Enough is Enough! A Case Study on the Effect of Data Size for Evaluation Using Universal Dependencies

Rob van der Goot, Zoey Liu, Max Müller-Eberstein
IT University of Copenhagen, University of Florida, IT University of Copenhagen
robv@itu.dk, liu.ying@ufl.edu, mamy@itu.dk

Abstract

When creating a new dataset for evaluation, one of the first considerations is the size of the dataset. If our evaluation data is too small, we risk making unsupported claims based on the results on such data. If, on the other hand, the data is too large, we waste valuable annotation time and costs that could have been used to widen the scope of our evaluation (i.e. annotate for more domains/languages). Hence, we investigate the effect of the size, and a variety of sampling strategies of evaluation data to optimize annotation efforts, using dependency parsing as a test case. We show that for in-language, in-domain datasets, 5,000 tokens is enough to obtain a reliable ranking of different parsers; especially if the data is distant enough from the training split (otherwise, we recommend 10,000). In cross-domain setups, the same amounts are required, but in cross-lingual setups much less (2,000 tokens) is enough.

Keywords: Evaluation Methodologies, Parsing, Grammar, Syntax, Treebank

1. Introduction

When creating a new dataset, it is standard procedure in Natural Language Processing (NLP), and more widely machine learning, to split your data into a training, development (also called validation or evaluation split), and test split to avoid overfitting. The training data is used to train a machine learning model and can be omitted in unsupervised (or cross-domain/lingual) setups. The development data is used in the development phase to design and tune the model(s) of interest. Finally, the test data is used to confirm the main conclusions and compare against previous work.[1] We will refer to the development and test split as evaluation splits, as they have a similar use-case (comparing models). Historically, splitting the data in 60%-20%-20% (train-dev-test) or 80%-10%-10% for larger datasets has been a popular strategy to obtain these data splits, and the full data size was decided based on budget availability.

Recent work on model evaluation has proposed to identify adequate sample sizes using statistical power analyses for classification and translation tasks (Card et al., 2020), which, in turn, would require large amounts of simulated data and scores. We use an alternative strategy, and use only real outputs from parsers and focus on a structured prediction task: dependency parsing. Although the methods presented can also be applied to other tasks and datasets.

We will use data from the Universal Dependencies dataset (Zeman et al., 2021), as it is (one of) the largest and most diverse annotated corpora available in NLP. UD uses the previously mentioned 80%-10%-10% for each treebank if there is enough data. For smaller treebanks (<100,000 tokens), they suggest to use 10,000 tokens for development and 10,000 for test.[2] We plotted the sizes of the development and test sets for all treebanks from UD2.9 (Zeman et al., 2021) in Figure 1. The guidelines for data splitting are clearly reflected, and implicitly suggest that 10,000 tokens are a good amount for an evaluation split when creating a new treebank.

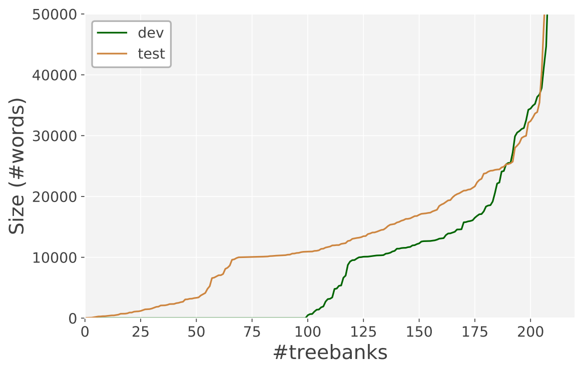

Figure 1: Plot of the sizes (cumulative) of the dev and test splits of all 217 official UD v2.9 treebanks (in 122 languages).

Since the use case of development data and test data is similar (they are both used to compare varieties of models, just in different stages of the research), in this work, we will assume that their ideal size is the same. Furthermore, we will evaluate data size on the token-level, as annotation and evaluation of UD is also done on the word-level.

In this work, we evaluate the effect of evaluation data size using two strategies: 1) Com-

[1] Recently, there have been more cases of using testing data during development (van der Goot, 2021).

[2] More detailed descriptions available at: https://universaldependencies.org/release_checklist.html.

LREC-COLING 2024, pages 6167–6176
20-25 May, 2024.

Source Treebank	#tokens	Domain-transfer treebanks
English-WSJ	1,173,766	EN-Atis, EN-ESL, EN-EWT, Naija-NSC
Italian-ISDT	278,429	IT-PoSTWITA
Russian-GSD	98,000	RU-Taiga
English-EWT	251,489	EN-Atis, EN-EWT, EN-WSJ, Naija-NSC
Italian-ISDT	119,342	IT-PoSTWITA
Russian-Taiga	138,908	RU-GSD

Table 1: List of used datasets. The top source treebanks are news, the bottom are web data.

pare rankings of models using weighted Kendall's Tau (Kendall, 1938) over rankings of parsers; 2) Compare model pairs with significance testing using Almost Stochastic Order (Del Barrio et al., 2018; Dror et al., 2019). Furthermore, we investigate cross-domain and cross-lingual setups in Section 3.2.[3]

2. Setup

2.1. Data

In the remainder of this paper, we focus on the treebanks listed in Table 1. We focus on the news and web domain, motivated by dataset availability. Large (>95,000 words) treebanks are available for Italian, and Russian from UD v2.9. To obtain a news treebank for English, we used the Stanford Converter on the Penn Treebank (Bies et al., 1995). Furthermore, we added cross-domain datasets with a size of >50,000 tokens where available.

We re-split the data from each treebank to gauge the effect of having different varieties of the development set.[4] Explorations with varying train sizes showed that a size of 50,000 tokens gave a good tradeoff in training time and accuracy, so we used this for our main experiments. Previous work has shown that training on random samples would lead to artificially high scores (Gorman and Bedrick, 2019; Çöltekin, 2020) because texts from the same document, writer, etc. would appear in both the training set and the development set. With that in mind, we used the first 50,000 tokens for training, and applied the following strategies for selecting the instances for the development data (visualized in Figure 2):

- SEQ: we took the last M samples as development data, based on the assumption that sentences in each treebank are ordered; this way there will be less chance the training and development data have overlap of documents/writers, etc.

SEQ	T	T	T				D	D
RAND	T	T	T	D		D		
RAND_SEQ	T	T	T		D	D		

Figure 2: Visualization of how we split existing datasets. The full bars are the original training data concatenated with the original development data, and each square represents a portion of this data. T = a portion of the training data, D = a portion of the development data.

- RAND: random sampling without replacement on the remaining instances. Note that this is different compared to the random splits proposed in Søgaard et al. (2021) and Çöltekin (2020), because the train split is from a separate range.

- RAND_SEQ: we took an ordered sequence of size M at a random starting point of the remaining data. Note that the web treebanks do not contain a chronological order, but the spoken treebanks do.

M was measured in numbers of tokens but we sampled whole sentences to not interrupt the context. We experimented with $M \in [100, 200, 500, 1000, 2000, 5000, 10000, 20000, treebank\_size]$.

2.2. Parsers

We use MaChAmp (v0.3beta (van der Goot et al., 2021)): A toolkit focused on multi-task learning. It uses a pre-trained language model as encoder and allows for multiple decoder heads (one for each task). However, we trained it with a single dependency head using a deep-biaffine parser (Dozat and Manning, 2017). We create different versions of the parser by iterating over all commonly-available multilingual language models that fit on our 32GB GPUs (33 in total, full list in Appendix A), and training with five different random initializations each. We used the standard Labeled Attachment Score (LAS), as implemented by Zeman et al. (2018) as the evaluation metric.

3. Results

3.1. In-treebank results

3.1.1. Comparison of rankings

For each target development set (size + split strategy), we train all 33 models using five initializations on each respective training split, and average the

[3] Code is available on: https://bitbucket.org/robvanderg/data_size/.

[4] If the train split alone was too small, we concatenated train-dev(-test).

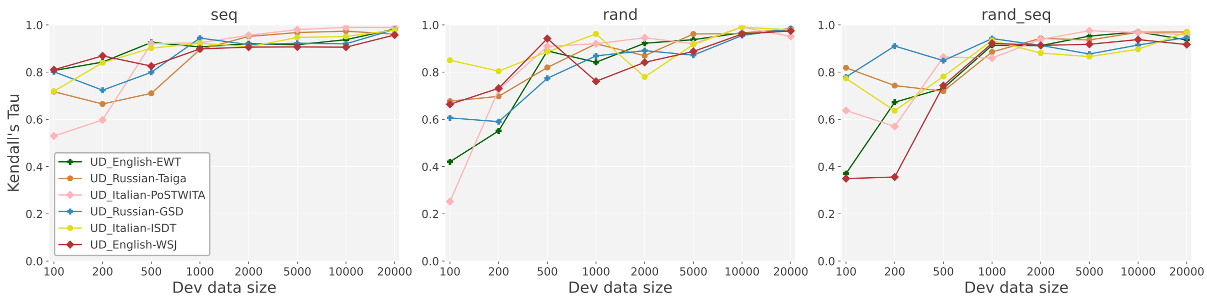

Figure 3: Kendall's Tau Scores for each treebank for all of our data splitting strategies. Note that the X-axis is divided based on our size sample, and is not scaled.

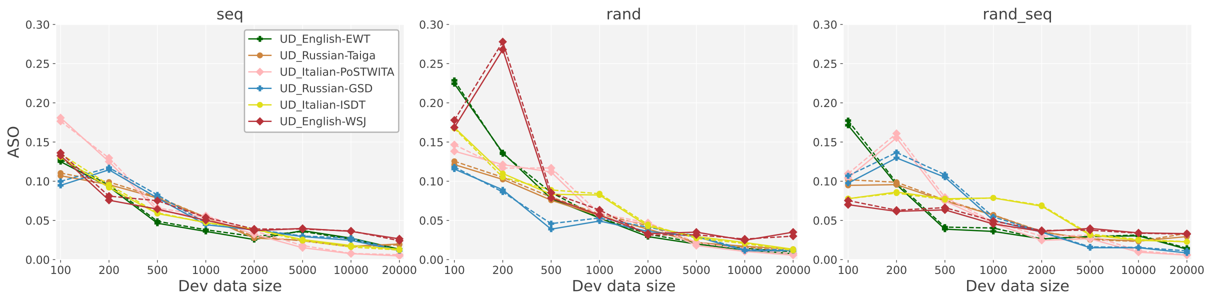

Figure 4: ASO distances. The dashed lines represent the average distance in ϵ_{min}, and the full line the percentage of cases where (binary) disagreement is found.

respective LAS from each run. The ranking obtained based on the largest possible development set is used as the gold ranking (i.e., the true order of the models), and is compared to the ranking of the smaller split sizes.

We used weighted Kendall's Tau (Kendall, 1938; Vigna, 2015) to quantify the differences between the rankings. Kendall's Tau measures correlations between rankings and returns a value between -1 and 1, where 1 indicates perfect agreement, and -1 means that the rankings are reversed. A value above 0.4 indicates a strong relation between both rankings (Botsch, 2011).

Figure 3 shows the Kendall's Tau scores for each treebank. The first observation is that the scores tend to converge (less variance for larger sizes). This indicates that a robust optimal ranking is found, and supports our design decision of considering the ranking at the maximum size as the gold standard. In general, the random splitting strategies (Rand and Rand_Seq) show higher correlations for smaller data sizes compared to the Seq strategy. Across treebanks, the Kendall's Tau seems to converge at around 2,000 or 5,000 instances, but strong correlations (>0.4) can already be found for much smaller evaluation splits.

We also investigate a more challenging setup by considering each seed as a separate model, so we are also charged with ranking the same language model that uses different seeds (a total of 5 $\times$ 33 parsers). Full results are reported in Appendix B; they show that all trends remain similar, except that Kendall's Tau scores are slightly lower. This indicates that our proposed method for estimating the effect of evaluation data set size is robust across random initializations.

3.1.2. Significance Testing

Now that we have established that there is a strong correlation between smaller development sizes and the maximum size, we next quantify this effect empirically by running significance tests between the performance of all model within each development size. Once again, we compare the results of the smaller development splits to the maximum-size development split. Intuitively, we compute a matrix for each data size that consists of a value for every model versus every other model, for which each significant difference is marked. We use the Almost Stochastic Order test (Dror et al., 2019; Del Barrio et al., 2018) as implemented by Ulmer et al. (2022) over the five random seeds to estimate significance. If ASO determines $\epsilon_{min} < 0.5$, we consider model A to be significantly better than model B.

We used two metrics to evaluate the consistency of the significance testing results: 1) The amount of disagreement of significance results, where ϵ_{min} is converted to a binary value ($\epsilon_{min} < 0.5$), and for which we count the number of different entries across two development data sizes (i.e., the overlap of two binary matrices). 2) The average absolute

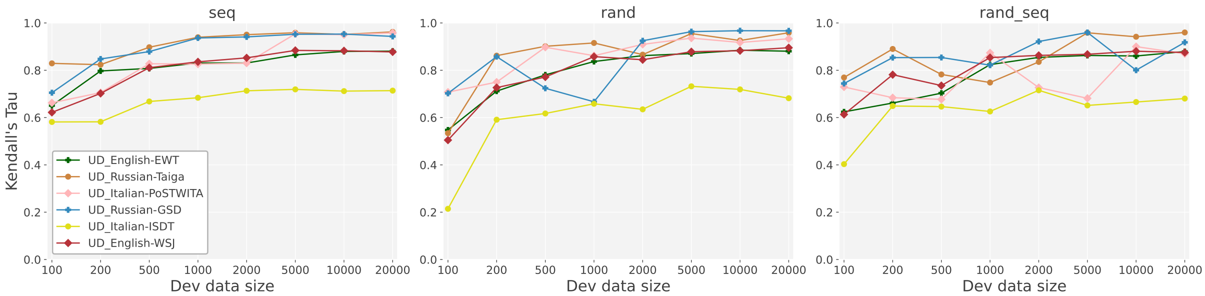

Figure 5: Cross-domain Kendall's Tau Scores for each treebank for all of our data splitting strategies.

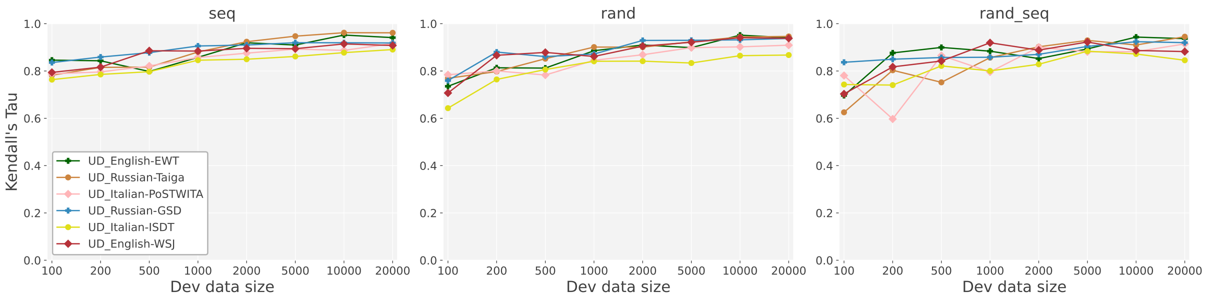

Figure 6: Cross-lingual Kendall's Tau Scores for each treebank for all of our data splitting strategies.

difference of the ϵ_{min} values for all model pairs across data sizes.

The results of ASO testing in Figure 4 show that these two metrics have an almost perfect correlation (Pearson correlation is 0.99), as is also visually indicated by the nearly overlapping dashed and full lines. Furthermore, the trends are highly similar to the Kendall's Tau scores (Figure 3). We again see a slightly earlier convergence for the random splitting strategies, but still see minor improvements for the larger sizes, especially with the seq sampling strategy. Since the trends are highly similar to the Kendall's Tau scores, and ASO is more computationally costly, we focus solely on Kendall's Tau in the following section.

3.2. Cross-lingual/domain Results

Only 131/217 of all UD v2.9 treebanks have training data, and as such, the evaluation of parsers on test-only treebanks relies on cross-treebank performance.[5] In order to estimate sufficient development set sizes for this common scenario, we additionally perform the Kendall's Tau experiments from Section 3.1.1 on the cross-lingual and cross-domain setups introduced in Section 2.1.

For cross-domain setups, Figure 5 shows that smaller dev-sizes are already more stable, but if the best possible ranking is desired, sizes should be similar compared to the in-domain results (Figure 3).

[5]Note that we only consider in-domain treebanks for the cross-lingual experiments.

The cross-lingual results (Figure 6) show that very minimal amounts of data lead to similar rankings as the largest development splits; a size of 500-2,000 tokens already leads to an almost perfect ranking. Interestingly, the most stable rankings are obtained with the `seq` strategy; taking the consecutive instances with the largest distance from the training data. It should be noted that our sample of languages is relatively closely related to each other; we expect that in cross-domain samples with more distinct languages, differences across parsers will be more profound and even smaller samples could be indicative enough.

4. Conclusion

We have investigated the effect of dataset size on evaluation for a variety of setups within dependency parsing. Across two measures of model performance rankings (Kendall's Tau and ASO), we have shown that the target size of the official UD guidelines of 10,000 tokens is a safe choice for ensuring representative model performance rankings, but that even smaller sizes of 2,000 to 5,000 tokens have sufficient predictive power in our sample of treebanks. Furthermore, if we target cross-domain setups, good rankings can be obtained using smaller sizes. Cross-lingually even smaller datasets down to around 500–2,000 tokens are sufficient for predicting final model rankings. For reducing these minimum data sizes even further, future work could investigate more targeted sampling strategies with a focus on increased data diversity.

5. Bibliographical References

Francesco Barbieri, Luis Espinosa-Anke, and Jose Camacho-Collados. 2021. A multilingual language model toolkit for twitter. *arXiv preprint arXiv:2104.12250*.

Ann Bies, Mark Ferguson, Karen Katz, Robert MacIntyre, Victoria Tredinnick, Grace Kim, Mary Ann Marcinkiewicz, and Britta Schasberger. 1995. Bracketing guidelines for Treebank II style Penn Treebank project. Technical report, University of Pennsylvania.

R Botsch. 2011. Chapter 12: Significance and measures of association. *Scopes and Methods of Political Science*.

Dallas Card, Peter Henderson, Urvashi Khandelwal, Robin Jia, Kyle Mahowald, and Dan Jurafsky. 2020. With little power comes great responsibility. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9263–9274, Online. Association for Computational Linguistics.

Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. 2021. InfoXLM: An information-theoretic framework for cross-lingual language model pre-training. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3576–3588, Online. Association for Computational Linguistics.

Hyung Won Chung, Thibault Fevry, Henry Tsai, Melvin Johnson, and Sebastian Ruder. 2020. Rethinking embedding coupling in pre-trained language models. In *International Conference on Learning Representations*.

Jonathan H. Clark, Dan Garrette, Iulia Turc, and John Wieting. 2022. Canine: Pre-training an efficient tokenization-free encoder for language representation. *Transactions of the Association for Computational Linguistics*, 10:73–91.

Çağrı Çöltekin. 2020. Verification, reproduction and replication of NLP experiments: a case study on parsing Universal Dependencies. In *Proceedings of the Fourth Workshop on Universal Dependencies (UDW 2020)*, pages 46–56, Barcelona, Spain (Online). Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Nicola De Cao, Ledell Wu, Kashyap Popat, Mikel Artetxe, Naman Goyal, Mikhail Plekhanov, Luke Zettlemoyer, Nicola Cancedda, Sebastian Riedel, and Fabio Petroni. 2022. Multilingual autoregressive entity linking. *Transactions of the Association for Computational Linguistics*, 10:274–290.

Eustasio Del Barrio, Juan A Cuesta-Albertos, and Carlos Matrán. 2018. An optimal transportation approach for assessing almost stochastic order. In *The Mathematics of the Uncertain*, pages 33–44. Springer.

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

Timothy Dozat and Christopher D. Manning. 2017. Deep Biaffine Attention for Neural Dependency Parsing. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Rotem Dror, Segev Shlomov, and Roi Reichart. 2019. Deep dominance - how to properly compare deep neural models. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 2773–2785. Association for Computational Linguistics.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Kyle Gorman and Steven Bedrick. 2019. We need to talk about standard splits. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2786–2791, Florence, Italy. Association for Computational Linguistics.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*.

Maurice G Kendall. 1938. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, et al. 2021. Few-shot learning with multilingual language models. *arXiv preprint arXiv:2112.10668*.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Markus Sagen. 2021. Large-context question answering with cross-lingual transfer. Master's thesis, Uppsala University, Department of Information Technology.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *5th Workshop on Energy Efficient Machine Learning and Cognitive Computing - NeurIPS*.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Anders Søgaard, Sebastian Ebert, Jasmijn Bastings, and Katja Filippova. 2021. We need to talk about random splits. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1823–1832, Online. Association for Computational Linguistics.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT – building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Lisboa, Portugal. European Association for Machine Translation.

Dennis Ulmer, Christian Hardmeier, and Jes Frellsen. 2022. deep-significance-easy and meaningful statistical significance testing in the age of neural networks. *arXiv preprint arXiv:2204.06815*.

Rob van der Goot. 2021. We need to talk about train-dev-test splits. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4485–4494, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Rob van der Goot, Ahmet Üstün, Alan Ramponi, Ibrahim Sharaf, and Barbara Plank. 2021. Massive choice, ample tasks (MaChAmp): A toolkit for multi-task learning in NLP. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 176–197, Online. Association for Computational Linguistics.

Sebastiano Vigna. 2015. A weighted correlation index for rankings with ties. In *Proceedings of the 24th international conference on World Wide Web*, pages 1166–1176.

Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. 2022. ByT5: Towards a token-free future with pre-trained byte-to-byte models. *Transactions of the Association for Computational Linguistics*, 10:291–306.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. 2020. LUKE: Deep contextualized entity representations with entity-aware self-attention. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6442–6454, Online. Association for Computational Linguistics.

Daniel Zeman, Jan Hajič, Martin Popel, Martin Potthast, Milan Straka, Filip Ginter, Joakim Nivre, and Slav Petrov. 2018. CoNLL 2018 shared task: Multilingual parsing from raw text to Universal Dependencies. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–21, Brussels, Belgium. Association for Computational Linguistics.

Daniel Zeman, Joakim Nivre, Mitchell Abrams, Elia Ackermann, Noëmi Aepli, Hamid Aghaei, Željko Agić, Amir Ahmadi, Lars Ahrenberg, Chika Kennedy Ajede, Gabrielė Aleksandravičiūtė, Ika Alfina, Lene Antonsen, Katya Aplonova, Angelina Aquino, Carolina Aragon, Maria Jesus Aranzabe, Bilge Nas Arıcan, Ĥórunn Arnardóttir, Gashaw Arutie, Jessica Naraiswari Arwidarasti, Masayuki Asahara, Deniz Baran Aslan, Luma Ateyah, Furkan Atmaca, Mohammed Attia, Aitziber Atutxa, Liesbeth Augustinus, Elena Badmaeva, Keerthana Balasubramani, Miguel Ballesteros, Esha Banerjee, Sebastian Bank, Verginica Barbu Mititelu, Starkaður Barkarson, Rodolfo Basile, Victoria Basmov, Colin Batchelor, John Bauer, Seyyit Talha Bedir, Kepa Bengoetxea, Gözde Berk, Yevgeni Berzak, Irshad Ahmad Bhat, Riyaz Ahmad Bhat, Erica Biagetti, Eckhard Bick, Agnė Bielinskienė, Kristín Bjarnadóttir, Rogier Blokland, Victoria Bobicev, Loïc Boizou, Emanuel Borges Völker, Carl Börstell, Cristina Bosco, Gosse Bouma, Sam Bowman, Adriane Boyd, Anouck Braggaar, Kristina Brokaitė, Aljoscha Burchardt, Marie Candito, Bernard Caron, Gauthier Caron, Lauren Cassidy, Tatiana Cavalcanti, Gülşen Cebiroğlu Eryiğit, Flavio Massimiliano Cecchini, Giuseppe G. A. Celano, Slavomír Čéplö, Neslihan Cesur, Savas Cetin, Özlem Çetinoğlu, Fabricio Chalub, Shweta Chauhan, Ethan Chi, Taishi Chika, Yongseok Cho, Jinho Choi, Jayeol Chun, Juyeon Chung, Alessandra T. Cignarella, Silvie Cinková, Aurélie Collomb, Çağrı Çöltekin, Miriam Connor, Marine Courtin, Mihaela Cristescu, Philemon Daniel, Elizabeth Davidson, Marie-Catherine de Marneffe, Valeria de Paiva, Mehmet Oguz Derin, Elvis de Souza, Arantza Diaz de Ilarraza, Carly Dickerson, Arawinda Dinakaramani, Elisa Di Nuovo, Bamba Dione, Peter Dirix, Kaja Dobrovoljc, Timothy Dozat, Kira Droganova, Puneet Dwivedi, Hanne Eckhoff, Sandra Eiche, Marhaba Eli, Ali Elkahky, Binyam Ephrem, Olga Erina, Tomaž Erjavec, Aline Etienne, Wograine Evelyn, Sidney Facundes, Richárd Farkas, Jannatul Ferdaousi, Marília Fernanda, Hector Fernandez Alcalde, Jennifer Foster, Cláudia Freitas, Kazunori Fujita, Katarína Gajdošová, Daniel Galbraith, Marcos Garcia, Moa Gärdenfors, Sebastian Garza, Fabrício Ferraz Gerardi, Kim Gerdes, Filip Ginter, Gustavo Godoy, Iakes Goenaga, Koldo Gojenola, Memduh Gökırmak, Yoav Goldberg, Xavier Gómez Guinovart, Berta González Saavedra, Bernadeta Griciūtė, Matias Grioni, Loïc Grobol, Normunds Grūzītis, Bruno Guillaume, Céline Guillot-Barbance, Tunga Güngör, Nizar Habash, Hinrik Hafsteinsson, Jan Hajič, Jan Hajič jr., Mika Hämäläinen, Linh Hà Mỹ, Na-Rae Han, Muhammad Yudistira Hanifmuti, Sam Hardwick, Kim Harris, Dag Haug, Johannes Heinecke, Oliver Hellwig, Felix Hennig, Barbora Hladká, Jaroslava Hlaváčová, Florinel Hociung, Petter Hohle, Eva Huber, Jena Hwang, Takumi Ikeda, Anton Karl Ingason, Radu Ion, Elena Irimia, Ọlájídé Ishola, Kaoru Ito, Siratun Jannat, Tomáš Jelínek, Apoorva Jha, Anders Johannsen, Hildur Jónsdóttir, Fredrik Jørgensen, Markus Juutinen, Sarveswaran K, Hüner Kaşıkara, Andre Kaasen, Nadezhda Kabaeva, Sylvain Kahane, Hiroshi Kanayama, Jenna Kanerva, Neslihan Kara, Boris Katz, Tolga Kayadelen, Jessica Kenney, Václava Kettnerová, Jesse Kirchner, Elena Klementieva, Elena Klyachko, Arne Köhn, Abdullatif Köksal, Kamil Kopacewicz, Timo Korkiakangas, Mehmet Köse, Natalia Kotsyba, Jolanta Kovalevskaitė, Simon Krek, Parameswari Krishnamurthy, Sandra Kübler, Oğuzhan Kuyrukçu, Aslı Kuzgun, Sookyoung Kwak, Veronika Laippala, Lucia Lam, Lorenzo Lambertino, Tatiana Lando, Septina Dian Larasati, Alexei Lavrentiev, John Lee, Phuong Lê Hồng, Alessandro Lenci, Saran Lertpradit, Herman Leung, Maria Levina, Cheuk Ying Li, Josie Li, Keying Li, Yuan Li, KyungTae Lim, Bruna Lima Padovani, Krister Lindén, Nikola Ljubešić, Olga Loginova, Stefano Lusito, Andry Luthfi, Mikko Luukko, Olga Lyashevskaya, Teresa Lynn, Vivien Macketanz, Menel Mahamdi, Jean Maillard, Aibek Makazhanov, Michael Mandl, Christopher Manning, Ruli Manurung, Büşra Marşan, Cătălina Mărănduc, David Mareček, Katrin Marheinecke, Héctor Martínez Alonso, Lorena Martín-Rodríguez, André Martins, Jan Mašek, Hiroshi Matsuda, Yuji Matsumoto, Alessandro Mazzei, Ryan McDonald, Sarah McGuinness, Gustavo Mendonça, Tatiana Merzhevich, Niko Miekka, Karina Mischenkova, Margarita Misirpashayeva, Anna Missilä, Cătălin Mititelu, Maria Mitrofan, Yusuke Miyao, AmirHossein Mojiri Foroushani, Judit Molnár, Amirsaeid Moloodi, Simonetta Montemagni, Amir More, Laura Moreno Romero, Giovanni Moretti, Keiko Sophie Mori, Shinsuke Mori, Tomohiko Morioka, Shigeki Moro, Bjartur Mortensen, Bohdan Moskalevskyi, Kadri Muischnek, Robert Munro, Yugo Murawaki, Kaili Müürisep, Pinkey Nainwani, Mariam Nakhlé, Juan Ignacio Navarro Horñiacek, Anna Nedoluzhko, Gunta Nešpore-Bērzkalne, Manuela Nevaci, Luong Nguyễn Thị, Huyền Nguyễn Thị Minh, Yoshihiro Nikaido, Vitaly Nikolaev, Rattima Nitisaroj, Alireza Nourian, Hanna Nurmi, Stina Ojala, Atul Kr. Ojha, Adédayọ̀ Olúòkun, Mai Omura, Emeka Onwuegbuzia, Petya Osenova, Robert Östling, Lilja Øvrelid, Şaziye Betül Özateş, Merve Özçelik, Arzucan Özgür, Balkız Öztürk Başaran, Hyunji Hayley

Park, Niko Partanen, Elena Pascual, Marco Passarotti, Agnieszka Patejuk, Guilherme Paulino-Passos, Angelika Peljak-Łapińska, Siyao Peng, Cenel-Augusto Perez, Natalia Perkova, Guy Perrier, Slav Petrov, Daria Petrova, Jason Phelan, Jussi Piitulainen, Tommi A Pirinen, Emily Pitler, Barbara Plank, Thierry Poibeau, Larisa Ponomareva, Martin Popel, Lauma Pretkalniņa, Sophie Prévost, Prokopis Prokopidis, Adam Przepiórkowski, Tiina Puolakainen, Sampo Pyysalo, Peng Qi, Andriela Rääbis, Alexandre Rademaker, Mizanur Rahoman, Taraka Rama, Loganathan Ramasamy, Carlos Ramisch, Fam Rashel, Mohammad Sadegh Rasooli, Vinit Ravishankar, Livy Real, Petru Rebeja, Siva Reddy, Mathilde Regnault, Georg Rehm, Ivan Riabov, Michael Rießler, Erika Rimkutė, Larissa Rinaldi, Laura Rituma, Putri Rizqiyah, Luisa Rocha, Eiríkur Rögnvaldsson, Mykhailo Romanenko, Rudolf Rosa, Valentin Roșca, Davide Rovati, Olga Rudina, Jack Rueter, Kristján Rúnarsson, Shoval Sadde, Pegah Safari, Benoît Sagot, Aleksi Sahala, Shadi Saleh, Alessio Salomoni, Tanja Samardžić, Stephanie Samson, Manuela Sanguinetti, Ezgi Sanıyar, Dage Särg, Baiba Saulīte, Yanin Sawanakunanon, Shefali Saxena, Kevin Scannell, Salvatore Scarlata, Nathan Schneider, Sebastian Schuster, Lane Schwartz, Djamé Seddah, Wolfgang Seeker, Mojgan Seraji, Syeda Shahzadi, Mo Shen, Atsuko Shimada, Hiroyuki Shirasu, Yana Shishkina, Muh Shohibussirri, Dmitry Sichinava, Janine Siewert, Einar Freyr Sigurðsson, Aline Silveira, Natalia Silveira, Maria Simi, Radu Simionescu, Katalin Simkó, Mária Šimková, Kiril Simov, Maria Skachedubova, Aaron Smith, Isabela Soares-Bastos, Shafi Sourov, Carolyn Spadine, Rachele Sprugnoli, Steinh̑ór Steingrímsson, Antonio Stella, Milan Straka, Emmett Strickland, Jana Strnadová, Alane Suhr, Yogi Lesmana Sulestio, Umut Sulubacak, Shingo Suzuki, Zsolt Szántó, Chihiro Taguchi, Dima Taji, Yuta Takahashi, Fabio Tamburini, Mary Ann C. Tan, Takaaki Tanaka, Dipta Tanaya, Samson Tella, Isabelle Tellier, Marinella Testori, Guillaume Thomas, Liisi Torga, Marsida Toska, Trond Trosterud, Anna Trukhina, Reut Tsarfaty, Utku Türk, Francis Tyers, Sumire Uematsu, Roman Untilov, Zdeňka Urešová, Larraitz Uria, Hans Uszkoreit, Andrius Utka, Sowmya Vajjala, Rob van der Goot, Martine Vanhove, Daniel van Niekerk, Gertjan van Noord, Viktor Varga, Eric Villemonte de la Clergerie, Veronika Vincze, Natalia Vlasova, Aya Wakasa, Joel C. Wallenberg, Lars Wallin, Abigail Walsh, Jing Xian Wang, Jonathan North Washington, Maximilan Wendt, Paul Widmer, Sri Hartati Wijono, Seyi Williams, Mats Wirén, Christian Wittern, Tsegay Woldemariam, Taksum Wong, Alina Wróblewska, Mary Yako, Kayo Yamashita, Naoki Yamazaki, Chunxiao Yan, Koichi Yasuoka, Marat M. Yavrumyan, Arife Betül Yenice, Olcay Taner Yıldız, Zhuoran Yu, Arlisa Yuliawati, Zdeněk Žabokrtský, Shorouq Zahra, Amir Zeldes, He Zhou, Hanzhi Zhu, Anna Zhuravleva, and Rayan Ziane. 2021. Universal dependencies 2.9. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Name	Citation
Helsinki-NLP/opus-mt-mul-en	Tiedemann and Thottingal (2020)
Peltarion/xlm-roberta-longformer-base-4096	Sagen (2021)
bert-base-multilingual-cased	Devlin et al. (2019)
bert-base-multilingual-uncased	Devlin et al. (2019)
bigscience/bloom-560m	Scao et al. (2022)
cardiffnlp/twitter-xlm-roberta-base	Barbieri et al. (2021)
distilbert-base-multilingual-cased	Sanh et al. (2019)
facebook/mbart-large-50	Liu et al. (2020)
facebook/mbart-large-50-many-to-many-mmt	Liu et al. (2020)
facebook/mbart-large-50-many-to-one-mmt	Liu et al. (2020)
facebook/mbart-large-50-one-to-many-mmt	Liu et al. (2020)
facebook/mbart-large-cc25	Liu et al. (2020)
facebook/mgenre-wiki	De Cao et al. (2022)
facebook/nllb-200-distilled-600M	Costa-jussà et al. (2022)
facebook/xglm-564M	Lin et al. (2021)
google/byt5-base	Xue et al. (2022)
google/byt5-small	Xue et al. (2022)
google/canine-c	Clark et al. (2022)
google/canine-s	Clark et al. (2022)
google/mt5-base	Xue et al. (2021)
google/mt5-small	Xue et al. (2021)
google/rembert	Chung et al. (2020)
microsoft/infoxlm-base	Chi et al. (2021)
microsoft/infoxlm-large	Chi et al. (2021)
microsoft/mdeberta-v3-base	He et al. (2021)
setu4993/LaBSE	Feng et al. (2022)
studio-ousia/mluke-base	Yamada et al. (2020)
studio-ousia/mluke-base-lite	Yamada et al. (2020)
studio-ousia/mluke-large	Yamada et al. (2020)
studio-ousia/mluke-large-lite	Yamada et al. (2020)
xlm-mlm-100-1280	Conneau et al. (2020)
xlm-roberta-base	Conneau et al. (2020)
xlm-roberta-large	Conneau et al. (2020)

Table 2: Language models used in our experiments.

A. Language models used

The multilingual language models we used as a basis for our parsers are listed in Table 2.

B. Results with separate seeds

To make the ranking more challenging, we also considered a setup in which each model initialization is treated as a different parser. So instead of taking the average over seeds for each language model, we have five parsers per language model in the final ranking. Results (Table 7) show that the Kendall's Tau scores are only slightly lower compared to the averaged results (Table 3), showcasing the robustness of our approach.

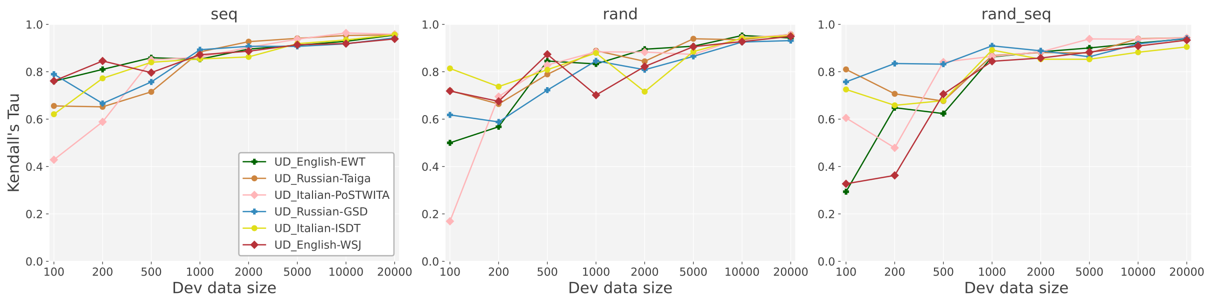

Figure 7: Kendall's Tau Scores for each treebank for all of our data splitting strategies when using each seed as a separate model. Note that the X-axis is divided based on our size sample, and is not scaled.