech	47.31	74.60	73.79	76.14	78.25	72.45	73.82	**81.48**	74.71	75.65
Russian	47.23	72.69	79.09	76.98	74.23	74.35	72.69	78.13	**76.93**	75.64
Slovene	46.64	78.97	73.19	75.66	74.42	72.21	**74.50**	79.09	74.40	75.30
Polish	49.20	71.87	73.59	80.17	73.03	72.35	71.31	78.13	74.08	74.32
Slovak	44.05	72.20	72.51	74.81	74.45	71.38	70.91	82.39	72.98	73.95
German	46.82	71.92	73.09	74.41	72.37	71.68	70.87	76.66	73.21	73.03
Belarusian	41.32	69.03	71.73	72.76	70.62	75.88	69.67	74.66	73.56	72.24
French	**49.12**	67.46	70.55	73.03	68.91	69.14	70.05	71.97	71.29	70.30
English	47.58	66.54	69.90	70.82	68.38	67.37	68.63	71.37	70.07	69.14
Spanish	48.96	67.57	69.04	71.88	68.14	67.54	68.45	70.81	69.48	69.11
Latvian	43.52	66.76	66.94	68.94	66.99	67.05	65.03	70.99	68.26	67.62
Finnish	44.25	64.52	65.04	65.84	63.78	64.44	63.69	67.95	65.22	65.06

Table 4: Average UUAS scores for probes trained using 5k sentences. The test languages are in columns, and the train languages in rows. Higher values indicate better syntax recally. The results are averaged over three independent runs with different random seeds. Standard deviations of results are not reported, since values are below 1 UUAS point.

	Baseline	Slovene	Russian	Polish	Czech	Belarusian	Croatian	Slovak	Ukrainian	Average
Slovene	44.17	75.27	70.12	72.23	**71.36**	69.30	**70.82**	75.73	70.80	**71.95**
Czech	44.26	**71.81**	69.14	73.01	74.40	70.47	70.24	**76.66**	69.79	71.94
Croatian	44.43	71.33	69.80	72.31	71.24	69.13	74.41	75.04	71.38	71.83
Russian	44.97	68.66	75.30	**73.22**	69.98	**71.30**	68.06	73.40	**73.02**	71.62
Ukrainian	43.18	68.37	**72.11**	72.80	69.55	71.23	68.51	73.57	73.40	71.19
Slovak	41.51	68.39	68.85	71.15	70.90	68.71	67.50	78.71	68.86	70.38
German	42.45	67.01	68.83	70.82	68.67	67.86	66.83	72.91	69.01	68.99
Polish	43.66	65.63	67.75	74.18	67.33	67.28	65.64	70.45	68.03	68.29
French	47.11	65.67	67.21	70.77	66.70	66.93	67.44	69.93	68.24	67.86
Belarusian	38.58	64.68	67.03	69.01	66.81	71.27	64.61	70.37	67.78	67.70
English	46.32	65.47	67.59	67.95	65.65	65.41	66.51	69.20	67.61	66.92
Spanish	**47.52**	64.25	66.79	70.20	65.51	65.33	65.17	68.63	67.26	66.64
Latvian	39.38	61.39	62.08	63.54	61.35	63.03	59.50	66.51	61.87	62.41
Indonesian	46.29	59.10	61.05	64.36	60.26	62.94	60.07	63.79	63.21	61.85
Finnish	41.09	60.94	60.31	61.70	59.68	60.08	59.41	64.04	60.49	60.83
Chinese	42.12	54.52	56.12	55.68	55.89	58.69	54.42	58.25	57.96	56.44

Table 5: Average UUAS scores for probes trained using 2.5k sentences. The test languages are in columns, and the train languages in rows. Higher values indicate better syntax recally. The results are averaged over three independent runs with different random seeds. Standard deviations of results are not reported, since values are below 1 UUAS point.

	Baseline	Slovene	Russian	Polish	Czech	Belarusian	Croatian	Slovak	Ukrainian	Average
Croatian	39.33	**64.04**	**63.10**	**65.33**	**63.81**	**63.89**	67.59	66.99	62.84	**64.70**
Slovene	39.90	66.04	62.39	64.69	63.48	63.52	**62.17**	**67.53**	**63.96**	64.22
Czech	38.46	59.71	60.39	63.56	64.71	61.49	60.12	66.76	62.84	62.45
Slovak	37.13	59.54	60.48	64.46	63.14	62.45	58.02	69.25	62.16	62.44
French	44.07	57.97	61.72	64.24	60.41	62.09	60.71	64.58	63.13	61.86
Ukrainian	36.96	59.42	61.24	63.38	59.82	63.01	57.96	63.81	64.16	61.60
Spanish	**44.47**	57.91	60.61	63.64	59.91	60.58	58.88	62.42	61.38	60.67
English	42.55	57.12	61.74	62.92	58.57	61.54	57.57	62.82	61.93	60.53
Belarusian	34.80	57.21	58.46	60.61	58.80	63.32	55.61	62.50	61.07	59.70
Russian	37.71	56.84	62.07	60.86	57.79	60.48	54.82	62.04	61.17	59.51
German	34.16	58.12	58.09	60.08	58.54	59.31	56.69	61.44	59.71	59.00
Indonesian	42.60	54.93	57.59	60.51	56.42	59.11	53.83	60.07	59.24	57.71
Lithuanian	36.82	53.85	54.51	57.00	54.05	57.00	52.85	59.97	55.96	55.65
Latvian	35.44	54.00	54.17	56.10	54.28	55.39	51.15	59.48	54.37	54.87
Finnish	35.24	52.42	53.47	54.76	52.48	54.64	50.86	55.99	53.49	53.51
Polish	35.38	49.85	52.59	57.84	50.06	55.03	49.22	54.27	52.77	52.70
Chinese	37.82	50.69	52.65	53.10	51.21	53.86	50.34	54.19	52.59	52.33

Table 6: Average UUAS scores for probes trained using 1k sentences. The test languages are in columns, and the train languages in rows. Higher values indicate better syntax recally. The results are averaged over three independent runs with different random seeds. Standard deviations of results are not reported, since values are below 1 UUAS point.

	Baseline	Slovene	Russian	Polish	Czech	Belarusian	Croatian	Slovak	Ukrainian	Average
Spanish	**26.62**	**37.49**	**42.61**	**43.93**	**41.23**	**43.83**	**37.26**	**44.33**	**42.11**	**41.60**
French	23.12	35.94	40.93	41.96	38.84	42.43	36.95	42.45	40.96	40.06
Slovene	22.10	34.23	38.29	40.61	38.13	41.03	35.81	42.88	38.19	38.65
Czech	20.20	35.26	38.26	39.94	36.60	41.19	35.41	41.89	38.61	38.39
Indonesian	22.66	33.74	37.95	39.79	37.05	41.07	34.60	41.10	38.48	37.97
Croatian	20.02	34.94	37.48	39.04	37.18	40.60	33.59	41.11	37.49	37.68
Ukrainian	22.31	34.03	37.02	39.45	36.91	40.34	33.73	40.79	36.83	37.39
Lithuanian	21.08	33.89	36.59	38.79	35.82	40.28	33.45	40.83	36.96	37.08
English	22.68	32.87	37.92	39.17	35.67	39.71	33.48	39.49	37.31	36.95
Polish	20.97	32.49	37.35	38.60	35.97	40.17	32.99	40.32	37.18	36.88
Chinese	23.30	33.48	36.68	37.88	36.17	40.11	33.99	40.41	35.90	36.83
Belarusian	21.82	33.30	36.12	37.92	36.15	40.24	32.26	39.63	36.19	36.48
German	19.89	32.68	36.73	38.22	35.49	39.57	33.07	39.49	36.41	36.46
Slovak	18.01	34.21	36.28	37.90	35.03	39.37	33.34	38.89	35.41	36.30
Latvian	22.21	33.59	35.78	37.58	35.60	39.35	32.48	39.79	35.86	36.25
Russian	20.71	32.35	35.06	37.24	35.60	38.33	32.33	39.55	35.58	35.76
Finnish	20.28	31.32	34.74	36.05	33.71	37.49	31.37	37.95	34.17	34.60

Table 7: Average UUAS scores for probes trained using 100 sentences. The test languages are in columns, and the train languages in rows. Higher values indicate better syntax recally. The results are averaged over three independent runs with different random seeds. Standard deviations of results are not reported, since values are below 1 UUAS point.