Word-order typology in Multilingual BERT:
A case study in subordinate-clause detection

Dmitry Nikolaev   **Sebastian Padó**

IMS, University of Stuttgart

dnikolaev@fastmail.com, pado@ims.uni-stuttgart.de

Abstract

The capabilities and limitations of BERT and similar models are still unclear when it comes to learning syntactic abstractions, in particular across languages. In this paper, we use the task of *subordinate-clause detection* within and across languages to probe these properties. We show that this task is deceptively simple, with easy gains offset by a long tail of harder cases, and that BERT's zero-shot performance is dominated by word-order effects, mirroring the SVO/VSO/SOV typology.

1 Introduction

Analysing the ability of pre-trained neural language models, such as BERT (Devlin et al., 2019), to abstract grammatical patterns from raw texts has become a prominent research question (Jawahar et al., 2019; Rogers et al., 2020). Results remain mixed. While BERT-based models have been shown to learn syntactic representations that are similarly structured across languages (Chi et al., 2020), some grammatical patterns, such as discontinuous constituents, remain challenging for them even when training data is plentiful (Kogkalidis and Winholds, 2022). In practical terms, zero-shot performance of BERT-based models is lower for typologically distant languages (Pires et al., 2019), and they can profit from direct exposure to typological features during fine-tuning (Bjerva and Augenstein, 2021).

In this study, we add another datapoint to the conversation by analysing the ability of BERT-based models to capture the *distinction between main and subordinate clauses* across languages. This task is promising for two reasons. First, it highlights variability in the way main and subordinate clauses are structured across languages, thus acting as an informative probe into the relationship between BERT and typological categories. Second, the task is arguably relevant for downstream performance on natural-language understanding,

where (some notion of) syntactic scope and compositionality should support tasks such as analysing commitment (Jiang and de Marneffe, 2019; Zhang and de Marneffe, 2021) or factuality (Lotan et al., 2013), text simplification (Sikka and Mago, 2020), or paraphrase detection (Timmer et al., 2021). In order to operationalise it in a cross-lingual fashion, we use the Universal Dependencies framework (UD; Nivre et al., 2020) with its large multilingual collection of corpora.

Our analysis proceeds in two stages. First, we survey the performance of BERT models fine-tuned and tested on the same language across 20 typologically diverse languages (§ 3). For the majority of languages, distinguishing main and subordinate clauses is easily solved with base-size models and relatively small training sets. However, some languages demonstrate a non-negligible number of errors, which we analyse.

Then we study the performance of Multilingual BERT (mBERT) in a zero-shot setting (§ 4), where we fine-tune the model on labeled data in 10 different languages and then test its performance on 31 datasets representing 27 different languages. We find that the performance of mBERT is dominated by word-order effects well known from the typological literature (Comrie, 1981): the Arabic model shows best-in-class performance on Irish, and the Japanese model has best-in-class performance on Korean, while both have poor performance overall. European languages with large training sets provide good inductive bias for typologically diverse languages but fail on SOV languages.

2 Experimental Setup

Data   To make our analysis maximally comparable across languages, we start from the Parallel Universal Dependencies (PUD) collection (Zeman et al., 2017), which contains translations for a set of 1000 English sentences. PUD only contains test corpora. As these are too small to be further

Language	Mandarin	Vietnamese	Korean	Arabic	Hindi	German	Armenian	Turkish	Welsh	Indonesian
Accuracy	88.7	90	90.4	91.2	93.6	94.1	94.3	95.1	95.6	96

Language	Basque	Spanish	Irish	English	Hebrew	Afrikaans	French	Japanese	Czech	Russian
Accuracy	96.9	97.1	97.4	97.9	98.2	98.8	99	99.1	99.6	99.7

Table 1: Performance of single-language models.

split into train/test subsets, we use other corpora to fine-tune the models. We also add corpora for languages not covered by PUD for better typological coverage. See Appendix § A.2 for the full list.

Model The experimental setup is identical in the single-language and zero-shot settings. A pre-trained mBERT model (a variant of `bert-base`) and several pre-trained single-language BERT models, all provided by HuggingFace (Wolf et al., 2020), are fine-tuned on the binary classification of predicates into main vs. subordinate clauses. We operationalize main clauses as those headed by predicates with the UD label `root` and subordinate clauses through the UD labels `acl`, `ccomp`, `advcl`, `csubj`, and `xcomp`. The last hidden state of the embedding model for the first subword of each predicate is fed to a two-layer MLP with a tanh activation after the first layer, and the model is fine-tuned using cross-entropy loss. For the single-language setup, the model is fine-tuned for five epochs, and we report the best result on the validation set. Most models begin overfitting after the second epoch, so in the zero-shot setting all models are fine-tuned for two epochs.

3 Single-Language Models

The main results obtained by the models fine-tuned and tested on the same language are shown in Table 1. Results are above 90% for almost all languages, while a majority baseline (always assign subordinate clause) attains an accuracy of 50–70% depending on the language. (Table 3 in the Appendix provides more details about the models and corpora, including exact baseline results.) At first glance, neither the size of the training set nor the size of the model seem to be a major factor: mBERT demonstrates better performance when fine-tuned on the small Afrikaans and Hebrew datasets than when trained on a bigger Chinese dataset. When fine-tuned on the English data, it attains the same performance as an English-only `bert-large`.[1]

[1]The mBERT result is reported in Table 2.

A more fundamental distinction seems to exist between major European languages, the results on which are generally at > 97% accuracy (except for German), and Mandarin Chinese, Vietnamese, and Korean where results are around 90%. Our analysis indicates that these differences are partly due to discrepancies in UD annotations across corpora but also due to genuine syntactic differences. An example of an annotation-related confound is the treatment of quotations. The PUD corpora that we use preferentially as test sets treat quotations as sentential complements of communication verbs. Some of the corpora we use for fine-tuning, however, analyse the cases where quotation precedes the verb of speech as parataxis. The head predicate of the quotation therefore receives the label `root` and becomes the main predicate of the whole sentence, leading to spurious mistakes in the analysis of PUD corpora, where they are annotated as `ccomp`'s. This discrepancy accounts for the lion's share of classification mistakes in German and some mistakes in Mandarin.

In contrast, an example of genuine ambiguity is provided by the Mandarin *gēnjù* construction. This construction means 'according to' and can incorporate both nominal and verbal constituents. Thus, *gēnjù shàng biǎogé zhōng qī gè yuánsù de guānxì* from the Mandarin GSD corpus, which we used for fine-tuning, means 'based on the relationship of the seven elements in the above table', and the annotation treats this construction as an oblique prepositional phrase. Cf. the following example from the Mandarin PUD corpus: *gēnjù kěxíng xìng yánjiū gūjì* 'according to the feasibility study / the feasibility study estimates that / as the feasbility study estimates'. The analysis of this sentence in PUD makes *gūjì* 'estimate' the main predicate of the sentence, while an alternative analysis would make it the head of an adverbial clause, and yet another analysis would label it as a nominal element. The ability of Mandarin words to act as different parts of speech in different contexts (especially in case of verbs, which can act as clause heads, auxiliaries, complementisers, and compound elements)

12

makes this kind of disambiguation difficult even for human annotators, which in turn makes it hard to formulate the exact rule that language models are supposed to extract from the data. A similar situation holds for Vietnamese.[2]

A different type of systematic ambiguity is presented by Korean, which also demonstrates poorer performance. Korean has about sixty markers connecting two clauses, and many of those allow for both coordinative and conjunctive readings, which makes either the first or the second clause the main one, respectively (Cho and Whitman, 2020, 220–227). Examples of this type are responsible for a large share of mistakes in Korean.

Overall, these results indicate that subordinate-clause detection is a long-tail task: major easily learnable patterns account for more than 90% of test cases for all languages, but in some languages there is an assortment of harder cases that prevent language models from efficiently generalising.

4 Zero-Shot Setting

4.1 Quantitative Results

We now turn to the analysis of the performance of the models in the zero-shot setting. The model described in § 2 is fine-tuned for two epochs on five European languages (English, Russian, Czech, French, and German) and five Eurasian languages (Standard Arabic, Mandarin Chinese, Turkish, Korean, and Japanese) with larger training corpora (the ones shown in Table 3). Each of the fine-tuned models is then applied in a zero-shot way to a range of test corpora from the UD collection.[3]

Based on the results in Table 2, several observations can be made. First, there is a set of European languages with large training corpora that can act as 'general approximators': they demonstrate high performance across the board. The best overall performance is attained by Russian, which has the second-largest training corpus (nearly 33k sentences). German, with the largest training corpus (nearly 56k sentences) performs worse than both Russian and English (the second best, with only

circa 6k training sentences). While this good result for English may be attributed to more informative pre-training (English Wikipedia is much larger than the German one), such a bias would also have favoured German compared to Russian. An alternative explanation is provided by the more idiosyncratic German word-order patterns (V2 in main clauses vs. V-last in subordinate clauses), which help it achieve best-in-class performance on the similar Afrikaans. Notably, Russian beats English even though PUD corpora were translated from English and therefore should contain some traces of its morphosyntactic patterns (Rabinovich et al., 2017; Nikolaev et al., 2020).

At the other end of the spectrum, we find mediocre general approximators (Arabic, Turkish) and outright bad ones (Japanese and Korean). At first glance, their performance could be an artefact of lower-quality annotations or suboptimal tokenisation (Mielke et al., 2021). This, however, does not explain a remarkable set of results that is clearly due to word-order patterns. While the fine-tuned model for Arabic, a VSO language, performs worse on its own test corpus than models fine-tuned on European languages, it provides best-in-class performance on Irish, another VSO language (96% accuracy). The English-based model is not far behind (95%), but given the overall large gap in performance between them across the board, it seems that congruent word-order patterns provide a strong inductive bias for subordinate-clause identification.

Unfortunately, VSO languages are rare,[4] and it is impossible to check if this pattern generalises to other language pairs. However, our test-corpus suite includes data on strict SOV languages (Japanese, Korean) and languages where SOV is the dominant (Hindi, Turkish) or a common (Mandarin, Basque) pattern. These provide us with a large number of language pairs with different degrees of word-order congruence and fairly clear patterns of model performance. First, universal approximators, despite good performance on VSO languages, struggle on strict SOV languages, especially Japanese, while SOV languages demonstrate consistently good performance among themselves. E.g., Korean demonstrates best-in-class performance on Turkish, tied with Turkish itself, while Japanese has best-in-class performance on Korean. Turkish also demonstrates decent perfor-

---

[2] 'Syntactic category classification for Vietnamese is still in debate. That lack of consensus is due to the unclear limit between the grammatical roles of many words as well as the frequent phenomenon of syntactic category mutation' (Nguyen et al., 2004).

[3] Where available, we experiment with two test sets for the same language to assess domain-induced variance. As Table 2 shows, the difference in scores between different testing corpora for the same language can reach 5–6%, but it does not change the overall pattern.

[4] Out of 1376 languages in WALS (Dryer and Haspelmath, 2013), 95 are VSO, 564 are SOV and 488 are SVO.

	ar padt	ga idt	af booms	de pud	cs pud	cy ccg	en ewt	en pud	es pud	fi pud	fr pud	he hdt	hy arm	id pud	is pud	it pud
English	**96**	95	93	94	94	86	**98**	**98**	**96**	96	96	93	93	**95**	**96**	**98**
Russian	95	94	86	93	95	90	94	96	95	**97**	**99**	94	**94**	93	**96**	96
Czech	94	94	83	**95**	**100**	**92**	92	93	94	94	98	**95**	88	89	92	94
French	94	91	84	92	95	90	92	97	95	95	**99**	94	87	93	**96**	96
German	94	87	**95**	94	88	82	90	95	94	92	93	91	90	89	95	96
Arabic	90	**96**	76	85	84	90	85	87	86	85	84	87	84	85	89	85
Mandarin	86	84	85	87	87	85	85	86	86	89	87	87	81	83	86	87
Turkish	67	61	61	65	71	<u>62</u>	64	69	75	77	71	73	73	68	68	74
Korean	<u>51</u>	<u>53</u>	63	<u>53</u>	61	<u>52</u>	<u>51</u>	<u>54</u>	<u>59</u>	65	<u>59</u>	59	<u>57</u>	<u>55</u>	<u>54</u>	61
Japanese	<u>55</u>	<u>56</u>	<u>41</u>	<u>39</u>	<u>52</u>	63	<u>51</u>	<u>52</u>	<u>52</u>	<u>54</u>	<u>55</u>	<u>54</u>	<u>58</u>	<u>53</u>	<u>51</u>	<u>54</u>

	pl pud	pt pud	ru pud	ru syntag	sv pud	eu bdt	hi pud	tr pud	ja gsd	ja pud	ko pud	vi vtb	th pud	zh gsd	zh pud	mean
English	95	**97**	93	93	96	**88**	87	83	<u>66</u>	<u>70</u>	67	**82**	84	<u>67</u>	71	88.9
Russian	98	95	**100**	99	95	**88**	90	90	<u>68</u>	72	70	79	80	<u>64</u>	69	**89.2**
Czech	97	93	94	96	92	87	88	88	<u>64</u>	<u>66</u>	68	78	79	<u>65</u>	71	87.5
French	**98**	96	96	97	94	85	**89**	86	<u>54</u>	<u>61</u>	66	77	76	<u>63</u>	69	87.0
German	89	94	93	88	**97**	81	86	78	<u>59</u>	62	<u>57</u>	78	78	<u>67</u>	68	85.2
Arabic	85	86	88	85	89	71	70	65	<u>63</u>	66	<u>59</u>	74	79	<u>66</u>	<u>65</u>	80.3
Mandarin	85	85	86	85	86	82	87	89	80	78	77	74	80	**91**	**86**	84.6
Turkish	76	73	71	71	69	79	83	**94**	82	83	88	<u>63</u>	<u>68</u>	72	71	72.3
Korean	66	<u>58</u>	<u>59</u>	<u>59</u>	<u>53</u>	74	76	**94**	87	88	88	<u>52</u>	61	<u>67</u>	<u>66</u>	63.1
Japanese	<u>54</u>	<u>52</u>	<u>55</u>	<u>55</u>	<u>50</u>	<u>57</u>	63	88	**99**	**98**	**95**	54	<u>70</u>	72	66	60.3

Table 2: Performance of zero-shot models. Rows: source languages; columns: target languages and corpora. Underlined values fail to beat the majority-class baseline (always predict subordinate clause). See § A.1 for language abbreviations and § A.2 for details about corpora.

mance on Hindi, with which it shares a relatively flexible SOV order.

Another language with strong SOV tendencies is Mandarin Chinese, which has been argued to be in transition from SVO to SOV order (Sun and Givón, 1985). Mandarin, which we already found difficult to model in § 3, is very hard to generalise to, with no source languages attaining accuracy above 71–72%. Tellingly, Turkish is the only other language with decent results on both Mandarin test sets. Mandarin is also the only language to always beat the majority-class baseline.

4.2 Case Study: English–Mandarin

In order to get a better understanding of the difficulties that models face in the zero-shot setting we analysed the mistakes that the English-based fine-tuned model made when making predictions on Mandarin data.

Setting aside errors stemming from annotation discrepancies,[5] the major source of model mistakes seems to be the fact that Mandarin complex sentences are predominantly right-headed: 99% of

advcl, 100% of acl, and 96% of dep[6] have their parent node to the right. In contrast, 75% of English advcl and 98% of English acl are left-headed in PUD. This makes an English-based zero-shot model prejudiced against finding root nodes in the final clause of the sentence, and it incorrectly analyses a wide range of right-headed Mandarin complex clauses. Statistically, there are 142 sentence-initial subordinate clauses mistakenly analysed as main clauses and only 6 reverse errors. By contrast, there are 278 sentence-final main clauses mistakenly analysed as subordinate ones and 82 reverse errors.

Sometimes this divergence further interacts with ways in which English and Mandarin alternate between clause coordination and subordination. Thus, Mandarin tends to describe sequences of events as a pair of an adverbial clause and a main clause (*after having taken a shower, he dried himself*) instead of as two coordinated clauses (*he took a shower and dried himself*). English UD treats the first conjoined clause as the matrix one, while it is often advcl in Mandarin, and the absence of overt unambiguous complementisers makes it hard

---

[5]E.g., as discussed in §3, the model expects direct quotes to have the form ccomp (quote) + root (verb of speech) and not root (quote) + parataxis (verb of speech).

[6]dep labels different kinds of hard-to-analyse relations and is frequent in Mandarin PUD (397 occurrences).

for the model to see beyond mere frequencies.

A similar situation obtains with some English postposed descriptive subordinate clauses, such as *it's X–that Y* constructions[7] and non-restrictive relative clauses.[8] In these cases, Mandarin uses a coordinative construction, in which the head, according to the UD analysis, is on the right conjunct, corresponding to the English `acl`, and the first conjunct is attached to it using the `dep` label. Again, the English-based model expects to find the root in the first of the two clauses, and there is no overt complementiser to suggest otherwise.

4.3 Attention Patterns

An analysis of the properties of the models underlying these findings is beyond the scope of this paper, but preliminary checks of the attention patterns show that successful models strongly attend to complementisers in the last two layers. As SVO and VSO languages tend to have complementisers before subordinate clauses and SOV languages after (Hawkins, 1990), fine-tuning biases models towards looking for them in only one direction. The attention of subordinate-clause heads to main-clause heads is weaker, presumably due to higher lexical variety in that position.

5 Related Work

Both aspects of our analysis – subordinate-clause detection and the study of word-order effects – have been addressed but not in conjunction and not in a multiple-source-language setting. Our study extends previous approaches by providing a ZS 'upper baseline' derived from the study of the performance of several monolingual models and then conducting a novel many-sources-to-many-target analysis of zero-shot performance.

Lin et al. (2019) test BERT on the auxiliary-classification task (main vs. subordinate clause) as part of their investigation of BERT's linguistic knowledge. Rönnqvist et al. (2019) extend this analysis to the multilingual setting with a focus on Nordic languages.

Word-order differences have been shown to impact the performance of English-based cross-lingual models, especially in the domain of syntactic parsing (Ahmad et al., 2019) and with tasks that rely on syntactic information (Liu et al., 2020;

Arviv et al., 2021), while reordering has been long known to be an efficient preprocessing step in syntactic transfer (Rasooli and Collins, 2019) and machine translation, both statistical (Wang et al., 2007) and neural (Chen et al., 2019).

6 Conclusion

We extend previous work on syntactic capabilities of BERT, mostly focusing on English, by providing a more comprehensive analysis of its performance on the task of subordinate-clause detection in multiple languages and language pairs in the zero-shot setting. We show that the performance of single-language models is uneven across languages: East and Southeast Asian languages with less rigid boundaries between POS categories and coordination and subordination prove harder to model. We also show that mBERT's performance in the zero-shot setting, while being largely correlated with the size of the pre-training and fine-tuning corpora, with Russian being the best source language across the board, is well aligned with the word-order typology: language pairs with congruent word orders demonstrate better results, with both SVO and SOV orders having higher in-group than across-group accuracies. A single pair of VSO languages in the data further corroborates this finding, showing that the verb-final order is not important *per se*.

The clause-initial position of complementisers in VSO languages partly blurs this effect and helps SVO languages with large training corpora serve as good sources for fine-tuning, but even Russian and English fail on SOV languages, where complementisers tend to be postposed and dependent-clause predicates never appear in the sentence-final position. This shows that at least for some tasks, training on a single source language is not enough. Moreover, our results from single-language modelling seem to indicate that even superficially simple syntactic tasks vary in difficulty across languages, which imposes a hard limit on how well cross-lingual projection can perform.

Acknowledgements

---

[7] *It's fantastic that they got the Paris Agreement, but...*

[8] *However, they could not find this same pattern in tissues such as the bladder, which are not directly exposed.*

References

Wasi Ahmad, Zhisong Zhang, Xuezhe Ma, Eduard Hovy, Kai-Wei Chang, and Nanyun Peng. 2019. On difficulties of cross-lingual transfer with order differences: A case study on dependency parsing. 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 2440–2452, Minneapolis, Minnesota. Association for Computational Linguistics.

Ofir Arviv, Dmitry Nikolaev, Taelin Karidi, and Omri Abend. 2021. On the relation between syntactic divergence and zero-shot performance. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4803–4817, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Eduard Bejček, Eva Hajičová, Jan Hajič, Pavlína Jínová, Václava Kettnerová, Veronika Kolářová, Marie Mikulová, Jiří Mírovský, Anna Nedoluzhko, Jarmila Panevová, Lucie Poláková, Magda Ševčíková, Jan Štěpánek, and Šárka Zikánová. 2013. Prague dependency treebank 3.0. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Riyaz Ahmad Bhat, Rajesh Bhatt, Annahita Farudi, Prescott Klassen, Bhuvana Narasimhan, Martha Palmer, Owen Rambow, Dipti Misra Sharma, Ashwini Vaidya, Sri Ramagurumurthy Vishnu, et al. 2017. The Hindi/Urdu treebank project. In *Handbook of Linguistic Annotation*. Springer Press.

Johannes Bjerva and Isabelle Augenstein. 2021. Does typological blinding impede cross-lingual sharing? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 480–486, Online. Association for Computational Linguistics.

Emanuel Borges Völker, Maximilian Wendt, Felix Hennig, and Arne Köhn. 2019. HDT-UD: A very large Universal Dependencies treebank for German. In *Proceedings of the Third Workshop on Universal Dependencies (UDW, SyntaxFest 2019)*, pages 46–57, Paris, France. Association for Computational Linguistics.

Kehai Chen, Rui Wang, Masao Utiyama, and Eiichiro Sumita. 2019. Neural machine translation with reordering embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1787–1799, Florence, Italy. 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.

Sungdai Cho and John Whitman. 2020. *Korean: A linguistic introduction*. Cambridge University Press, Cambridge; New York.

Jayeol Chun, Na-Rae Han, Jena D. Hwang, and Jinho D. Choi. 2018. Building Universal Dependency treebanks in Korean. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Bernard Comrie. 1981. *Language universals and linguistic typology*. University of Chicago Press, Chicago, IL.

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.

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

Bruno Guillaume, Marie-Catherine de Marneffe, and Guy Perrier. 2019. Conversion et améliorations de corpus du français annotés en universal dependencies. *Revue TAL*, 60(2):71–95.

Jan Hajič, Otakar Smrž, Petr Zemánek, Petr Pajas, Jan Šnaidauf, Emanuel Beška, Jakub Kracmar, and Kamila Hassanová. 2009. Prague arabic dependency treebank 1.0. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

John A Hawkins. 1990. A parsing theory of word order universals. *Linguistic inquiry*, 21(2):223–261.

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.

Nanjiang Jiang and Marie-Catherine de Marneffe. 2019. Evaluating BERT for natural language inference: A case study on the CommitmentBank. 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 6086–6091, Hong Kong, China. Association for Computational Linguistics.

Konstantinos Kogkalidis and Gijs Winholds. 2022. Discontinuous constituency and BERT: A case study of Dutch. *arXiv preprint arXiv:2203.01063*.

Yongjie Lin, Yi Chern Tan, and Robert Frank. 2019. Open sesame: Getting inside BERT's linguistic knowledge. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 241–253, Florence, Italy. Association for Computational Linguistics.

Zihan Liu, Genta Indra Winata, Samuel Cahyawijaya, Andrea Madotto, Zhaojiang Lin, and Pascale Fung. 2020. On the importance of word order information in cross-lingual sequence labeling. *arXiv preprint arXiv:2001.11164*.

Amnon Lotan, Asher Stern, and Ido Dagan. 2013. TruthTeller: Annotating predicate truth. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 752–757, Atlanta, Georgia. Association for Computational Linguistics.

Sabrina J Mielke, Zaid Alyafeai, Elizabeth Salesky, Colin Raffel, Manan Dey, Matthias Gallé, Arun Raja, Chenglei Si, Wilson Y Lee, Benoît Sagot, and Samson Tan. 2021. Between words and characters: A brief history of open-vocabulary modeling and tokenization in NLP. *arXiv preprint arXiv:2112.10508*.

Phuong-Thai Nguyen, Xuan-Luong Vu, Thi-Minh-Huyen Nguyen, Van-Hiep Nguyen, and Hong-Phuong Le. 2009. Building a large syntactically-annotated corpus of Vietnamese. In *Proceedings of the Third Linguistic Annotation Workshop (LAW III)*, pages 182–185, Suntec, Singapore. Association for Computational Linguistics.

Thanh Bon Nguyen, Thi Minh Huyen Nguyen, Laurent Romary, and Xuan Luong Vu. 2004. Lexical descriptions for Vietnamese language processing. In *The 1st International Joint Conference on Natural Language Processing - IJCNLP'04 / Workshop on Asian Language Resources*, Sanya, Hainan Island, China. Colloque avec actes et comité de lecture. internationale.

Dmitry Nikolaev, Taelin Karidi, Neta Kenneth, Veronika Mitnik, Lilja Saeboe, and Omri Abend. 2020. Morphosyntactic predictability of translationese. *Linguistics Vanguard*, 6(1).

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.

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.

Ella Rabinovich, Noam Ordan, and Shuly Wintner. 2017. Found in translation: Reconstructing phylogenetic language trees from translations. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 530–540, Vancouver, Canada. Association for Computational Linguistics.

Mohammad Sadegh Rasooli and Michael Collins. 2019. Low-resource syntactic transfer with unsupervised source reordering. 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 3845–3856, Minneapolis, Minnesota. Association for Computational Linguistics.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Samuel Rönnqvist, Jenna Kanerva, Tapio Salakoski, and Filip Ginter. 2019. Is multilingual BERT fluent in language generation? In *Proceedings of the First NLPL Workshop on Deep Learning for Natural Language Processing*, pages 29–36, Turku, Finland. Linköping University Electronic Press.

Punardeep Sikka and Vijay Mago. 2020. A survey on text simplification. *arXiv preprint arXiv:2008.08612*.

Chao-Fen Sun and Talmy Givón. 1985. On the so-called SOV word order in Mandarin Chinese: A quantified text study and its implications. *Language*, pages 329–351.

Roelien C Timmer, David Liebowitz, Surya Nepal, and Salil S Kanhere. 2021. Can pre-trained transformers be used in detecting complex sensitive sentences? – A Monsanto case study. In *2021 Third IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications (TPS-ISA)*, pages 90–97. IEEE.

Reut Tsarfaty. 2013. A unified morpho-syntactic scheme of Stanford dependencies. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 578–584, Sofia, Bulgaria. Association for Computational Linguistics.

Chao Wang, Michael Collins, and Philipp Koehn. 2007. Chinese syntactic reordering for statistical machine translation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 737–745, Prague, Czech Republic. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen,

Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Daniel Zeman, Martin Popel, Milan Straka, Jan Hajič, Joakim Nivre, Filip Ginter, Juhani Luotolahti, Sampo Pyysalo, Slav Petrov, Martin Potthast, Francis Tyers, Elena Badmaeva, Memduh Gokirmak, Anna Nedoluzhko, Silvie Cinková, Jan Hajič jr., Jaroslava Hlaváčová, Václava Kettnerová, Zdeňka Urešová, Jenna Kanerva, Stina Ojala, Anna Missilä, Christopher D. Manning, Sebastian Schuster, Siva Reddy, Dima Taji, Nizar Habash, Herman Leung, Marie-Catherine de Marneffe, Manuela Sanguinetti, Maria Simi, Hiroshi Kanayama, Valeria de Paiva, Kira Droganova, Héctor Martínez Alonso, Çağrı Çöltekin, Umut Sulubacak, Hans Uszkoreit, Vivien Macketanz, Aljoscha Burchardt, Kim Harris, Katrin Marheinecke, Georg Rehm, Tolga Kayadelen, Mohammed Attia, Ali Elkahky, Zhuoran Yu, Emily Pitler, Saran Lertpradit, Michael Mandl, Jesse Kirchner, Hector Fernandez Alcalde, Jana Strnadová, Esha Banerjee, Ruli Manurung, Antonio Stella, Atsuko Shimada, Sookyoung Kwak, Gustavo Mendonça, Tatiana Lando, Rattima Nitisaroj, and Josie Li. 2017. CoNLL 2017 shared task: Multilingual parsing from raw text to Universal Dependencies. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–19, Vancouver, Canada. Association for Computational Linguistics.

Xinliang Frederick Zhang and Marie-Catherine de Marneffe. 2021. Identifying inherent disagreement in natural language inference. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4908–4915, Online. Association for Computational Linguistics.

A Appendix

A.1 Abbreviations

af Afrikaans

ar Standard Arabic

cs Czech

cy Welsh

de German

en English

es Spanish

eu Basque

ga Irish

fi Finnish

fr French

he Hebrew

hi Hindi

hy Eastern Armenian

id Indonesian

is Icelandic

it Italian

ja Japanese

ko Korean

pl Polish

pt Portuguese

ru Russian

sv Swedish

th Thai

tr Turkish

vi Vietnamese

zh Mandarin Chinese

A.2 Corpora

In addition to the Parallel Universal Dependencies collection (Zeman et al., 2017), the following corpora were used to train and/or validate models:

- Afribooms: UD Afrikaans-AfriBooms, https://github.com/ UniversalDependencies/UD_ Afrikaans-AfriBooms

- ArmTDP: Universal Dependencies treebank for Eastern Armenian, https://github.com/ UniversalDependencies/UD_ Armenian-ArmTDP

- BDT: Basque UD treebank, https://github.com/ UniversalDependencies/UD_ Basque-BDT

- CCG: Corpws Cystrawennol y Gymraeg (Syntactic Corpus of Welsh), https://github.com/ UniversalDependencies/UD_ Welsh-CCG

- EWT: Universal Dependencies English Web Treebank, https://github.com/ UniversalDependencies/UD_ English-EWT

- GSD (French): UD French GSD, https://github.com/ UniversalDependencies/UD_ French-GSD (Guillaume et al., 2019)

- GSD (Japanese): UD Japanese Treebank, https://github.com/ UniversalDependencies/UD_ Japanese-GSD

- GSD (Korean): Google Korean Universal Dependency Treebank, https://github.com/ UniversalDependencies/UD_ Korean-GSD (Chun et al., 2018)

- GSD (Mandarin): Traditional Chinese Universal Dependencies Treebank, https://github.com/ UniversalDependencies/UD_ Chinese-GSD

- GSD (Spanish): Spanish UD treebank,
  `https://github.com/UniversalDependencies/UD_Chinese-GSD`

- HDT: UD version of the Hamburg Dependency Treebank,
  `https://github.com/UniversalDependencies/UD_German-HDT` (Borges Völker et al., 2019)

- HDTB: Hindi Universal Dependency Treebank,
  `https://github.com/UniversalDependencies/UD_Hindi-HDTB` (Bhat et al., 2017)

- HTB: Universal Dependencies Corpus for Hebrew,
  `https://github.com/UniversalDependencies/UD_Hebrew-HTB` (Tsarfaty, 2013)

- IDT: Irish UD Treebank,
  `https://github.com/UniversalDependencies/UD_Irish-IDT`

- KENET: Turkish-Kenet UD Treebank,
  `https://github.com/UniversalDependencies/UD_Turkish-Kenet`

- PADT: UD version of the Prague Arabic Dependency Treebank,
  `https://github.com/UniversalDependencies/UD_Arabic-PADT` (Hajič et al., 2009)

- PDT: UD version of the Prague Dependency Treebank,
  `https://github.com/UniversalDependencies/UD_Czech-PDT` (Bejček et al., 2013)

- Syntagrus: SynTagRus Dependency Treebank,
  `https://github.com/UniversalDependencies/UD_Russian-SynTagRus`

- VTB: UD version of the VLSP constituency treebank,
  `https://github.com/`

`UniversalDependencies/UD_Vietnamese-VTB` (Nguyen et al., 2009)

A.3 Single-language model results

The results attained by the models fine-tuned and tested on the same language are shown in Table 3. See § A.2 for the details about the train and test corpora.

Language	Train corpus	Test corpus	Model	#Train	#Test	Main-Main	Main-Sub	Sub-Main	Sub-Sub	Acc.
Mandarin	GSD-Train	PUD	mBERT	3196	736	556	180	122	1364	86.4
Mandarin	GSD-Train	PUD	HFL-BERT-WWM	3196	736	570	166	85	1401	88.7
Vietnamese	VTB-Train	VTB-Dev	mBERT	1105	619	510	109	90	1283	90
Korean	GSD-Train	PUD	mBERT	2201	618	603	15	149	936	90.4
Arabic	PADT-Train	PUD	mBERT	3755	520	436	84	31	752	91.2
Hindi	HDTB-Train	PUD	mBERT	5167	565	506	59	32	831	93.6
German	HDT-Train	PUD	mBERT	55938	441	427	14	49	578	94.1
Armenian	ArmTDP-Train	ArmTDP-Dev	mBERT	1165	149	145	4	21	269	94.3
Turkish	KENET-Train	PUD	mBERT	6784	731	653	78	25	1338	95.1
Welsh	CCG-Train	CCG-Dev	mBERT	377	341	315	26	27	824	95.6
Indonesian	GSD-Train	PUD	mBERT	2770	572	553	19	42	923	96
Basque	BDT-Train	BDT-Dev	mBERT	3181	1029	979	50	39	1758	96.9
Spanish	GSD-Train	PUD	mBERT	7247	548	513	35	5	824	97.1
Irish	IDT-Train	IDT-Dev	mBERT	2323	236	226	10	8	441	97.4
English	EWT-Train	PUD	BERT-LARGE-CASED	5968	556	529	27	4	915	97.9
Hebrew	HTB-Train	HTB-Dev	mBERT	2342	206	201	5	4	297	98.2
Afrikaans	Afribooms-Train	Afribooms-Train	mBERT	643	97	96	1	2	142	98.8
French	GSD-Train	PUD	mBERT	7712	572	563	9	6	956	99
Japanese	GSD-Train	PUD	TOHOKU-BERT-LARGE	5101	844	838	8	18	2090	99.1
Czech	PDT-Train	PUD	mBERT	26277	504	502	2	3	779	99.6
Russian	Syntagrus-Train	PUD	mBERT	32851	595	593	2	2	961	99.7

Table 3: Performance of single-language models across languages. #Train and #Test denote the number of sentences in the train and test corpus respectively. In the 'Main-Main', 'Main-Sub', 'Sub-Main', and 'Sub-Sub' columns, the part before the hyphen is the gold label of a predicate (main/subordinate clause) and the second part is the guessed label. Acc: Accuracy.