Cross-linguistic complexity analysis using UD treebanks

Anonymous ACL submission

Abstract

The aim of this project is to explore the use of UD treebanks in cross-linguistic complexity research. While it is not this paper's objective to make strong theoretical claims on complexity in its own right, it examines complexity measures computed on various UD treebanks across 4 languages. This work focuses on treebanks for English, German, Hungarian, and Chinese, in an attempt to determine whether complexity measures stay consistent within treebanks for the same language, and whether some languages consistently show higher complexity than others. These claims are verified with official UD corpora, as well as with a parallel corpus of 50 sentences for each language mentioned above.

1 Introduction

What constitutes linguistic complexity is an elusive multi-faceted inquiry, and quantifying it is equally challenging. Previous approaches have included text-based metrics as well as eye-tracking studies (e.g. Gordon et al., 2007), but a general consensus has not been reached, especially when it comes to the notion of "overall complexity" of a language. In the case that an experiment produces numeric values as a proxy for complexity, it is not obvious that results computed on two different languages are comparable.

In this work, the objective is not to attempt to resolve these foundational question about linguistic complexity, but rather to accept that there are certain well documented text-based proxies for computing it - such as lexical, morphological, and syntactic complexity measures - and to explore how the Universal Dependencies framework can be employed to further the subject matter.

Indeed, the potential of UD-annotated data appears to be largely untapped in linguistic complexity research. A convincing contribution to this field was made by Berdicevskis et al., 2018, estimating robustness of complexity values computed on a variety of treebanks, while accounting for additional obstacles such as language-specific annotation conventions that might reduce comparability.

While this project is more restricted in scope and claims, the goal is to showcase the use of UD treebanks in linguistic complexity research along the same lines as Berdicevskis et al., 2018. For this purpose, we use a variety of treebanks across four languages and a set of text-based complexity metrics inspired in part by readability research, and in part by complexity research more broadly. These metrics are computed on two datasets, one of which is a collection of official UD treebanks, and the other being a small in-house parallel corpus comprising 50 sentences of each of the 4 target languages.

2 Dataset

Table 1 shows the list of UD web treebanks used in this experiment. For the English language, which has many treebanks available to it, we selected UD English-ESL/TLE (Berzak et al., 2016, Yannakoudakis et al., 2011, GitHub[1]), GUM (Berzak et al., 2016, GitHub[2]), and EWT (Silveira et al., 2014, GitHub[3]). For German, we opted for GSD (McDonald et al., 2013, GitHub [4]) and HDT (Borges Völker et al., 2019, Hennig and Köhn, 2017, Foth et al., 2014, Foth, 2006, GitHub [5]).

[1] https://github.com/UniversalDependencies/UD_English-ESL

[2] https://github.com/UniversalDependencies/UD_English-GUM

[3] https://github.com/UniversalDependencies/UD_English-EWT

[4] https://github.com/UniversalDependencies/UD_German-GSD

[5] https://github.com/UniversalDependencies/UD_German-HDT

Treebank	Language	N. sentences	Abbreviation
UD English-ESL/TLE	English	4124	en_esl
UD English-GUM	English	6911	en_gum
UD English-EWT	English	12544	en_ewt
UD German-GSD	German	13814	de_gsd
UD German-HDT	German	75617	de_hdt
UD Chinese-GSD	Chinese	3997	cmn_gsd
UD Chinese-PUD	Chinese	1000	cmn_pud
UD Hungarian-Szeged	Hungarian	455, 455	hun_szeged_1, hun_szeged_2

Table 1: UD web treebanks and sources

For Chinese, GSD (GitHub[6]) and PUD (GitHub[7]). For Hungarian, only the Szeged treebank (Vincze et al., 2010, Github[8]) was available - therefore in order to have at least two treebanks for each language, the sentences annotated in the Szeged treebank were randomly distributed across two new treebanks (hun_szeged_1, hun_szeged_2).

An additional parallel treebank[9] made by Nino Meisinger, Qin Gu, Lisa Wang, and Aron Winkler as coursework at the University of Tübingen was also employed. Although its limited size of only 50 sentences per language prevents meaningful conclusions from being drawn from measurements computed on it, its parallel nature nevertheless allows interesting observations to be made in reference to the complexity metric values sourced from the official treebanks.

References

Aleksandrs Berdicevskis, Çağrı Çöltekin, Katharina Ehret, Kilu von Prince, Daniel Ross, Bill Thompson, Chunxiao Yan, Vera Demberg, Gary Lupyan, Taraka Rama, and Christian Bentz. 2018. Using universal dependencies in cross-linguistic complexity research. *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 8–17.

Yevgeni Berzak, Jessica Kenney, Carolyn Spadine, Jing Xian Wang, Lucia Lam, Keiko Sophie Mori, Sebastian Garza, and Boris Katz. 2016. Universal dependencies for learner english. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 737–746. 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.

Kilian Foth, Arne Köhn, Niels Beuck, and Wolfgang Menzel. 2014. Because size does matter: The hamburg dependency treebank. In *Proceedings of the Language Resources and Evaluation Conference 2014*, pages 2326–2333. Reykjavik, Iceland: European Language Resources Association (ELRA).

Kilian A. Foth. 2006. Eine umfassende constraint-dependenz-grammatik des deutschen. *Universität Hamburg*.

Peter C. Gordon, Hanjung Lee, and Yoonhyoung Lee. 2007. Linguistic complexity and information structure in korean: Evidence from eye-tracking during reading. *Cognition*, 104(3):495–534.

Felix Hennig and Arne Köhn. 2017. Dependency tree transformation with tree transducers. In *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*, pages 58–66, Gothenburg, Sweden. Association for Computational Linguistics.

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

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

Veronika Vincze, Dóra Szauter, Attila Almási, György Móra, Zoltán Alexin, and János Csirik. 2010. Hungarian dependency treebank. In *Proceedings of the*

[6] https://github.com/UniversalDependencies/UD_Chinese-GSD

[7] https://github.com/UniversalDependencies/UD_Chinese-PUD

[8] https://github.com/UniversalDependencies/UD_Hungarian-Szeged

[9] https://github.com/iscl-dtdp/ParallelTreebank-FinalProject

Seventh Conference on International Language Resources and Evaluation (LREC'10).

Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. 2011. A new dataset and method for automatically grading esol texts. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 180–189. Association for Computational Linguistics.

A Example Appendix

This is an appendix.