Cross-linguistic complexity analysis using UD treebanks

Aron Winkler
University of Tübungen
email@domain

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, a variety of treebanks across four languages were used, in combination with 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 and metrics

2.1 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, the Atis treebank (Github[1]), the gum GUM (Berzak et al., 2016, GitHub[2]), and the EWT treebank (Silveira et al., 2014, GitHub[3]) were ised. For German, this work employs 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]). For Chinese, GSD (GitHub[6]) and

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

Treebank	Language	N. sentences	Abbreviation
UD English-Atis	English	4274	en_atis
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

PUD (GitHub[7]) were the final choice. 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 obtained from the official treebanks. Sentences for this parallel treebank were sourced from Tatoeba[10], more specifically by selecting the 50 longest sentences that had translations for all 4 target languages. If more than one translation was available for any given language, then the choice was left up to the annotator. The final treebank was then manually annotated according to UD standards by annotators native or very proficient in the target language.

2.2 Metrics

Table 2 details the complexity metrics used in this experiment. Most of them were inspired by readability research, while a couple were adopted from (Berdicevskis et al., 2018). Sentence-level metrics (labelled as "sentence" in the "Level" column in the table) were computed at the sentence level, then averaged for the treebank. Treebank-level metrics (labelled as "treebank" in the "Level" column in the table), were calculated regardless of sentence boundaries.

A balance between morphological metrics (e.g. number of word forms per lemma) and syntactic metrics (e.g. number of clauses per sentence) was one of the objectives of this selection. A number of primitive lexical complexity metrics (e.g. type token ratio, token count) were also included. In this sense, noun to verb ratio is commonly used in readability research, as a preponderance of verbs signals more complex sentence. A similar case can be made for the number of clauses per sentence, as a sentence with many subordinate clauses will be naturally harder to process than a singular main clause.

We discuss also a pair of metrics not commonly seen in research and only enabled by the presence of dependency annotation, namely number of *ccomp* and *xcomp* relations in a sentence respectively. The intuition behind these metrics is that processing *ccomp* as opposed to *xcomp* relations would entail a lower cognitive load. Thus, a higher usage of *xcomp* might suggest higher overall complexity. Length of longest dependecy link is likewise enabled by the nature of the annotation, and gives an idea of how close related elements are on the surface level - academic Hungarian, as an example, tends do displace items across long sequences of clauses, often separating elements of the main clause by multiple lines of text.

Although there are more sophisticated metrics on the market, this set was ultimately selected for ease of calculation and because their interpretation is generally approachable without further analytics, which woudln't necesserily be the case for more nuanced strategies.

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

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

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

[10] https://tatoeba.org/en/

Metric	Description	Abbreviation	Level
Token Count	Token length of sentence	*tc*	sentence
Type Token Ratio	Simple type token ratio, without preprocessing	*ttr*	sentence
Parse Tree Depth	Depth of the parse tree obtained from the sentence	*ptd*	sentence
Length of longest dependency link	Surface distance in tokens between the most distant related elements in a sentence	*lldl*	sentence
Noun to verb ratio	Number of nouns divided by the number of verbs in a sentence	*n2v*	sentence
Number of clauses per sentence	Number of clauses per sentence - clauses are identified by the presence of verbs	*cxc*	sentence
Number of xcomp relations per sentence	Number of xcomp relations per sentence	*xcomp*	sentence
Number of ccomp relations per sentence	Number of ccomp relations per sentence	*ccomp*	sentence
Ratio of verbs with explicit subject	Number of verbs in a sentence with explicit *nsubj* or *csubj* relations, divided by the number of verbs	*ves*	sentence
Ratio of verbs with explicit subject	Number of verbs in a sentence with explicit *nsubj* or *csubj* relations, divided by the number of verbs	*ves_t*	treebank
Word forms per lemma	Number of UD FORM variants for every LEMMA	*wfpl*	treebank
POS sequence variability	Variability of sequences of three POS tags	*vps*	treebank

Table 2: Complexity metrics

3 Results

3.1 Official treebanks

Table 3 shows the values related to each metric and each treebank, with the top 3 values highlighted for each metric. For the *cmn_pud* treebank, *wfpl* is not reported, as this treebank does not include lemma information.

Similary to (Berdicevskis et al., 2018), multiple treebanks were used for each language to get a sense of metric robustness, i.e. whether the metric returns consistent results for different data in the same language. Most metrics are well-behaved in this sense, with *en_atis* being the biggest outlier across the board. The values from this treebank differ substantially from *en_gum* and *en_ewt* in most metrics. One possible justification for this outcome is the nature of how the data was sourced for *en_atis*, namely by collecting and transcribing user interactions with automated inquiry systems related to flight information.

Outside of this case, most metrics yield comparable outputs for most language treebank sets, a somewhat unexpected outcome in reference to the forecasted effect of at times wildly different treebank sizes. The biggest difference was measured between the two German treebanks, *de_gsd* and *de_hdt*, with the latter being over five times larger than the former. Despite this difference, outputs for the two treebanks are similar for all metrics except *wfpl* and *vps*, which were already projected to suffer most from the effect of treebank size (more sentences represent more opportunities for rarer POS tag sequences and lemma forms to appear).

For the robustness aspect, it should be noted that the two treebanks for Hungarian score very similarly across the board, however this is a somewhat unremarkable outcome. The two treebanks were in fact derived from splitting a single original one in half, therefore they are smaller than all other tree-

Treebank	tc	ttr	ptd	lldl	n2v	cxc
en_atis	11.3839	**0.9785**	3.358	6.0374	3.2123	0.5138
en_ewt	14.4258	**0.9449**	3.4635	12.6864	1.2317	1.5512
en_gum	15.6093	0.9279	3.676	13.7041	1.5472	1.6286
de_gsd	16.6016	0.9436	3.8252	14.7274	**3.3466**	1.0269
de_hdt	16.0206	**0.9566**	3.7007	13.9952	2.4751	0.8688
hun_szeged_1	**18.9604**	0.9254	**4.3341**	**16.7121**	**3.2302**	1.7165
hun_szeged_2	**19.022**	0.9234	**4.411**	**16.7956**	**3.3951**	**1.7187**
cmn_gsd	**21.2667**	0.8983	4.286	**18.624**	2.6473	**2.7671**
cmn_pud	18.513	0.9248	**4.288**	16.602	2.0688	**2.217**

Treebank	xcomp	ccomp	ves	ves_t	wfpl	vps
en_atis	0.0962	0.0082	0.2757	0.3578	1.1892	801.0
en_ewt	**0.2439**	**0.1685**	0.4147	0.5304	**1.4458**	**2877.0**
en_gum	0.2396	0.1296	0.4347	0.5092	1.3522	**2634.0**
de_gsd	0.121	0.0443	0.5698	**0.6447**	1.2279	2293.0
de_hdt	0.1122	0.0928	**0.6587**	**0.7232**	**2.5082**	**2685.0**
hun_szeged_1	0.1429	0.0308	**0.5972**	0.5786	1.3355	1081.0
hun_szeged_2	0.1802	0.0374	**0.6212**	**0.6108**	1.3456	1020.0
cmn_gsd	**0.3665**	**0.4353**	0.451	0.4088	1.0003	1756.0
cmn_pud	**0.476**	**0.403**	0.4937	0.4462	**5478.0**	1370.0

Table 3: Results on official UD treebanks. Top 3 values are highlighted for each metric.

banks in the experiment, and, unlike those other treebanks, have the same source. Conclusions derived from this observation are therefore necessarily weaker than those drawn from, for example, the two German treebanks.

When it comes to the specific outputs generated during the experiment, it is generally the case that Hungarian and Chinese treebanks yield values that suggest a higher degree of overall complexity. All metrics except *ttr* (type token ratio), *wfpl* (word forms per lemma) and *vps* (variability of 3 POS tag sequences) show at least either language in the top ranks. Notably, two of these metrics - *wfpl* and *vps* - were projected to be most affected by treebank size, and Hungarian and Chinese have the smallest treebanks in this project's dataset. It should also not go without mention that there is a relation between *tc* and *ttr*: *tc* values are highest for exactly the 4 Hungarian and Chinese treebanks, thus partially explaining the lower expression of *ttr*.

For the metric "ratio of verbs with explicit subject", two variants were reported - *ves* and *ves_t*, the former computed at the sentence level and the latter at the treebank level. This decision was motivated by the presence of several nominal sentences in the treebanks, which made this metric unstable at the sentence level. Seeing the results, even though the numbers and rankings obtained differ, German and Hungarian are still the top languages. The fact that English (which usually requires verbal subjects to be present) has fewer explicit subjects on average than Hungarian (which doesn't) is an unexpected outcome. However, this might be caused by discrepancies in annotation conventions, and not necessarily a true difference between the languages.

The metrics *ccomp* (average number of ccomp relations per sentence) and *xcomp* (average number of xcomp relations per sentence) were a somewhat novel selection in this experiment. An argument was made above that a usage of the xcomp dependency relation might suggest higher overall complexity. Althugh for both *ccomp* and *xcomp* English and Chinese were the highest values, the *xcomp* metric is higher than *ccomp* in all treebanks except *cmn_gsd*. In this sense *cmn_gsd* is the only instance of higher ccomp relation usage, suggesting Chinese to be the least complex language according to the presuppositions related to these metrics - but this case cannot be confidently made without more detailed analysis of how these relations function in the specific treebanks.

Treebank	tc	ttr	ptd	lldl	n2v	cxc
en	**14.6**	0.9264	3.58	**12.44**	1.6202	1.54
deu	13.38	0.9311	3.38	11.44	1.6287	1.18
hun	10.8	**0.9604**	3.2	6.68	**1.711**	1.3
cmn	14.52	0.9117	**3.66**	12.1	1.2491	**1.96**

Treebank	xcomp	ccomp	ves	ves_t	wfpl	vps
en	0.14	0.16	0.6117	0.6548	1.2075	311.0
deu	0.16	0.14	**0.6267**	**0.75**	1.1977	305.0
hun	0.1	**0.24**	0.3417	0.378	**1.2524**	236.0
cmn	**0.5**	0.22	0.4892	0.4745	1.0051	**330.0**

Table 4: Results on in-house parallel treebank. Top values are highlighted for each metric.

3.2 Parallel treebank

To corroborate the observations drawn from the offical UD treebanks, the same set of metrics were computed on the parallel corpus introduced above. Table 4 outlines the results of this second part.

Hungarian and Chinese remain the holders of the highest values in 8 out of 12 (11 if we group *ves* and *ves_t*), lending credence to the theory that these languages are more complex in relation to these metrics.

The metrics *vps* and *wfpl* were highlighted as affected by treebank size in the section above, therefore conferming them on same-size treebanks, however small, might still be a valuable undertaking. Indeed, Hungarian, a language with complex morphology, takes the lead in *wfpl* and Chinese outranks German and English, which have a more constrained word order.

Generally, while the results on the parallel treebank are more or less in line with the official treebanks, the sample size in this case is too small. Thus, meaningful conclusions cannot be drawn. Nevertheless the goal was to highlight the usage of parallel treebanks as a strategy to control for as many factors as possible and isolate aspects related to complexity.

4 Discussion and future research

This project attempted to showcase the use of treebanks in the UD framework for complexity research. To achieve this, a set of treebanks across four languages and a set of text-level metrics related to complexity were selected. The previous sections detailed the results of the experiment.

As mentioned before, not all treebanks are created equal. In the case of this experiment, it was both the case that the treebank had different sizes, and that it is not a guarantee that annotation conventions across language frameworks are harmonic by default. (Berdicevskis et al., 2018), a study similar to this work, took into account this latter obstacle by smoothing the inter-language annotation differences, while the current work took no such approaches. An idea for future work on the topic might be to look at [AUX + VERB + (N/C)SUBJ] structures in relation to the *ves* metric, as the subject can be confusingly attached to either VERB or AUX.

(Berdicevskis et al., 2018) employed a variety of nuanced metrics, which were certainly better suited to the task of gauging language complexity. Although the selection of metrics here can be defended in terms of interpretability and low hardware requirements, future work on the topic should err toward the side of significance rather than interpretability.

On the topic of significance, statistical tests are absent from this experiment, although the results obtained would have benefitted from them. Breaking academic formality for this sentence, I will do my best to learn how to perform them for my future projects.

While no claims are made with regard to overall language complexity or even treebank complexity, perhaps this work eases the way to future experimentation with UD treebanks and the general field of language complexity.

Acknowledgements

I thank Dr. Çağrı Çöltekin for his patient counselling and for providing insight into his 2018 paper, which served as inspiration for most of this work.

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. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 8–17, Brussels, Belgium. Association for Computational Linguistics.

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

Resources

All code for this paper is available at `https://github.com/waron97/dtdp-final-paper`.