Examining Cross-lingual Contextual Embeddings with Orthogonal Structural Probes

Tomasz Limisiewicz and **David Mareček**
Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics
Charles University, Prague, Czech Republic
{limisiewicz, marecek}@ufal.mff.cuni.cz

Abstract

State-of-the-art contextual embeddings are obtained from large language models available only for a few languages. For others, we need to learn representations using a multilingual model. There is an ongoing debate on whether multilingual embeddings can be aligned in a space shared across many languages. The novel *Orthogonal Structural Probe* (Limisiewicz and Mareček, 2021) allows us to answer this question for specific linguistic features and learn a projection based only on mono-lingual annotated datasets. We evaluate syntactic (UD) and lexical (WordNet) structural information encoded in mBERT's contextual representations for nine diverse languages.[1] We observe that for languages closely related to English, no transformation is needed. The evaluated information is encoded in a shared cross-lingual embedding space. For other languages, it is beneficial to apply orthogonal transformation learned separately for each language. We successfully apply our findings to zero-shot and few-shot cross-lingual parsing.

1 Introduction

The representation learned by language models has been successfully applied in various NLP tasks. Multilingual pre-training allows utilizing the representation for various languages, including low-resource ones. There is an open discussion about to what extent contextual embeddings are similar across languages (Søgaard et al., 2018; Hartmann et al., 2019; Vulić et al., 2020). The motivation for our work is to answer: **Q1** Is linguistic information uniformly encoded in the representations of various languages? And if this assumption does not hold: **Q2** Is it possible to learn orthogonal transformation to align the embeddings?

We probe for the syntactic and lexical structures encoded in multilingual embeddings with the new *Orthogonal Structural Probes* (Limisiewicz and Mareček, 2021). Previously, Chi et al. (2020) employed *structural probing* (Hewitt and Manning, 2019) to evaluate cross-lingual syntactic information in mBERT and visualize how it is distributed across languages. Our approach's advantage is learning an orthogonal transformation that maps the embeddings across languages based on monolingual linguistic information: dependency syntax and lexical hypernymy. This new capability allows us to test different probing scenarios. We measure how adding assumptions of isomorphism and uniformity of the representations across languages affect probing results to answer our research questions.

2 Related Work

Probing It is a method of evaluating linguistic information encoded in pre-trained NLP models. Usually, a simple classifier for the probing task is trained on the frozen model's representation (Linzen et al., 2016; Belinkov et al., 2017; Blevins et al., 2018). The work of Hewitt and Manning (2019) introduced structural probes that linearly transform contextual embeddings to approximate the topology of dependency trees. Limisiewicz and Mareček (2021) proposed new structural tasks and introduced orthogonal constraint allowing to decompose projected embeddings into parts correlated with specific linguistic features. Kulmizev et al. (2020) probed different languages to examine what type of syntactic dependency annotation is captured in an LM. Hall Maudslay et al. (2020) modify the loss function, improving syntactic probes' ability to parse.

Cross-lingual embeddings There is an essential branch of research studying relationships of embeddings across languages. Mikolov et al. (2013)

[1] English, Spanish, Slovene, Indonesian, Chinese, Finnish, Arabic, French, and Basque

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 4589–4598
November 7–11, 2021.

showed that distributions of the word vectors in different languages could be aligned in shared space. Following research analyzed various methods of aligning cross-lingual static embeddings (Faruqui and Dyer, 2014; Artetxe et al., 2016; Smith et al., 2017) and gradually dropped the requirement of parallel data for alignment (Artetxe et al., 2018; Zhang et al., 2017; Lample et al., 2018).

Significant attention was also devoted to the analysis of multilingual and contextual embeddings of MBERT (Pires et al., 2019; Libovický et al., 2020). There is also no conclusive answer to whether the alignment of such representations is beneficial to cross-lingual transfer. Wang et al. (2019) show that the alignment facilitates zero-shot parsing, while results of Wu and Dredze (2020) for multiple tasks put in doubt the benefits of the alignment.

3 Method

The *Structural Probe* (Hewitt and Manning, 2019) is a gradient optimized linear projection of the contextual word representations produced by a pre-trained neural model (e.g. BERT Devlin et al. (2019), ELMo Peters et al. (2018)).

In a *Distance Probe*, the Euclidean distance between projected word vectors approximates the distance between words in a dependency tree:

$$d_B(h_i, h_j)^2 = (B(h_i - h_j))^T(B(h_i - h_j)), \quad (1)$$

B is the *Linear Transformation* matrix and h_i, h_j are the vector representations of words at positions i and j.

Another type of a probe is a *Depth Probe*, where the token's depth in a dependency tree is approximated by the Euclidean norm of a projected word vector:

$$||h_i||_B^2 = (Bh_i)^T(Bh_i) \quad (2)$$

Orthogonal Structural Probes Limisiewicz and Mareček (2021) proposed decomposing matrix B and then gradient optimizing a vector and orthogonal matrix. The new formulation of an *Orthogonal Distance Probe* is[2]:

$$\begin{aligned} &d_{\bar{d}V^T}(h_i, h_j)^2 \\ &= (\bar{d} \odot V^T(h_i - h_j))^T(\bar{d} \odot V^T(h_i - h_j)), \end{aligned} \quad (3)$$

where V is an orthogonal matrix (*Orthogonal Transformation*) and $\bar{d}$ is a *Scaling Vector*, which can be changed during optimization for each task to allow multi-task joint probing.

This procedure allowed optimizing a separate *Scaling Vector* $\bar{d}$ for a specific objective, allowing probing for multiple linguistic tasks simultaneously. In this work, an individual *Orthogonal Transformation* V is trained for each language, facilitating multi-language probing. This approach assumes that the representations are isomorphic across languages; we examine this claim in our experiments.

Our implementation is available at GitHub: `https://github.com/Tom556/OrthogonalTransformerProbing`.

4 Experiments

We examine vector representations obtained from multilingual cased BERT (Devlin et al., 2019).

4.1 Data and Probing Objectives

We probe for syntactic structure annotated in Universal Dependencies treebanks (Nivre et al., 2020) and for lexical hypernymy trees from WordNet (Miller, 1995). We optimize depth and dependency probes in both types of structures jointly.

For both dependency and lexical probes, we use sentences from UD treebanks in nine languages. For each treebank, we sampled 4000 sentences to diminish the effect of varying size datasets in probe optimization. Lexical depths and distances for each sentence are obtained from hypernymy trees that are available for each language in Open Multilingual Wordnet (Bond and Foster, 2013).[3]

Choice of Layers We probe the representations of the 7th layer for dependency information and representations of the 5th layer for lexical information. These layers achieve the highest performance for the respective features.

4.2 Multilingual Evaluation

We utilize the new joint optimization capability of *Orthogonal Structural Probes* to analyze how the encoding of linguistic phenomena are expressed across different languages in MBERT representations.

To answer our research question, we evaluate three settings of multilingual *Orthogonal Structural Probe* training. The approaches are sorted by expressiveness; the most expressive one makes the

[2]Reformulation of an *Orthogonal Depth Probe* is analogical.

[3]List of all the datasets used in this work can be found in Appendix.

Approach	EN	ES	SL	ID	ZH	FI	AR	FR	EU	AVERAGE I-E	AVERAGE N-I-E	AVERAGE All
Dependency Distance Spearman's Correlation												
IN-LANG	.812	.858	.857	.841	.830	.788	.838	.856	.769	.846	.813	.828
Chi et al.	.817	.859	-	.807	.777	.812	.822	.864	-	.847	.805	.823
Δ MAPPEDL	.000	-.001	.001	-.003	.000	.001	-.001	-.002	.001	-.001	.000	.000
Δ ALLL	.000	-.007	-.006	-.013	-.039	.000	-.027	-.006	-.032	-.005	-.022	-.015
Chi et al.	-.011	-.011	-	-.018	-.060	-.010	-.037	-.011	-	-.011	-.031	-.023
Dependency Depth Spearman's Correlation												
IN-LANG	.843	.868	.867	.855	.844	.822	.865	.877	.797	.864	.837	.849
Δ MAPPEDL	-.004	-.003	-.002	-.002	.000	-.002	.001	-.002	-.001	-.002	-.001	-.002
Δ ALLL	-.006	-.007	-.008	-.011	-.035	-.005	-.031	-.010	-.031	-.008	-.023	-.016
Lexical Distance Spearman's Correlation												
IN-LANG	.756	.841	.639	.719	.800	.657	.733	.794	.679	.757	.717	.735
Δ MAPPEDL	-.003	.005	-.011	-.001	.010	.001	.042	.001	-.008	-.002	.009	.004
Δ ALLL	-.038	-.025	-.042	-.051	-.014	-.043	.025	-.013	-.063	-.030	-.029	-.030
Lexical Depth Spearman's Correlation												
IN-LANG	.853	.881	.779	.852	.875	.784	.906	.844	.842	.839	.850	.845
Δ MAPPEDL	.004	-.005	.013	-.011	.006	.023	-.024	.007	.021	.004	.005	.005
Δ ALLL	-.027	-.048	-.040	-.124	-.068	-.006	-.305	-.032	-.020	-.037	-.103	-.079

Table 1: Spearman's correlation between gold and predicted depths and distances. Δ denotes the differences from IN-LANG results. Each of our results is an average of 6 randomly initialized probing experiments. Statistically significant differences are circled. The three last columns present averages for **I**ndo-**E**uropean, **N**on-**I**ndo-**E**uropean, and all languages. The evaluation is not zero-shot, we use data in a target language. Correlations for dependency distance are compared with *Standard Structural Probes* reported by Chi et al. (2020).

weakest assumption about the likeness of representations across languages:

IN-LANG no assumption We train a separate instance of *Orthogonal Structural Probe* for each language. Neither *Scaling Vector* nor *Orthogonal Transformation* is shared between languages.

MAPPEDLANGS isomorphity assumption We train a shared *Scaling Vector* for each probing task and a separate *Orthogonal Transformation* per language. If the embedding subspaces are orthogonal across languages, the orthogonal mapping will be learned during probe training, and the setting will achieve similar results as the previous one.

ALLLANGS: uniformity assumption Both the *Scaling Vector* and *Orthogonal Transformation* are shared across languages. If the same embedding subspace encodes the probed information across languages, the results of this setting will be on par with the first approach.

The first and the last approach was proposed analyzed for *Structural Probes* by Chi et al. (2020). MAPPEDLANGS setting is possible thanks to the new probing formulation of Limisiewicz and Mareček (2021). For evaluation, we compute Spearman's correlations between predicted and gold depths and distances. In this evaluation, we use supervision for a target language. Furthermore, we analyze the impact of two language-specific features on the results: a) size of the MBERT training corpus in a given language; b) typological similarity to English. The former is expressed in the number of tokens in Wikipedia. The latter is a Hamming similarity between features in WALS (Dryer and Haspelmath, 2013).[4]

4.3 Zero- and Few-shot Parsing

We extract directed trees from the predictions of dependency probes. For that purpose, we employ the Maximum Spanning Tree algorithm on the predicted distances and the algorithm's extension of

[4]In this work, we consider all the features in the areas: Nominal Categories, Verb Categories, and Lexicon for computing a lexical typological similarity, and features in the areas: Nominal Syntax, Word Order, Simple Clauses, and Complex Sentences as a syntactic typological similarity. Each area contains multiple typological features.

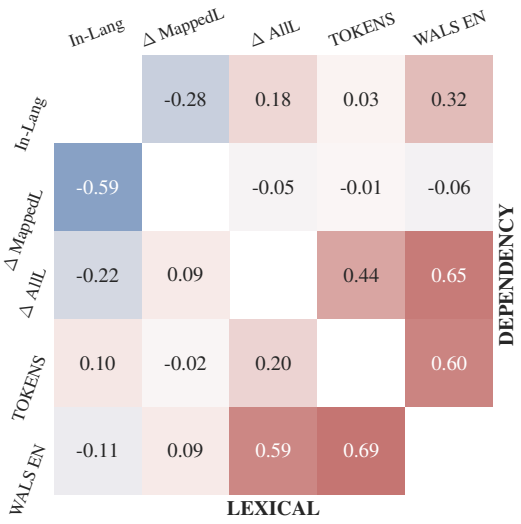


Figure 1: Pearson's correlation between results from Table 1 for each language and two language-specific features: typological similarity to English and number of tokens in Wikipedia. Correlations for dependency probes are in the upper-right triangle and for lexical probes in the lower-left triangle.

	N	ZH	EU	SL	FI	AR
Lauscher+*	0	51.41	50.31	-	65.66	44.46
Wang et al.		-	-	67.86	65.45	-
+CLBT**		-	-	69.04	67.96	-
+FT* **		-	-	69.16	**69.16**	-
MAPPEDL		34.44	39.10	35.44	37.33	40.95
ALLL		**52.92**	**58.77**	**70.76**	64.60	**57.47**
Lauscher+*	10	57.73	57.23	-	65.13	71.00
MAPPEDL		37.01	39.63	35.77	40.15	36.81
ALLL		53.12	58.51	70.85	64.98	68.59
Lauscher+*	50	66.78	66.73	-	69.26	75.84
MAPPEDL		45.07	50.02	55.09	49.32	57.77
ALLLANGS		53.63	59.07	70.43	65.02	68.81
Lauscher+*	100	69.91	65.70	-	70.25	78.50
MAPPEDL		50.27	56.07	60.00	52.86	62.36
ALLL		53.71	60.23	70.54	64.83	68.71
Lauscher+*	1000	80.12	74.75	-	78.00	83.85
MAPPEDL		60.57	65.98	72.81	63.80	68.85
ALLL		57.17	63.49	72.35	66.05	69.57

Table 2: UAS of extracted dependency trees. Our two approaches are compared to the previous works that use a biaffine parser (Lauscher et al., 2020; Wang et al., 2019). We probed the representations of the 7th layer. *): fine-tuning of MBERT is used. **): the multilingual dictionary is used to align the embeddings.

Kulmizev et al. (2020) to extract directed trees based on predicted depths.

We examine cross-lingual transfer for parsing sentences in Chinese, Basque, Slovene, Finnish, and Arabic. For each of them, we train the probe on the remaining eight languages. In a few-shot setting, we also optimize on 10 to 1000 examples from the target language.

5 Results

Sperman's correlation Using IN-LANG probes for each language gives high Spearman's correlations across the languages. The MAPPEDLANGS approach brings only a slight difference for most of the configuration while imposing uniformity constraint (ALLLANGS) deteriorates the results for some of the languages, as shown in Table 1. The drop in correlation is especially high for Non-Indo-European languages (except for lexical distance where the difference between Indo-European and Non-Indo-European groups is small).

In Fig. 1, we present the Pearson's correlations between results from Table 1 and two language-specific features. The key observation is that topological similarity to English is strongly correlated with ΔALLLANGS. Hence, a shared probe achieves relatively good for English, Spanish, and French. It shows that lexical and dependency information is uniformly distributed in the embedding space for those languages. We bear in mind that the European languages are over-represented in the MBERT's pre-training corpus. However, the size of pre-training corpora is correlated to a lesser extent with ΔALLLANGS than WALS similarity, suggesting that the latter has a more prominent role than the former. There is no significant correlation between ΔMAPPEDLANGS and typological similarity; the embeddings of diverse languages can be similarly well mapped into a shared space. Notably, we observe that some languages with the lower performance of IN-LANG probes can benefit from mapping (e.g., Slovene, Finnish, and Basque in the lexical depth). We view it as a benefit of cross-lingual transfer from more resourceful languages.

Zero-shot Parsing For all languages except Finnish in zero-shot configuration, our ALLLANGS approach is better than other works that utilize a biaffine parser (Dozat and Manning, 2017) on top of MBERT representations, shown in Table 2. Without any supervision, our MAPPEDLANGS approach performs poorly because mapping cannot be learned effectively. When some annotated data is added to the training, the difference between ALLLANGS and MAPPEDLANNGS decreases. We

observe that between 100 and 1000 training samples are needed to learn the *Orthogonal Transformation* effectively. Also, with higher supervision, we observe that the results reported by (Lauscher et al., 2020) notably outperform our approach. The outcome was anticipated because they fine-tune MBERT and use biaffine with a larger capacity than a probe. For their approach, the introduction of even small supervision is more advantageous than for probing.

6 Conclusions

We propose an effective way to multilingually probe for syntactic dependency (UD) and lexical hypernymy (WordNet). Our algorithm learns probes for multiple tasks and multiple languages jointly. The formulation of *Orthogonal Structural Probe* allows learning cross-lingual transformation based on mono-lingual supervision. Our comparative evaluation indicates that the evaluated information is similarly distributed in the MBERT's representations for languages typologically similar to English: Spanish, French, and Finnish. We show that aligning the embeddings with *Orthogonal Transformation* improves the results for other examined languages, suggesting that the representations are isomorphic. We show that the probe can be utilized in zero- and few-shot parsing. The method achieves better UAS results for Chinese, Slovene, Basque, and Arabic in a zero-shot setting than previous approaches, which use a more complex biaffine parser.

Limitations In our choice of languages, we wanted to ensure diversity. Nevertheless, four of the analyzed languages belong to an Indo-European family that could facilitate finding shared encoding subspace for those languages.

Acknowledgments

We thank anonymous EMNLP reviewers for their valuable comments and suggestions for improvement. This work has been supported by grant 338521 of the Charles University Grant Agency and by Progress Q48 grant of Charles University. We have been using language resources and tools developed, stored, and distributed by the LINDAT/CLARIAH-CZ project of the Ministry of Education, Youth and Sports of the Czech Republic (project LM2018101).

References

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2016. Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2289–2294, Austin, Texas. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 789–798, Melbourne, Australia. 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.

Francis Bond and Ryan Foster. 2013. Linking and extending an open multilingual Wordnet. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1352–1362, Sofia, Bulgaria. Association for Computational Linguistics.

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.

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.

Kaja Dobrovoljc, Tomaž Erjavec, and Simon Krek. 2017. The Universal Dependencies treebank for Slovenian. In *Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing*, pages 33–38, Valencia, Spain. Association for Computational Linguistics.

Timothy Dozat and Christopher D. Manning. 2017. Deep biaffine attention for neural dependency parsing. *ArXiv*, abs/1611.01734.

Matthew S. Dryer and Martin Haspelmath, editors. 2013. *WALS Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Sabri Elkateb, William Black, Horacio Rodriguez, Musa Alkhalifa, Piek Vossen, Adam Pease, and Christiane Fellbaum. 2006. Building a wordnet for Arabic. In *In Proceedings of The fifth international conference on Language Resources and Evaluation (LREC 2006)*.

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

Darja Fišer, Jernej Novak, and Tomaž. 2012. sloWNet 3.0: development, extension and cleaning. In *Proceedings of 6th International Global Wordnet Conference (GWC 2012)*, pages 113–117. The Global WordNet Association.

Aitor Gonzalez-Agirre, Egoitz Laparra, and German Rigau. 2012. Multilingual central repository version 3.0: upgrading a very large lexical knowledge base. In *Proceedings of the 6th Global WordNet Conference (GWC 2012)*, Matsue.

Rowan Hall Maudslay, Josef Valvoda, Tiago Pimentel, Adina Williams, and Ryan Cotterell. 2020. A tale of a probe and a parser. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7389–7395, Online. Association for Computational Linguistics.

Mareike Hartmann, Yova Kementchedjhieva, and Anders Sø gaard. 2019. Comparing unsupervised word translation methods step by step. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Katri Haverinen, Jenna Nyblom, Timo Viljanen, Veronika Laippala, Samuel Kohonen, Anna Missila, Stina Ojala, Tapio Salakoski, and Filip Ginter. 2014. Building the essential resources for finnish: the turku dependency treebank. *Lang. Resour. Evaluation*, 48(3):493–531.

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.

Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *International Conference on Learning Representations*.

Artur Kulmizev, Vinit Ravishankar, Mostafa Abdou, and Joakim Nivre. 2020. Do neural language models show preferences for syntactic formalisms? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4077–4091, Online. Association for Computational Linguistics.

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

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From zero to hero: On the limitations of zero-shot language transfer with multilingual Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4483–4499, Online. Association for Computational Linguistics.

Jindřich Libovický, Rudolf Rosa, and Alexander Fraser. 2020. On the language neutrality of pre-trained multilingual representations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1663–1674, Online. Association for Computational Linguistics.

Tomasz Limisiewicz and David Mareček. 2021. Introducing orthogonal constraint in structural probes. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 428–442, Online. Association for Computational Linguistics.

Krister Lindén and Lauri Carlson. 2010. Finnwordnet — wordnet påfinska via översättning. *LexicoNordica — Nordic Journal of Lexicography*, 17:119–140. In Swedish with an English abstract.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

Aranzabe M., Atutxa A., Bengoetxea K., Díaz de Ilarraza A., Goenaga I., Gojenola K., and Uria L. 2015. Automatic conversion of the basque dependency treebank to universal dependencies. *14th International Workshop on Treebanks and Linguistic Theories*.

Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, Claudia Bedini, Núria Bertomeu Castelló, and Jungmee Lee. 2013. Universal Dependency annotation for multilingual parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 92–97, Sofia, Bulgaria. Association for Computational Linguistics.

Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. 2013. Exploiting similarities among languages for machine translation. *CoRR*, abs/1309.4168.

George A. Miller. 1995. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41.

Nurril Hirfana Mohamed Noor, Suerya Sapuan, and Francis Bond. 2011. Creating the open Wordnet Bahasa. In *Proceedings of the 25th Pacific Asia Conference on Language, Information and Computation (PACLIC 25)*, pages 258–267, Singapore.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. 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)*, New Orleans, Louisiana. Association for Computational Linguistics.

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

Elisabete Pociello, Eneko Agirre, and Izaskun Aldezabal. 2011. Methodology and construction of the Basque wordnet. *Language Resources and Evaluation*, 45(2):121–142.

Benoît Sagot and Darja Fišer. 2008. Building a free French wordnet from multilingual resources. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco.

Natalia Silveira, Timothy Dozat, Marie-Catherine de Marneffe, Samuel Bowman, Miriam Connor, John Bauer, and Christopher D. Manning. 2014. A gold standard dependency corpus for English. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*.

Samuel L. Smith, David H. P. Turban, Steven Hamblin, and Nils Y. Hammerla. 2017. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*.

Otakar Smrž, Viktor Bielický, Iveta Kouřilová, and Jakub Kráčmar Zemánek. 2008. Dependency treebank : A word on the million words.

Anders Søgaard, Sebastian Ruder, and Ivan Vulić. 2018. On the limitations of unsupervised bilingual dictionary induction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 778–788, Melbourne, Australia. Association for Computational Linguistics.

Mariona Taulé, Maria Antònia Martí, and Marta Recasens. 2008. Ancora: Multilevel annotated corpora for catalan and spanish. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2008, 26 May - 1 June 2008, Marrakech, Morocco*. European Language Resources Association.

Ivan Vulić, Sebastian Ruder, and Anders Søgaard. 2020. Are all good word vector spaces isomorphic? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3178–3192, Online. Association for Computational Linguistics.

Shan Wang and Francis Bond. 2013. Building the chinese open wordnet (cow): Starting from core synsets. In *Sixth International Joint Conference on Natural Language Processing*, pages 10–18.

Yuxuan Wang, Wanxiang Che, Jiang Guo, Yijia Liu, and Ting Liu. 2019. Cross-lingual BERT transformation for zero-shot dependency parsing. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5721–5727, Hong Kong, China. Association for Computational Linguistics.

Shijie Wu and Mark Dredze. 2020. Do explicit alignments robustly improve multilingual encoders? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4471–4482, Online. Association for Computational Linguistics.

Meng Zhang, Yang Liu, Huanbo Luan, and Maosong Sun. 2017. Earth mover's distance minimization for unsupervised bilingual lexicon induction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1934–1945, Copenhagen, Denmark. Association for Computational Linguistics.

A WALS similarities

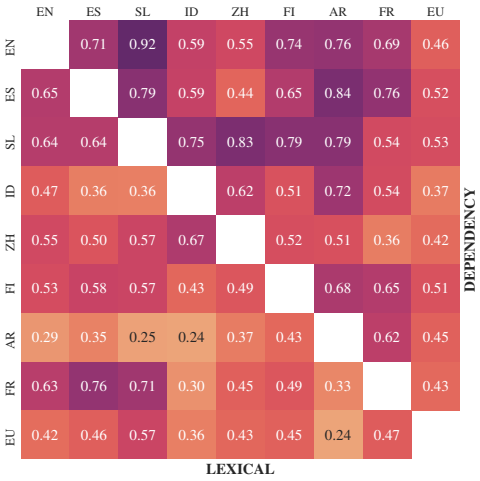


Figure 2: Typological (WALS) similarities between languages. Dependency similarities in the upper-right triangle and lexical similarities in the lower-left triangle.

In Fig. 2, we present typological similarities between languages. Bases on Fig. 3 we observe that typological similarity to languages related to English: Spanish, Finnish, French is correlated to ΔALLLANGS. Moreover, the correlation between similarity to these languages and the number of tokens in Wikipedia is smaller than for English[5]. It supports our claim that typological similarity is more important for uniformity assumption than the size of the pre-training corpus.

B Pre-training corpus size

Sizes of Wikipedia in eight analyzed languages are presented in Table 3.

C Datasets

In Table 4 we aggregate all the datasets used in our experiments.

D Information separation

In line with the findings of Limisiewicz and Mareček (2021) we have observed that in multilingual setting *Orthogonal Structural Probes* disentangle the subspaces responsible for encoding lexical and dependency structures Table 5.

[5]English is especially over-represented in the pre-trained corpus

Language	Articles	Tokens
English	6,171,405	2,622,505,044
French	2,255,469	823,362,731
Spanish	1,631,829	688,970,215
Chinese	1,151,113	269,492,468
Arabic	1,069,379	169,126,089
Finnish	494,487	98,712,322
Indonesian	547,825	96,356,452
Basque	365,301	46,487,007
Slovene	169,604	42,511,205

Table 3: The number of articles and tokens in Wikipedia for analyzed languages. The data come from https://github.com/mayhewsw/multilingual-data-stats/tree/main/wiki

E Probing setup

We use the same setup for training the *Orthogonal Structural Probe* as Limisiewicz and Mareček (2021), i.e. Adam optimizer (Kingma and Ba, 2014), initial training rate 0.02, and learning rate decay. We use *Double Soft Orthogonality Regularization* to coerce orthogonality of the matrix V.

E.1 Number of Parameters

A *Scaling Vector* for each of 4 objectives has a size 768×1 and an *Orthogonal Transformation* for each language is a matrix of size 768×768. In MAPPEDLANGS, our largest memory-wise setting, we train 8 *Orthogonal Transformations*. In this configuration, our probe has $4,721,664$ parameters.

E.2 Computation Time

We optimized probes on a GPU core *GeForce GTX 1080 Ti*. Training a probe in MAPPEDLANGS configuration takes about 3 hours.

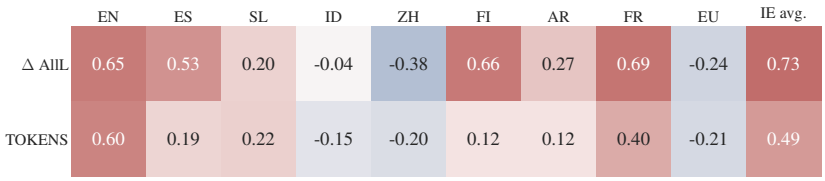
F Supplementary Results

F.1 UUAS results

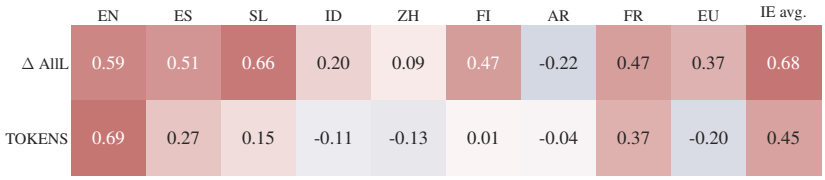
The Table 6 contains the results for undirected dependency trees. We use the same probing setting as in Section 3.2 without assigning directions to the edges. Similarly to Chi et al. (2020), we exclude punctuation from the evaluation.

F.2 Validation Results

In Table 7, we present the validation results corresponding to the test results in Table 1 of the main paper.



(a) Dependency



(b) Lexical

Figure 3: Pearson's correlation between WALS similarity to a specific language and ΔALLLANGS, the number of tokens in Wikipedia. "IE avg." stands for average similarity to analyzed Indo-European languages, i.e., English, Spanish, French, Slovene.

Language	Dependency Name	Dependency Reference	Lexical Name	Lexical Reference
English	EWT	Silveira et al. (2014)	Princeton Wordnet	Miller (1995)
French	GSD	McDonald et al. (2013)	Wordnet Libre du Français	Sagot and Fišer (2008)
Spanish	Ancora	Taulé et al. (2008)	Multilingual Central Repository	Gonzalez-Agirre et al. (2012)
Chinese	GSD	McDonald et al. (2013)	Chinese Open Wordnet	Wang and Bond (2013)
Arabic	PADT	Smrž et al. (2008)	Arabic WordNet	Elkateb et al. (2006)
Finnish	TDT	Haverinen et al. (2014)	FinnWordNet	Lindén and Carlson. (2010)
Indonesian	GSD	McDonald et al. (2013)	Wordnet Bahasa	Mohamed Noor et al. (2011)
Basque	BDT	M. et al. (2015)	Multilingual Central Repository	Pociello et al. (2011)
Slovene	SSJ	Dobrovoljc et al. (2017)	sloWNet	Fišer et al. (2012)

Table 4: The datasets used for training dependency and lexical probes.

		DEP		LEX	
		Depth	Dist.	Depth	Dist.
DEP	Depth	98	65	1	0
	Dist.		142	0	0
LEX	Depth			22	13
	Dist.				58

Table 5: The number of shared dimensions selected by *Scaling Vector* after the joint training of probe in MAPPEDLANGS setting on top of the 7th layer.

	N	ZH	EU	SL	FI	AR
Chi et al.	0	51.30	-	-	70.70	70.40
MAPPEDL		39.99	46.96	41.58	43.91	40.95
ALLL		57.82	64.59	75.06	68.70	68.70
MAPPEDL	10	42.37	47.06	41.07	46.38	36.81
ALLL		58.06	64.65	75.30	69.06	68.59
MAPPEDL	50	51.64	56.67	59.34	53.53	57.77
ALLL		58.73	65.18	74.99	69.08	68.81
MAPPEDL	100	62.36	62.44	64.51	57.95	62.36
ALLL		68.71	66.00	75.16	68.97	68.71
MAPPEDL	1000	66.43	70.50	76.10	67.08	68.85
ALLL		62.36	68.60	76.79	69.73	69.57

Table 6: UUAS of extracted dependency trees in zero- and few-shot setting. The result of *Structural Probe* reported by Chi et al. (2020) for reference.

Approach	EN	ES	SL	ID	ZH	FI	AR	FR	EU
Dependency Distance Spearman's Correlation									
IN-LANG	.816	.861	.844	.822	.815	.803	.835	.872	.749
Δ MAPPEDLANGS	.000	-.002	.000	.001	-.001	-.001	-.002	-.002	.002
Δ ALLLANGS	-.001	-.007	-.004	-.011	-.041	.000	-.022	-.010	-.021
Dependency Depth Spearman's Correlation									
IN-LANG	.847	.868	.857	.853	.837	.807	.864	.893	.786
Δ MAPPEDLANGS	-.003	-.002	-.004	.000	.002	-.005	-.002	-.003	-.001
Δ ALLLANGS	-.004	-.005	-.004	-.013	-.034	-.004	-.027	-.007	-.033
Lexical Distance Spearman's Correlation									
IN-LANG	.898	.880	.867	.857	.777	.664	.726	.810	.714
Δ MAPPEDLANGS	.000	.001	-.001	.003	.001	.001	.027	.008	-.005
Δ ALLLANGS	-.005	-.005	-.017	-.009	-.001	-.053	.004	-.024	-.082
Lexical Depth Spearman's Correlation									
IN-LANG	.844	.882	.792	.869	.862	.784	.884	.879	.847
Δ MAPPEDLANGS	.010	.002	.009	-.013	.006	.020	.011	-.006	.012
Δ ALLLANGS	-.010	-.067	-.079	-.108	-.055	.000	-.259	-.043	-.034

Table 7: Validation Spearman's correlation between gold and predicted depths and distances. We probe the representations of 7th layer for dependency information and representations of 5th layer for lexical information.