Verification, Reproduction and Replication of NLP Experiments: a Case Study on Parsing Universal Dependencies

Çağrı Çöltekin
Department of Linguistics
University of Tübingen
ccoltekin@sfs.uni-tuebingen.de

Abstract

As in any field of inquiry that depends on experiments, the verifiability of experimental studies is important in computational linguistics. Despite increased attention to verification of empirical results, the practices in the field are unclear. Furthermore, we argue, certain traditions and practices that are seemingly useful for verification may in fact be counterproductive. We demonstrate this through a set of multi-lingual experiments on parsing Universal Dependencies treebanks. In particular, we show that emphasis on exact replication leads to practices (some of which are now well established) that hide the variation in experimental results, effectively hindering verifiability with a false sense of certainty. The purpose of the present paper is to highlight the magnitude of the issues resulting from these common practices with the hope of instigating further discussion. Once we, as a community, are convinced about the importance of the problems, the solutions are rather obvious, although not necessarily easy to implement.

1 Introduction

The practice of independent verification of empirical findings has been at the core of the modern science. There have been, however, recent failures of reproduction in many fields (Open Science Collaboration, 2015; Freedman et al., 2015), which also instilled public interest with the popular name *reproducibility crisis* (Fidler and Wilcox, 2018). In computational linguistics and natural language processing (NLP), worries about reproducibility have been voiced for over a decade with a notable increase in recent years (Pedersen, 2008; Fokkens et al., 2013; Mieskes, 2017; Branco et al., 2017; Cohen et al., 2017; Reimers and Gurevych, 2017; Branco et al., 2020; Huber and Çöltekin, 2020, just to name a few).

As some of these studies point out, it is often unclear what is meant by the terms *replication* and *reproduction*. The definitions of these terms are often blurred, and they may even be used in opposite meanings in different studies (Cohen et al., 2018). In this paper, we use the term (exact) *replication* to refer to the activity of running the same code on the same data set with the aim of producing the same measurements reported in the original publication. We use the term *reproduction* to refer to the activity of verifying the claims or findings by varying the experimental settings in meaningful ways. As argued earlier (Drummond, 2009), the scientifically interesting and useful activity is *reproduction*, while *replication* has rather little or no use for the purpose of scientific verification.

The confusion, however, is not only limited to the usage of these terms. In many studies, the aim of the activity is also unclear, or often understood as obtaining the same numbers reported in an original study (in our terms *exact replication*). The emphasis on state-of-the-art scores (even with small increase over the previous state of the art) also motivates the exact, 'fair', comparisons based on a single-best score.

On the other hand, many of the statistical data-driven methods involve a number of inevitable sources of variation. For example, any machine learning method will produce varied results when trained and tested on different parts of the data, and many others are also sensitive to other stochastic processes, such as random initialization of their parameters or the order of training instances. Even though this variation is important for evaluating and comparing statistical models, there has been an increasing emphasis on

exact replication of the published scores. For example, the SemEval 2020 reviewer form includes the following scoring instructions for reviewers (emphasis ours):

> 4 = could mostly reproduce the results described here, although there may be some variation *because of sample variance* or minor variations in their interpretation of the protocol or method.
> 5 = could easily reproduce the results and verify the correctness of the results described here.

Hence, although lightly, not being *exactly* replicable due to sample variation is penalized, encouraging participants to eliminate, or hide, all sources of variation. The same or similar reviewing criteria have recently become standard for major publication venues for computational linguistics.

In this paper, we show that the emphasis on exact replication, in fact, hinders the idea of verification of the results. We support this claim through parsing experiments on Universal Dependencies treebanks (Nivre et al., 2016, UD), investigating the effects of two common or well-established practices that aim to facilitate replication. Namely, fixing random seed of the pseudo random number generator used during experiments, and standard training, development and test set splits.

2 Experimental Setting

Our experiments consist of replicating and reproducing parsing scores reported with version 1.2 of UDPipe (Straka and Straková, 2017; Straka and Straková, 2019) on UD version 2.5 treebanks (Zeman et al., 2019). UDPipe is an open-source pipeline for tokenization, tagging, and dependency parsing designed particularly for the UD annotation scheme. The system obtained near-top results in the recent CoNLL shared tasks on dependency parsing from raw text (Zeman et al., 2017; Zeman et al., 2018). Besides its performance and the ease of use, another interesting aspect of UDPipe for the current study is the fact that maintainers provide pre-trained models for the majority of the UD treebanks, where each model is trained and tested on the standard UD splits. The authors also publish common performance metrics, and provide scripts for replication.[1]

Our experiments consist of (1) exact replication of the reported results using the same code and the standard splits, (2) reproducing the results by varying the random seed, and (3) reproducing the results on random treebank splits. To facilitate exact replication of the published scores, the UDPipe model distribution fixes the random seed used for initializing the model parameters and other sources of variation (e.g., shuffling of training instances). For our main experiments, the UDPipe is modified to set the random seed based on the time stamp (a reasonably random quantity for our purposes). We use the same embeddings distributed with the UDPipe models, and the same hyperparameter settings as the original models. The experiments are run using the scripts provided in the UDPipe 2.5 models package, with slight modifications for parametrizing the input, and parallelizing it with GNU parallel (Tange, 2011). We report results for all treebanks for which a UDPipe 2.5 model was published, except for the largest five treebanks.[2]

For the experiments with different data splits, we simply combine all the sentences from the standard splits, and create 10 versions of each treebank by randomly splitting the sentences into train, dev, and test sets of the same size as the original split. In all cases, since the parser does not use any extra-sentential information, we sample the sentences randomly, without paying attention to document boundaries even if they are marked. In the main text, we only present and compare labeled attachment scores (LAS). Other metrics, including tokenization and tagging scores, are presented in Appendix A.

3 Results

We first verify the exact replication of the published scores. Besides the fixed treebank splits, since original study eliminates all sources of random variation, our results exactly match theirs.

To establish the expected variation of the parser on the same treebank splits, we perform multiple experiments with standard splits after allowing random variation in model initialization. The black bars

[1] `http://ufal.mff.cuni.cz/udpipe/models`. The (modified) scripts used in this study are accessible at `https://github.com/coltekin/udpipe-reproduction`.

[2] For the sake of the impact on the environment as well as on the patience of the author. The experiments reported here require over 10 CPU-months on a relatively recent architecture, and 5 additional data points provide little additional support for present discussion, while extending the computational cost considerably.

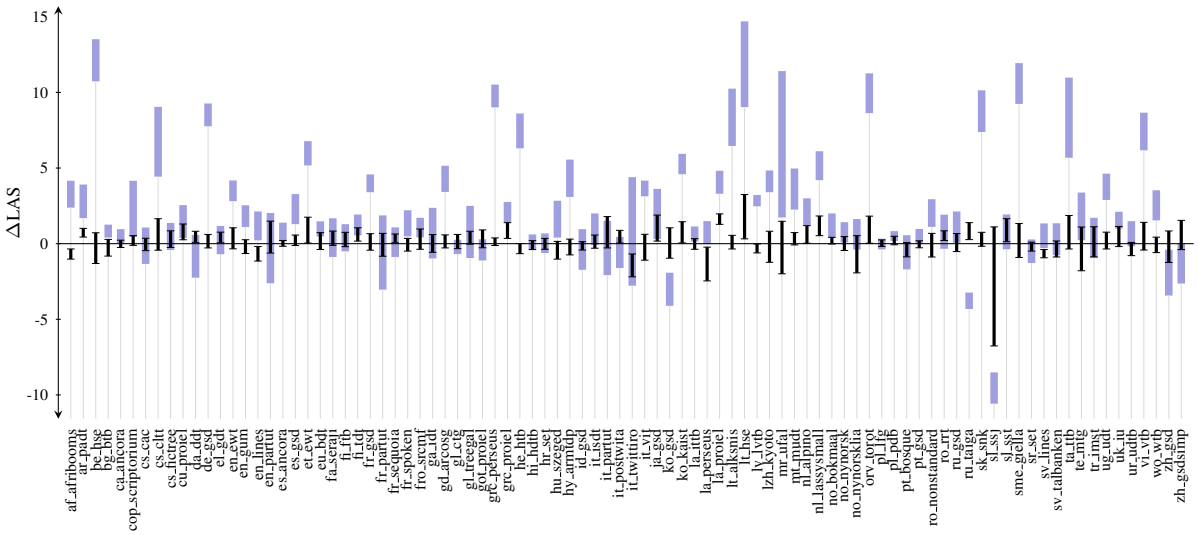

Figure 1: Percent LAS difference from published results and 10 experiments on standard splits (black bars) and 10 experiments with random train/dev/test splits (thick light blue bars). A positive value indicates a score higher than the published score. The bars indicate the complete range of scores.

in Figure 1 present the range of LAS obtained in 10 experiments on the standard splits (all LAS figures presented in this paper are percentages).[3] All scores in Figure 1 are relative to the scores published by the original study, a positive value indicating a score higher than the published score on a particular treebank. The scores do not diverge from the original study substantially. Nevertheless, there is considerable variation. The range of scores are between 0.40 and 7.89 LAS points with an average of 1.30 LAS. The range of LAS scores from 10 experiments with different random seeds do not contain the published result in 21 of the 90 experiments. It is also worth noting that in cases the reported value is within the range, it is not necessarily at the center of the range.

As expected, the range of scores obtained on random treebank splits (shown as light blue bars in Figure 1) indicate an even wider spread. The ranges vary between 0.65 and 9.68 LAS points, with an average of 2.25. The scores of random-split experiments also differ from the published results considerably. The LAS on random splits and the published results differ 2.72 points on average, with a maximum of 12.30. In the majority of the cases, the published results are not within the range of scores obtained on random-split experiments, only 33 out of 90 differences falling within this range. Most differences however, are positive. Minimum LAS obtained in random-split experiments are higher than the published results for 53 treebanks, indicating that standard splits underestimate the success of the parser. This is expected, since most treebanks do not sample test sets randomly, conflicting with the i.i.d. assumption of many machine learning models. The published results are above the random-split range only for 4 treebanks, probably due to the official splits with easier to parse test sets.

4 Discussion

The experiments presented above demonstrate the effects of two sources of variation on parsing scores. The variation displayed in Figure 1 is rarely reported in the literature. Furthermore, common practices promoted for enabling exact replication leads researchers to eliminate this variation from their experiments, reporting single-figure scores without an indication of the expected variation. Such single-figure scores, as demonstrated above, are often biased in an arbitrary direction.

[3]The reason for presenting the range is not to amplify the readers' perception of the variation. As an anonymous reviewer pointed out, the standard deviation or the standard error are better, more robust measures of variation from the mean. A study reporting a mean score with its variation over *multiple* experiments should report such a robust measure of variation. However, since the point in question here is the practice of reporting a *single* score from a single model instance, the demonstration of the range is more interesting for our purposes.

The first type of variation we discuss is due to model training and evaluation, e.g., because of random initialization of the model parameters, or random ordering of the data before online or batch training. Any non-trivial statistical model produces different results when trained and tested on the same data. This is typically negligible for convex optimization procedures.[4] The variation is naturally higher for models trained with non-convex optimization, such as neural networks which are the dominant approach to parsing as well as many other NLP tasks. The variation stemming from these two sources is rarely reported in the literature. On the contrary, due to heavy emphasis on replicability in the field (Pedersen, 2008; Wieling et al., 2018, for example), when researchers release their software, they often fix the variation in an arbitrary way, e.g., by fixing the initial random seed used by pseudo random number generators as in our case study above. The point here is not to promote systems with large variation. We want to *reduce* the variation when possible. For example, ensembling or model averaging are useful for reducing the variation. However, there will still be some variation based on initialization differences and sampling, and reporting this variation is useful for understanding the success of the model, as well as for meaningful model comparison.

Our experiments indicate the range of LAS obtained by UDPipe on UD treebanks over 10 random initializations is 1.30 on average, and goes up to 7.89 on individual treebanks. To facilitate a rough comparison, the LAS (averaged over all treebanks) differences between UDPipe and the systems above and below UDPipe in CoNLL 2018 shared task are less than 0.20, indicating that the arbitrary choice of random seed could have made the difference in this ranking.

The solution to this problem is clear. Rather than 'hiding' the variation by settling on a fixed model instance, we should report the variation caused by the inevitable sources of variability. As the quote from the review form in Section 1 demonstrate, reporting results with sampling variation is perceived as a negative property of a paper submitted to many computational linguistics conferences. Hence, to be able to increase chances of acceptance of a paper, the authors are encouraged to fix all sources of (natural) variation of models in an arbitrary manner. The present paper, hopefully, makes it clear that this is wrong. However, since the incorrect behavior is currently being promoted by the community at large. The real solution lies in further discussion of what should really be expected from individual experimental studies.

The second issue we discuss is the adverse effects of standard training, development and test set splits, which has also been brought up by others before (Gorman and Bedrick, 2019). The standard splits are common among all NLP data sets, but tradition is probably older for treebanks, going back to early parsing research.[5] The recent treebanking efforts, e.g., UD treebanks, continued this tradition with official splits of training, development, and test sets. The standard test sets are often justified by replicability of the results obtained on these data sets, and *fair* comparisons between different systems. Another motivation for test split is to avoid 'peeking into test data' (Jurafsky and Martin, 2009, p.189). However, the latter goal is difficult to entertain for a standard test that remains fixed for decades.

Independent of the motivation of the standard splits, the experiments reported above show that the use of standard treebank splits hides a large amount of variation, and it often biases the parsing results arbitrarily. When the scores are calculated over multiple random splits, the variation can be as large as 9.68 LAS and the average scores obtained can differ from the reported results up to 12.30 LAS. Furthermore, for the majority of the treebanks, scores reported in the original study are not in the range of scores obtained in random-split experiments. Clearly, these differences blur model comparisons. More importantly, in any practical use of these tools, the user should be aware of the expected variation. Fortunately, in our experiments, most of the results would come as a pleasant surprise, since we obtained ranges of scores better than the original report. This also means the standard test set splits are biased in certain ways, and, hence, model comparison on these sets may arbitrarily favor certain types of parsers.

The solution is, again, obvious: reporting results (with their variation) on different data splits. In other words, abandoning standard splits. However, this goes against the well-established traditions. Again, true solution require more discussion of the issues in the research community. Many data sets include

[4]The most common reason for a potential variance is the convergence criterion based on a fixed threshold. Different initializations may result in stopping at (slightly) different points on the error surface close to the minimum.

[5]As of this writing, a search on ACL anthology for keywords "Penn Treebank" and "section 23" returns over 1000 hits. Many models, most of them parsers, have been compared on this de facto standard test set.

a standard training, development and test set split. As a result, for fair comparison, researchers are encouraged, or sometimes required to report results on a single standard test set. For example, it is unlikely for a paper on parsing to be accepted if it reports results using non-standard splits. Hence, similar to reporting sampling variation, individual authors are forced to follow current practices. The real solution requires community-wide changes to acceptable practices.

The issues discussed above are orthogonal to the issue addressed by statistical significance testing. Although statistical assurances for significant differences have their merits, it is important to note that they indicate statistically significant differences on a *single test set*. Hence, in their typical application, they may even contribute to a false sense of significant difference that disappears when the systems are tested on different test sets, or even the same model trained with different initializations. Interestingly enough, the differences between the minimum and the maximum LAS scores in our experiments conducted on the standard test sets with varying random seeds are statistically significant for all treebanks.[6] Hence, a statistically significant finding reported, e.g., for two alternative training settings, may very well be due to the random initialization at least with a $10\,\%$ chance according to our experiments. The use of above-chance statistical differences are only helpful when the systems are tested on multiple data sets (Dror et al., 2017). However, in most tasks, multiple data sets are not a luxury we enjoy, especially when the research or application at hand is in a particular domain. Using multiple random splits of available data or use of cross-validation is often the closest we can get to a general solution.

Hidden in the issues related to variation with different data splits is the size and quality of the data sets. For many languages or tasks, the data sets are small and sometimes low quality, which eventually affect the models trained and tested on them. Creating, annotating and curating data sets is a labor-intensive, expensive, yet error-prone task. On the other hand, creating more data sets for which there are already existing resources, and improving the existing data sets are not rewarding tasks. Again, to be able to test our models on multiple and better data sets, we, as a community, need ways to encourage and reward creation and maintenance of more and high-quality data sets.

Although our case study is based on parsing results with UDPipe, the problems outlined are neither specific to UDPipe nor parsing. In any statistical, data-driven NLP experiment, some amount of variation is inevitable in experimental results, and the standard test set splits and avoiding variation due to other factors in model training hides this variation.

5 Concluding remarks

This paper presented a demonstration of adverse effects of avoiding variation in the results of NLP experiments. We demonstrate the effects of two sources of variation in NLP experiments: the variation due to sampling and initialization. For sampling variation, a well-established practice in the literature is to use a fixed standard training–test split across different studies. The only clear advantage of using a single standard test set is allowing fair comparisons between different studies. A *fair* comparison is definitely desirable in a competition, but for research we should perhaps emphasize *meaningful* comparisons more. The results presented above demonstrate that comparisons on a single (standard) tests set are not necessarily meaningful. Another, less-established but still common practice is to fix the random seed of the pseudo random number generator used during the experiments. We show that both practices hide the inherent variation in parsing scores, and bias the published results in arbitrary ways. As a result, these practices also defeat their very own purpose, making it difficult to interpret the comparison between different models outside this fixed setting.

The purpose of this paper is to point out the problem and promote further discussion of the topic. Our outcomes indicate that for insightful evaluations, we should acknowledge and report the variation in experimental results, rather than fixing the variation in arbitrary ways. This includes, (1) letting the model initializations vary, and (2) running the experiments on multiple data splits rather than a single training/development/test split. For both cases, reporting the variation based on multiple experiments. However, the implementation of the solutions is not easy. It requires community-wide changes beyond the power of individual researchers.

[6] At $p < 0.05$ calculated based on $10\,000$ bootstrap samples.

References

António Branco, Kevin Bretonnel Cohen, Piek Vossen, Nancy Ide, Nicoletta Calzolari, et al. 2017. Replicability and reproducibility of research results for human language technology: introducing an LRE special section. *Language Resources and Evaluation*, 51(1):221–247.

António Branco, Nicoletta Calzolari, Piek Vossen, Gertjan Van Noord, Dieter van Uytvanck, João Silva, Luís Gomes, André Moreira, and Willem Elbers. 2020. A shared task of a new, collaborative type to foster reproducibility: A first exercise in the area of language science and technology with reprolang2020. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 5539–5545, Marseille, France, May. European Language Resources Association.

Kevin Cohen, Aurélie Névéol, Jingbo Xia, Negacy Hailu, Lawrence Hunter, and Pierre Zweigenbaum. 2017. Reproducibility in Biomedical Natural Language Processing. In *AMIA Annual Symposium*, Washington, DC, United States, November.

K. Bretonnel Cohen, Jingbo Xia, Pierre Zweigenbaum, Tiffany Callahan, Orin Hargraves, Foster Goss, Nancy Ide, Aurélie Névéol, Cyril Grouin, and Lawrence E. Hunter. 2018. Three dimensions of reproducibility in natural language processing. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Rotem Dror, Gili Baumer, Marina Bogomolov, and Roi Reichart. 2017. Replicability analysis for natural language processing: Testing significance with multiple datasets. *Transactions of the Association for Computational Linguistics*, 5:471–486.

Chris Drummond. 2009. Replicability is not reproducibility: nor is it good science. In *Proceedings of the Evaluation Methods for Machine Learning Workshop at the 26th ICML*, pages 14–18.

Fiona Fidler and John Wilcox. 2018. Reproducibility of scientific results. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, winter 2018 edition.

Antske Fokkens, Marieke van Erp, Marten Postma, Ted Pedersen, Piek Vossen, and Nuno Freire. 2013. Offspring from reproduction problems: What replication failure teaches us. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1691–1701, Sofia, Bulgaria, August. Association for Computational Linguistics.

Leonard P. Freedman, Iain M. Cockburn, and Timothy S. Simcoe. 2015. The economics of reproducibility in preclinical research. *PLOS Biology*, 13(6):1–9, 06.

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, July. Association for Computational Linguistics.

Eva Huber and Çağrı Çöltekin. 2020. Reproduction and replication: A case study with automatic essay scoring. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 5603–5613, Marseille, France.

Daniel Jurafsky and James H. Martin. 2009. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Pearson Prentice Hall, second edition.

Margot Mieskes. 2017. A quantitative study of data in the NLP community. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 23–29, Valencia, Spain, April. Association for Computational Linguistics.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 23–28.

Open Science Collaboration. 2015. Estimating the reproducibility of psychological science. *Science*, (349):943–951.

Ted Pedersen. 2008. Empiricism is not a matter of faith. *Computational Linguistics*, 34(3):465–470.

Nils Reimers and Iryna Gurevych. 2017. Reporting score distributions makes a difference: Performance study of LSTM-networks for sequence tagging. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 338–348, Copenhagen, Denmark, September. Association for Computational Linguistics.

Milan Straka and Jana Straková. 2017. Tokenizing, pos tagging, lemmatizing and parsing UD 2.0 with UDPipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99, Vancouver, Canada, August. Association for Computational Linguistics.

Milan Straka and Jana Straková. 2019. Universal dependencies 2.5 models for UDPipe (2019-12-06). LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Ole Tange. 2011. GNU parallel - the command-line power tool. *;login: The USENIX Magazine*, 36(2):42–47.

Martijn Wieling, Josine Rawee, and Gertjan van Noord. 2018. Reproducibility in computational linguistics: Are we willing to share? *Computational Linguistics*, 44(4):641–649, December.

Daniel Zeman, Martin Popel, Milan Straka, Jan Hajic, Joakim Nivre, Filip Ginter, Juhani Luotolahti, Sampo Pyysalo, Slav Petrov, Martin Potthast, Francis Tyers, Elena Badmaeva, Memduh Gokirmak, Anna Nedoluzhko, Silvie Cinkova, Jan Hajic jr., Jaroslava Hlavacova, Václava Kettnerová, Zdenka Uresova, 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 Mendonca, 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, August. 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, October. Association for Computational Linguistics.

Daniel Zeman, Joakim Nivre, Mitchell Abrams, Noëmi Aepli, Željko Agić, Lars Ahrenberg, Gabrielė Aleksandravičiūtė, Lene Antonsen, Katya Aplonova, Maria Jesus Aranzabe, Gashaw Arutie, Masayuki Asahara, Luma Ateyah, Mohammed Attia, Aitziber Atutxa, Liesbeth Augustinus, Elena Badmaeva, Miguel Ballesteros, Esha Banerjee, Sebastian Bank, Verginica Barbu Mititelu, Victoria Basmov, Colin Batchelor, John Bauer, Sandra Bellato, Kepa Bengoetxea, Yevgeni Berzak, Irshad Ahmad Bhat, Riyaz Ahmad Bhat, Erica Biagetti, Eckhard Bick, Agnė Bielinskienė, Rogier Blokland, Victoria Bobicev, Loïc Boizou, Emanuel Borges Völker, Carl Börstell, Cristina Bosco, Gosse Bouma, Sam Bowman, Adriane Boyd, Kristina Brokaitė, Aljoscha Burchardt, Marie Candito, Bernard Caron, Gauthier Caron, Tatiana Cavalcanti, Gülşen Cebiroğlu Eryiğit, Flavio Massimiliano Cecchini, Giuseppe G. A. Celano, Slavomír Čéplö, Savas Cetin, Fabricio Chalub, Jinho Choi, Yongseok Cho, Jayeol Chun, Alessandra T. Cignarella, Silvie Cinková, Aurélie Collomb, Çağrı Çöltekin, Miriam Connor, Marine Courtin, Elizabeth Davidson, Marie-Catherine de Marneffe, Valeria de Paiva, Elvis de Souza, Arantza Diaz de Ilarraza, Carly Dickerson, Bamba Dione, Peter Dirix, Kaja Dobrovoljc, Timothy Dozat, Kira Droganova, Puneet Dwivedi, Hanne Eckhoff, Marhaba Eli, Ali Elkahky, Binyam Ephrem, Olga Erina, Tomaž Erjavec, Aline Etienne, Wograine Evelyn, Richárd Farkas, Hector Fernandez Alcalde, Jennifer Foster, Cláudia Freitas, Kazunori Fujita, Katarína Gajdošová, Daniel Galbraith, Marcos Garcia, Moa Gärdenfors, Sebastian Garza, Kim Gerdes, Filip Ginter, Iakes Goenaga, Koldo Gojenola, Memduh Gökırmak, Yoav Goldberg, Xavier Gómez Guinovart, Berta González Saavedra, Bernadeta Griciūtė, Matias Grioni, Normunds Grūzītis, Bruno Guillaume, Céline Guillot-Barbance, Nizar Habash, Jan Hajič, Jan Hajič jr., Mika Hämäläinen, Linh Hà Mỹ, Na-Rae Han, Kim Harris, Dag Haug, Johannes Heinecke, Felix Hennig, Barbora Hladká, Jaroslava Hlaváčová, Florinel Hociung, Petter Hohle, Jena Hwang, Takumi Ikeda, Radu Ion, Elena Irimia, Ọlájídé Ishola, Tomáš Jelínek, Anders Johannsen, Fredrik Jørgensen, Markus Juutinen, Hüner Kaşıkara, Andre Kaasen, Nadezhda Kabaeva, Sylvain Kahane, Hiroshi Kanayama, Jenna Kanerva, Boris Katz, Tolga Kayadelen, Jessica Kenney, Václava Kettnerová, Jesse Kirchner, Elena Klementieva, Arne Köhn, Kamil Kopacewicz, Natalia Kotsyba, Jolanta Kovalevskaitė, Simon Krek, Sookyoung Kwak, Veronika Laippala, Lorenzo Lambertino, Lucia Lam, Tatiana Lando, Septina Dian Larasati, Alexei Lavrentiev, John Lee, Phuong Lê H`ông, Alessandro Lenci, Saran Lertpradit, Herman Leung, Cheuk Ying Li, Josie Li, Keying Li, KyungTae Lim, Maria Liovina, Yuan Li, Nikola Ljubešić, Olga Loginova, Olga Lyashevskaya, Teresa Lynn, Vivien Macketanz, Aibek Makazhanov, Michael Mandl, Christopher Manning, Ruli Manurung, Cătălina Mărănduc, David Mareček, Katrin Marheinecke, Héctor Martínez Alonso, André Martins, Jan Mašek, Yuji Matsumoto, Ryan McDonald, Sarah McGuinness, Gustavo Mendonça, Niko Miekka, Margarita Misirpashayeva, Anna Missilä, Cătălin Mititelu, Maria Mitrofan, Yusuke Miyao, Simonetta Montemagni, Amir More, Laura Moreno Romero, Keiko Sophie Mori, Tomohiko Morioka, Shinsuke Mori, Shigeki Moro, Bjartur Mortensen, Bohdan Moskalevskyi, Kadri Muischnek, Robert Munro, Yugo Murawaki, Kaili Müürisep, Pinkey Nainwani, Juan Ignacio Navarro Horñiacek, Anna Nedoluzhko, Gunta Nešpore-Bērzkalne,

Luong Nguy˜ên Thị, Huy`ên Nguy˜ên Thị Minh, Yoshihiro Nikaido, Vitaly Nikolaev, Rattima Nitisaroj, Hanna Nurmi, Stina Ojala, Atul Kr. Ojha, Adédayọ̀ Olúòkun, Mai Omura, Petya Osenova, Robert Östling, Lilja Øvrelid, Niko Partanen, Elena Pascual, Marco Passarotti, Agnieszka Patejuk, Guilherme Paulino-Passos, Angelika Peljak-Łapińska, Siyao Peng, Cenel-Augusto Perez, Guy Perrier, Daria Petrova, Slav Petrov, 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, Loganathan Ramasamy, Taraka Rama, Carlos Ramisch, Vinit Ravishankar, Livy Real, Siva Reddy, Georg Rehm, Ivan Riabov, Michael Rießler, Erika Rimkutė, Larissa Rinaldi, Laura Rituma, Luisa Rocha, Mykhailo Romanenko, Rudolf Rosa, Davide Rovati, Valentin Roșca, Olga Rudina, Jack Rueter, Shoval Sadde, Benoît Sagot, Shadi Saleh, Alessio Salomoni, Tanja Samardžić, Stephanie Samson, Manuela Sanguinetti, Dage Särg, Baiba Saulīte, Yanin Sawanakunanon, Nathan Schneider, Sebastian Schuster, Djamé Seddah, Wolfgang Seeker, Mojgan Seraji, Mo Shen, Atsuko Shimada, Hiroyuki Shirasu, Muh Shohibussirri, Dmitry Sichinava, Aline Silveira, Natalia Silveira, Maria Simi, Radu Simionescu, Katalin Simkó, Mária Šimková, Kiril Simov, Aaron Smith, Isabela Soares-Bastos, Carolyn Spadine, Antonio Stella, Milan Straka, Jana Strnadová, Alane Suhr, Umut Sulubacak, Shingo Suzuki, Zsolt Szántó, Dima Taji, Yuta Takahashi, Fabio Tamburini, Takaaki Tanaka, Isabelle Tellier, Guillaume Thomas, Liisi Torga, Trond Trosterud, Anna Trukhina, Reut Tsarfaty, Francis Tyers, Sumire Uematsu, Zdeňka Urešová, Larraitz Uria, Hans Uszkoreit, Andrius Utka, Sowmya Vajjala, Daniel van Niekerk, Gertjan van Noord, Viktor Varga, Eric Villemonte de la Clergerie, Veronika Vincze, Lars Wallin, Abigail Walsh, Jing Xian Wang, Jonathan North Washington, Maximilan Wendt, Seyi Williams, Mats Wirén, Christian Wittern, Tsegay Woldemariam, Tak-sum Wong, Alina Wróblewska, Mary Yako, Naoki Yamazaki, Chunxiao Yan, Koichi Yasuoka, Marat M. Yavrumyan, Zhuoran Yu, Zdeněk Žabokrtský, Amir Zeldes, Manying Zhang, and Hanzhi Zhu. 2019. Universal dependencies 2.5. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

A Further comparisons

This section presents additional scores obtained in our experiments in all stages of the parsing pipeline. Unlike Figure 1 in the main text, the figures presented in this section present absolute values. The scores reported by the original study are indicated by a triangle or diamond on the figures. For ease of interpretation, a blue triangle pointing up indicates that the random-split experiments yielding larger scores than the original report, a orange triangle pointing down indicates random-split experiments that yield results lower than the original report, and a black diamond shape indicates that the original scores fall within the range of scores in random-sample experiments. Similar to Figure 1, the black bars indicate the experiments with standard splits, while the light blue bars represent experiments with random splits.

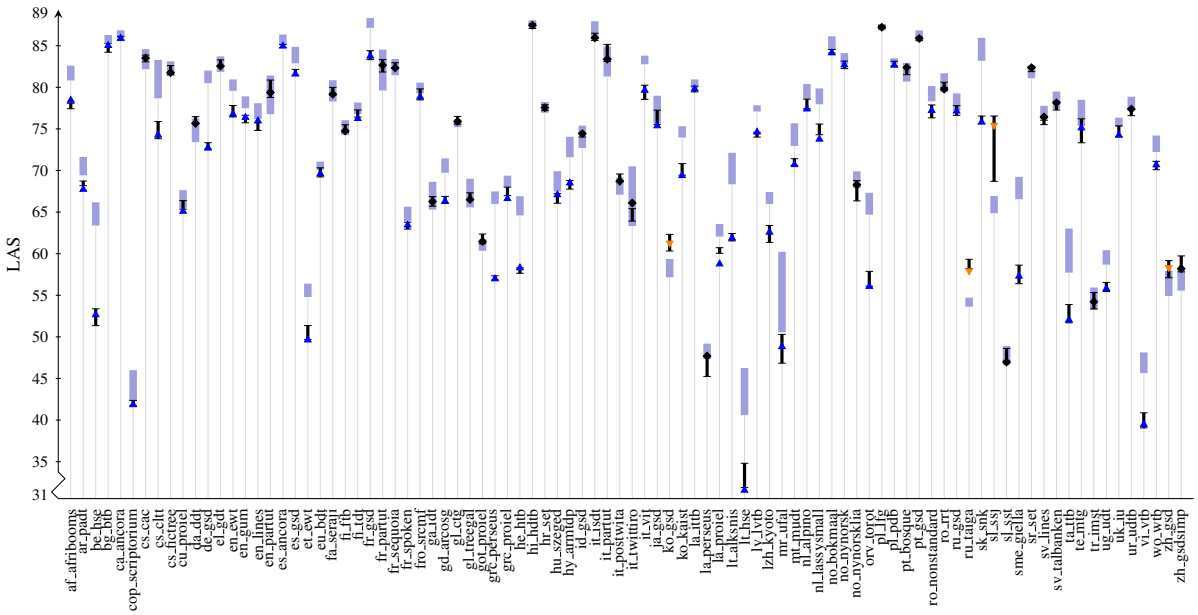

Figure 2: Absolute LAS scores. This figure present the same data in Figure 1, but in an absolute scale rather than relative to the scores published in the original study.

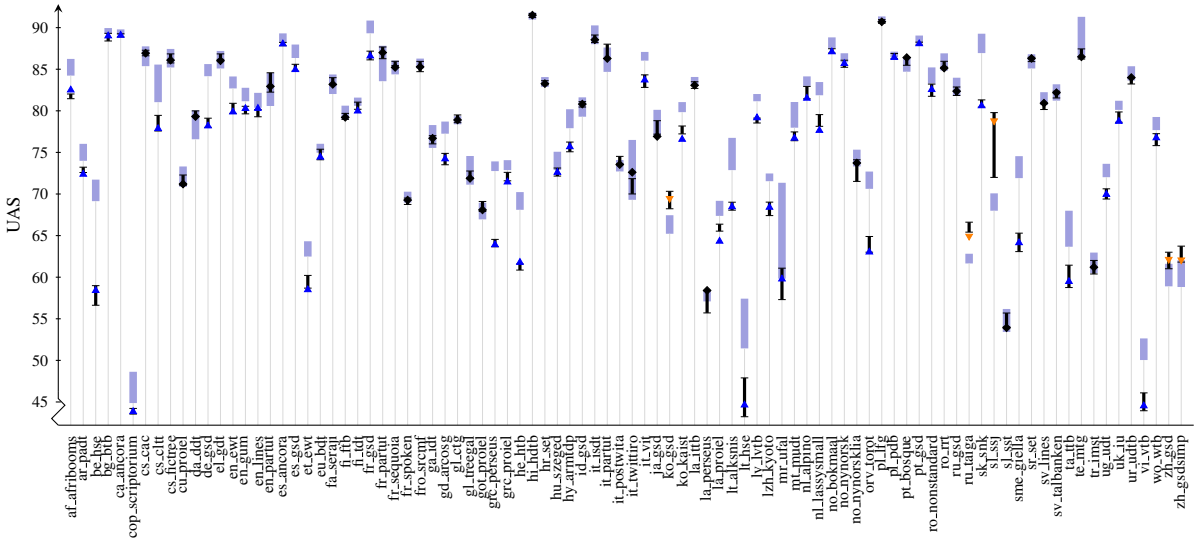

Figure 3: UAS scores.

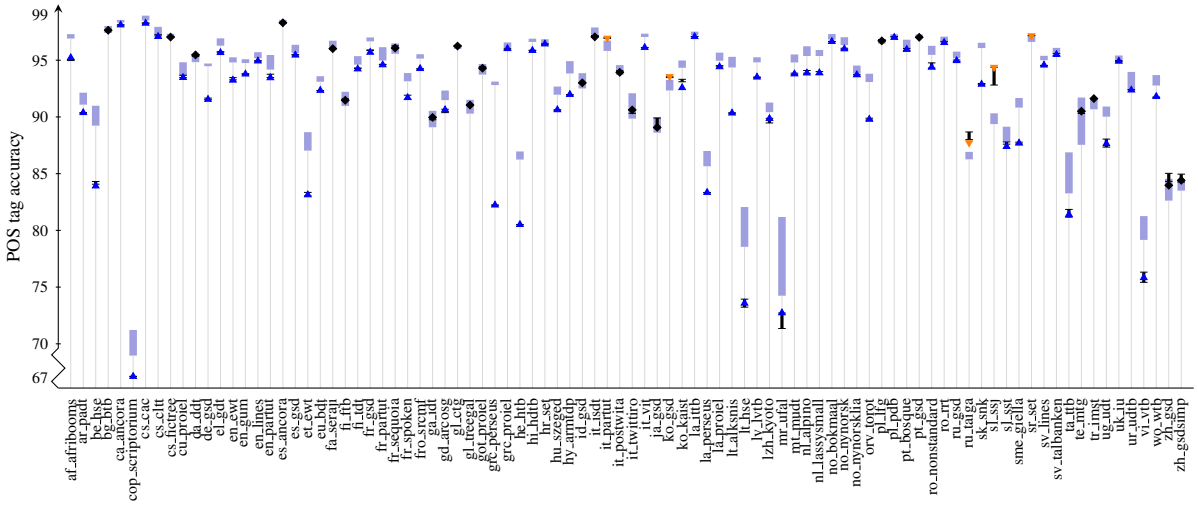

Figure 4: POS tagging accuracy.

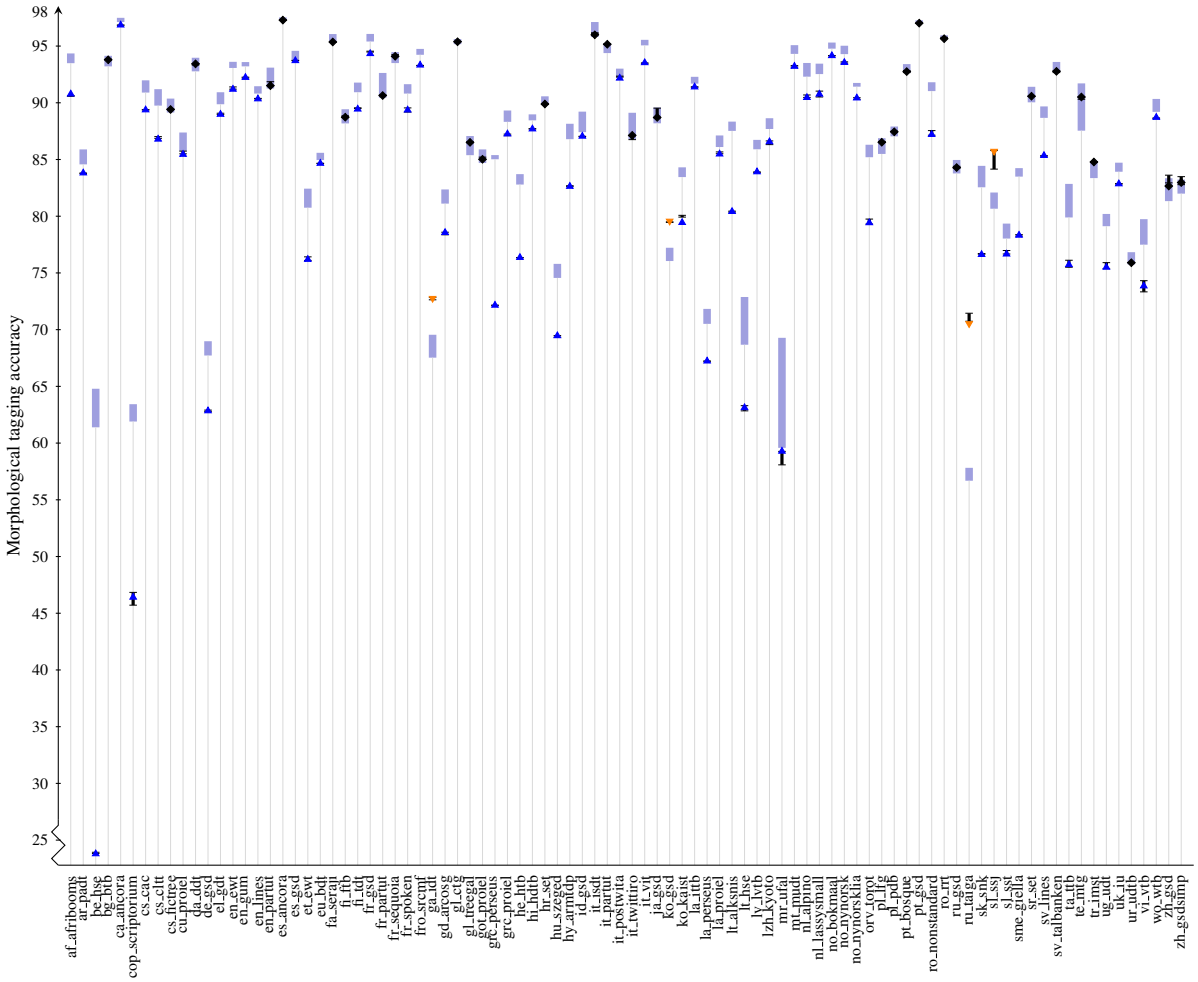

Figure 5: Full morphological tagging accuracy (including POS tags).

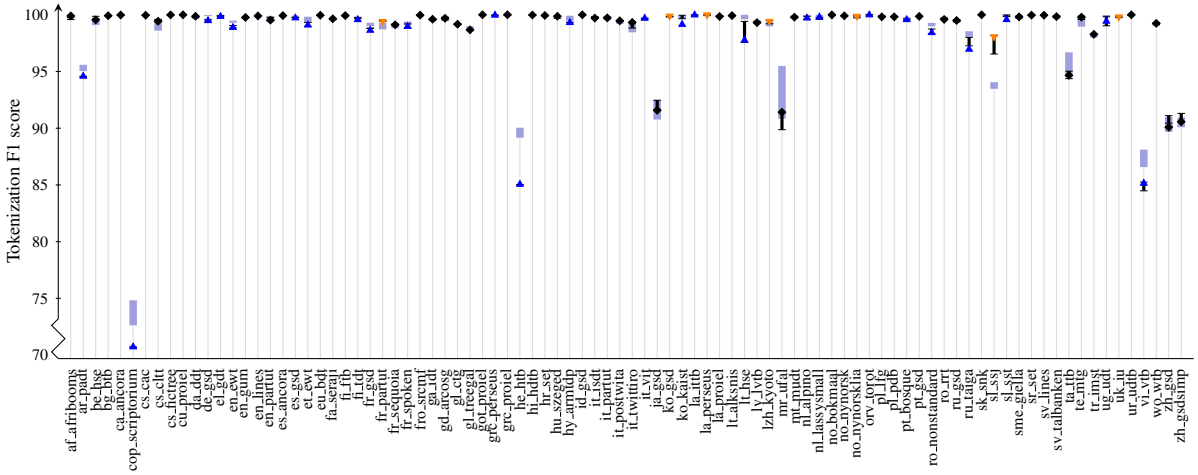

Figure 6: Tokenization F1 score.

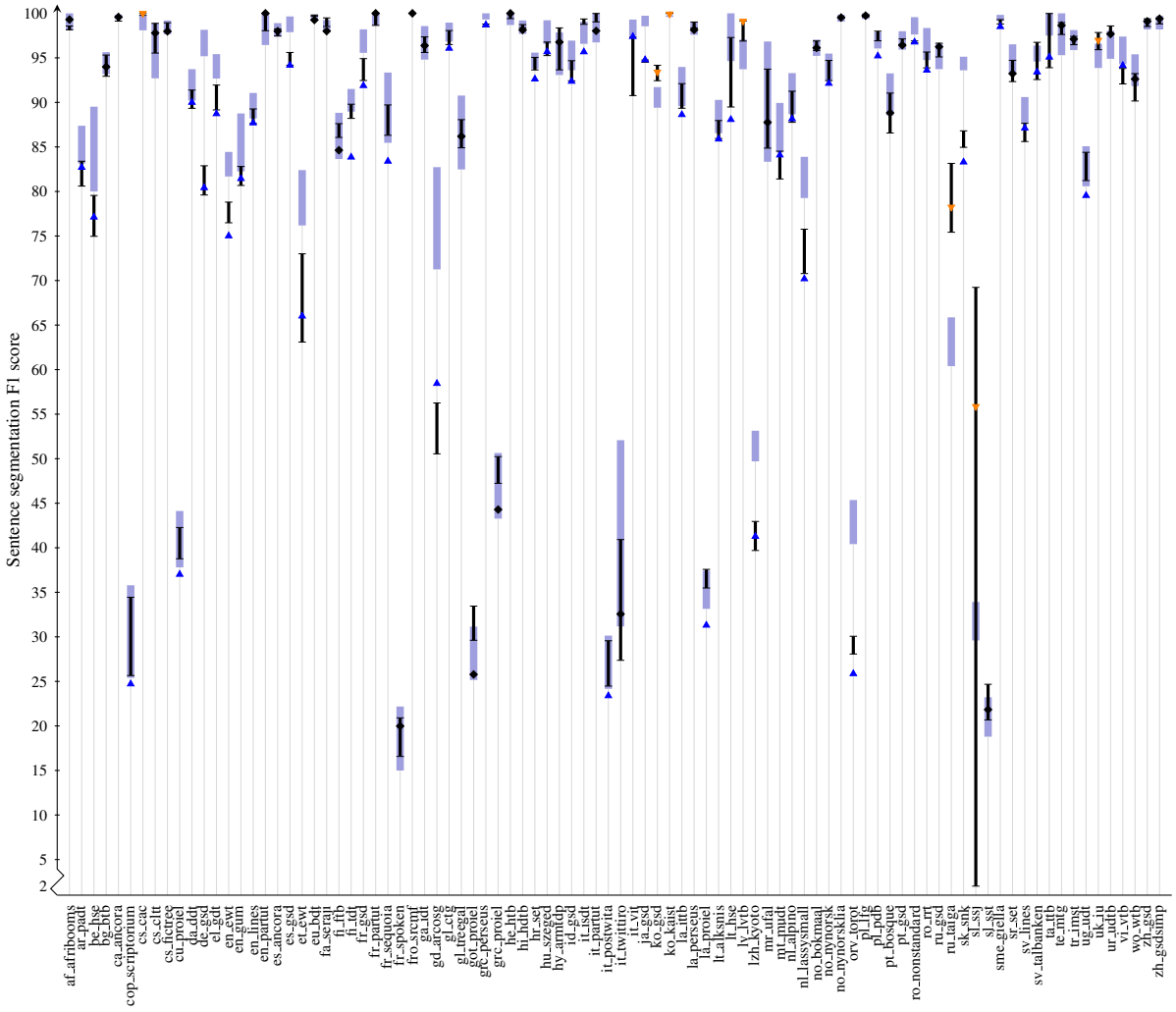

Figure 7: Sentence splitting F1 score.