Using Universal Dependencies in cross-linguistic complexity research

Aleksandrs Berdicevskis[1], Çağrı Çöltekin[2], Katharina Ehret[3], Kilu von Prince[4,5], Daniel Ross[6], Bill Thompson[7], Chunxiao Yan[8], Vera Demberg[5,9], Gary Lupyan[10], Taraka Rama[11] and Christian Bentz[2]

[1]Department of Linguistics and Philology, Uppsala University
[2]Department of Linguistics, University of Tübingen
[3]Department of Linguistics, Simon Fraser University
[4]Department of German Studies and Linguistics, Humboldt-Universität
[5]Department of Language Science and Technology, Saarland University
[6]Linguistics Department, University of Illinois at Urbana-Champaign
[7]Department of Psychology, University of California, Berkeley
[8]MoDyCo, Université Paris Nanterre & CNRS
[9]Department of Computer Science, Saarland University
[10]Department of Psychology, University of Wisconsin-Madison
[11]Department of Informatics, University of Oslo
aleksandrs.berdicevskis@lingfil.uu.se

Abstract

We evaluate corpus-based measures of linguistic complexity obtained using Universal Dependencies (UD) treebanks. We propose a method of estimating robustness of the complexity values obtained using a given measure and a given treebank. The results indicate that measures of syntactic complexity might be on average less robust than those of morphological complexity. We also estimate the validity of complexity measures by comparing the results for very similar languages and checking for unexpected differences. We show that some of those differences that arise can be diminished by using parallel treebanks and, more importantly from the practical point of view, by harmonizing the language-specific solutions in the UD annotation.

1 Introduction

Analyses of linguistic complexity are gaining ground in different domains of language sciences, such as sociolinguistic typology (Dahl, 2004; Wray and Grace, 2007; Dale and Lupyan, 2012), language learning (Hudson Kam and Newport, 2009; Perfors, 2012; Kempe and Brooks, 2018), and computational linguistics (Brunato et al., 2016). Here are a few examples of the claims that are being made: creole languages are simpler than "old" languages (McWhorter, 2001); languages with high proportions of non-native speakers tend to simplify morphologically (Trudgill, 2011); morphologically rich languages seem to be more difficult to parse (Nivre et al., 2007).

Ideally, strong claims have to be supported by strong empirical evidence, including quantitative evidence. An important caveat is that complexity is notoriously difficult to define and measure, and that there is currently no consensus about how proposed measures themselves can be evaluated and compared.

To overcome this, the first shared task on measuring linguistic complexity was organized in 2018 at the EVOLANG conference in Torun. Seven teams of researchers contributed overall 34 measures for 37 pre-defined languages (Berdicevskis and Bentz, 2018). All corpus-based measures had to be obtained using Universal Dependencies (UD) 2.1 corpora (Nivre et al., 2017).

The shared task was unusual in several senses. Most saliently, there was no gold standard against which the results could be compared. Such a benchmark will in fact never be available, since we cannot know what the *real* values of the constructs we label "linguistic complexity" are.

In this paper, we attempt to evaluate corpus-based measures of linguistic complexity in the absence of a gold standard. We view this as a small step towards exploring how complexity varies

Proceedings of the Second Workshop on Universal Dependencies (UDW 2018), pages 8–17
Brussels, Belgium, November 1, 2018.

Measure ID	Description	Relevant annotation levels
Morphological complexity		
CR_TTR	Type-token ratio	T, WS
CR_MSP	Mean size of paradigm, i.e., number of word forms per lemma	T, WS, L
CR_MFE	Entropy of morphological feature set	T, WS, F, L
CR_CFEwm	Entropy (non-predictability) of word forms from their morphological analysis	T, WS, F, L
CR_CFEmw	Entropy (non-predictability) of morphological analysis from word forms	T, WS, F, L
Eh_Morph	Eh_Morph and Eh_Synt are based on Kolmogorov complexity which is approximated with off-the shelf compression programs; combined with various distortion techniques compression algorithms can estimate morphological and syntactic complexity. Eh_Morph is a measure of word form variation. Precisely, the metric conflates to some extent structural word from (ir)regularity (such as, but not limited to, inflectional and derivational structures) and lexical diversity. Thus, texts that exhibit more word form variation count as more morphologically complex.	T, WS
TL_SemDist	TL_SemDist and TL_SemVar are measures of morphosemantic complexity, they describe the amount of semantic work executed by morphology in the corpora, as measured by traversal from lemma to wordform in a vector embedding space induced from lexical co-occurence statistics. TL_SemDist measures the sum of euclidian distances between all unique attested lemma-wordform pairs.	T, WS, L
TL_SemVar	See TL_SemDist. TL_SemVar measures the sum of by-component variance in semantic difference vectors (vectors that result from subtracting lemma vector from word form vector).	T, WS, L
Syntactic complexity		
CR_POSP	Perplexity (variability) of POS tag bigrams	T, WS, P
Eh_Synt	See Eh_Morph. Eh_Synt is a measure of word order rigidity: texts with maximally rigid word order count as syntactically complex while texts with maximally free word order count as syntactically simple. Eh_Synt relates to syntactic surface patterns and structural word order patterns (rather than syntagmatic relationships).	T, WS
PD_POS_tri	Variability of sequences of three POS tags	T, WS, P
PD_POS _tri_uni	Variability of POS tag sequences without the effect of differences in POS tag sets	T, WS, P
Ro_Dep	Total number of dependency triplets (P, RL, and P of related word). A direct interpretation of the UD corpus data, measuring the variety of syntactic dependencies in the data without regard to frequency.	T, WS, P, ST, RL
YK_avrCW_AT	Average of dependency flux weight combined with dependency length	T, WS, P, ST
YK_maxCW_AT	Maximum value of dependency flux weight combined with dependency length	T, WS, P, ST

Table 1: Complexity measures discussed in this paper. Annotation levels: T = tokenization, WS = word segmentation, L = lemmatization, P = part of speech, F = features, ST = syntactic tree, RL = relation labels. More detailed information can be found in Çöltekin and Rama, 2018 (for measures with the CR prefix), Ehret, 2018 (Eh), von Prince and Demberg, 2018 (PD), Ross, 2018 (Ro), Thompson and Lupyan, 2018 (TL), Yan and Kahane, 2018 (YK).

across languages and identifying important types of variation that relate to intuitive senses of "linguistic complexity". Our results also indicate to what extent UD in its current form can be used for cross-linguistic studies. Finally, we believe that the methods we suggest in this paper may be relevant not only for complexity, but also for other quantifiable typological parameters.

Section 2 describes the shared task and the proposed complexity measures, Section 3 describes the evaluation methods we suggest and the results they yield, Section 4 analyzes whether some of the problems we detect are corpus artefacts and can be eliminated by harmonizing the annotation and/or using the parallel treebanks, Section 5 concludes with a discussion.

2 Data and measures

For the shared task, participants had to measure the complexities of 37 languages (using the "original" UD treebanks, unless indicated otherwise in parentheses): Afrikaans, Arabic, Basque, Bulgarian, Catalan, Chinese, Croatian, Czech, Danish, Greek, Dutch, English, Estonian, Finnish, French, Galician, Hebrew, Hindi, Hungarian, Italian, Latvian, Norwegian-Bokmål, Norwegian-Nynorsk, Persian, Polish, Portuguese, Romanian, Russian (SynTagRus), Serbian, Slovak, Slovenian, Spanish (Ancora), Swedish, Turkish, Ukrainian, Urdu and Vietnamese. Other languages from the UD 2.1 release were not included because they were represented by a treebank which either was too small (less than 40K tokens), or lacked some levels of annotation, or was suspected (according to the information provided by the UD community) to contain many annotation errors. Ancient languages were not included either. In this paper, we also exclude Galician from consideration since it transpired that its annotation was incomplete.

The participants were free to choose which facet of linguistic complexity they wanted to focus on, the only requirement was to provide a clear definition of what is being measured. This is another peculiarity of the shared task: different participants were measuring different (though often related) constructs.

All corpus-based measures had to be applied to the corpora available in UD 2.1, but participants were free to decide which level of annotation (if any) to use. The corpora were obtained by merging together train, dev and test sets provided in the release.

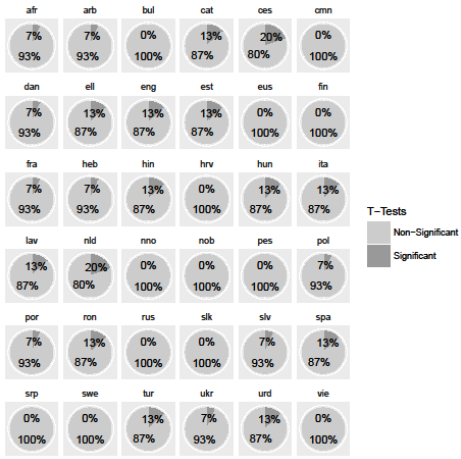

Figure 1: Non-robustness of treebanks. Languages are denoted by their ISO codes.

From every contribution to the shared task, we selected those UD-based measures that we judged to be most important. Table 1 lists these measures and briefly describes their key properties, including those levels of treebank annotation on which the measures are directly dependent (this information will be important in Section 4). We divide measures into those that gauge morphological complexity and those that gauge syntactic complexity, although these can of course be inter-dependent.

In Appendix A, we provide the complexity rank of each language according to each measure.

It should be noted that all the measures are in fact gauging complexities of treebanks, not complexities of languages. The main assumption of corpus-based approaches is that the former are reasonable approximations of the latter. It can be questioned whether this is actually the case (one obvious problem is that treebanks may not be representative in terms of genre sample), but in this paper we largely abstract away from this question and focus on testing quantitative approaches.

3 Evaluation

We evaluate *robustness* and *validity*. By robustness we mean that two applications of the same measure to the same corpus of the same language should ideally yield the same results. See Section 3.1 for the operationalization of this desideratum and the results.

To test validity, we rely on the following idea: if we take two languages that we know from qualitative typological research to be very similar

Figure 2: Non-robustness of measures

to each other (it is not sufficient that they are phylogenetically close, though it is probably necessary) and compare their complexities, the difference should on average be lower than if we compare two random languages from our sample. For the purposes of this paper we define *very similar* as 'are often claimed to be variants of the same language'. Three language pairs in our sample potentially meet this criterion: Norwegian-Bokmål and Norwegian-Nynorsk; Serbian and Croatian; Hindi and Urdu. For practical reasons, we focus on the former two in this paper (one important problem with Hindi and Urdu is that vowels are not marked in the Urdu UD treebank, which can strongly affect some of the measures, making the languages seem more different than they actually are). Indeed, while there certainly are differences between Norwegian-Bokmål and Norwegian-Nynorsk and between Serbian and Croatian, they are structurally very close (Sussex and Cubberley, 2006; Faarlund, Lie and Vannebo, 1997) and we would expect their complexities to be relatively similar. See section 3.2 for the operationalization of this desideratum and the results.

See Appendix B for data, detailed results and scripts.

3.1 Evaluating robustness

For every language, we randomly split its treebank into two parts containing the same number of sentences (the sentences are randomly drawn from anywhere in the corpus; if the total number of sentences is odd, then one part contains one extra sentence), then apply the complexity measure of interest to both halves, and repeat the procedure for n iterations ($n = 30$). We want the measure to yield similar results for the two halves, and we test whether it does by performing a paired t-test on the two samples of n measurements each (some of the samples are not normally distributed, but paired t-tests with sample size 30 are considered robust to non-normality, see Boneau, 1960). We also calculate the effect size (Cohen's d, see Kilgarriff, 2005 about the insufficience of significance testing in corpus linguistics). We consider the difference to be significant and non-negligible if p is lower than 0.10 and the absolute value of d is larger than 0.20. Note that our cutoff point for p is higher than the conventional thresholds for significance (0.05 or 0.01), which in our case means more conservative approach. For d, we use the conventional threshold, below which the effect size is typically considered negligible.

We consider the proportion of cases when the difference is significant and non-negligible a measure of *non-robustness*. See Figure 1 for the non-robustness of treebanks (i.e. the proportion of measures that yielded a significant and non-negligible difference for a given treebank according to the resampling test); see Figure 2 for

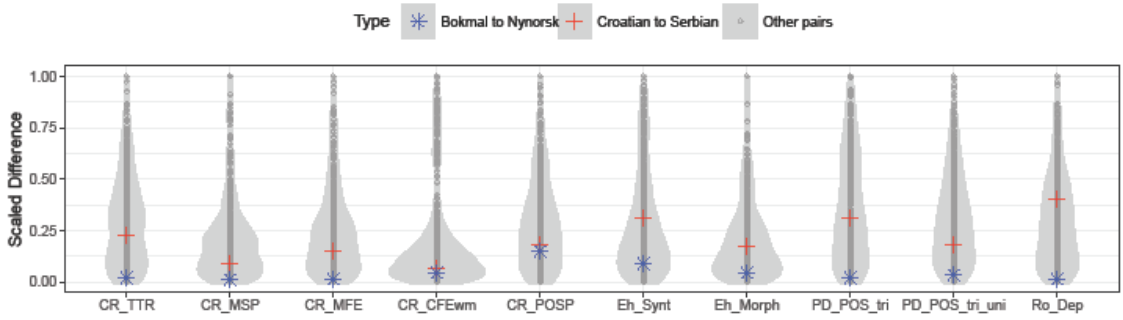

Figure 3: Distributions of pairwise absolute differences between all languages (jittered). Red dots: differences between Serbian and Croatian; blue dots: differences between Norwegian-Bokmål and Norwegian-Nynorsk.

the non-robustness of measures (i.e. the proportion of treebanks for which a given measure yielded a significant and non-negligible difference according to the resampling test).

The Czech and Dutch treebanks are the least robust according to this measure: resampling yields unwanted differences in 20% of all cases, i.e. for three measures out of 15. 12 treebanks exhibit non-robustness for two measures, 9 for one, 13 are fully robust.

It is not entirely clear which factors affect treebank robustness. There is no correlation between non-robustness and treebank size in tokens (Spearman's $r = 0.14$, $S = 6751.6$, $p = 0.43$). It is possible that more heterogeneous treebanks (e.g. those that contain large proportions of both very simple and very complex sentences) should be less robust, but it is difficult to measure heterogeneity. Note also that the differences are small and can be to a large extent random.

As regards measures, CR_POSP is least robust, yielding unwanted differences for seven languages out of 36, while TL_SemDist, TL_SemVar and PD_POS_TRI_UNI are fully robust. Interestingly, the average non-robustness of morphological measures (see Table 1) is 0.067, while that of syntactic is 0.079 (our sample, however, is neither large nor representative enough for any meaningful estimation of significance of this difference). A probable reason is that syntactic measures are likely to require larger corpora. Ross (2018: 28–29), for instance, shows that no UD 2.1 corpus is large enough to provide a precise estimate of RO_DEP. The heterogeneity of the propositional content (i.e. genre) can also affect syntactic measures (this has been shown for EH_SYNT, see Ehret, 2017).

3.2 Evaluating validity

For every measure, we calculate differences between all possible pairs of languages. Our prediction is that differences between Norwegian-Bokmål and Norwegian-Nynorsk and between Serbian and Croatian will be close to zero or at least lower than average differences. For the purposes of this section, we operationalize *lower than average* as 'lying below the first (25%) quantile of the distribution of the differences'.

The Serbian-Croatian pair does not satisfy this criterion for CR_TTR, CR_MSP, CR_MFE, CR_CFEWM, CR_POSP, EH_SYNT, EH_MORPH, PD_POS_TRI, PD_POS_TRI_UNI and RO_DEP. The Norwegian pair fails the criterion only for CR_POSP.

We plot the distributions of differences for these measures, highlighting the differences between Norwegian-Bokmål and Norwegian-Nynorsk and between Serbian and Croatian (see Figure 3).

It should be noted, however, that the UD corpora are not parallel and that the annotation, while meant to be universal, can in fact be quite different for different languages. In the next section, we explore if these two issues may affect our results.

Issue	Instances	Action taken
nob has feature "Voice" (values: "Pass")	1147	Feature removed
nob has feature "Reflex" (values: "Yes")	1231	Feature removed
Feature "Case" can have value "Gen,Nom" in nob	2	None
Feature "PronType" can have value "Dem,Ind" in nob	1	None

Table 2: Harmonization of the Norwegian-Bokmål (nob) and Norwegian-Nynorsk (nno) treebanks.

Issue	Instances	Action taken
hrv has POS DET (corresponds to PRON in srp)	7278	Changed to PRON
hrv has POS INTJ (used for interjections such as e.g. *hajde* 'come on', which are annotated as AUX in srp)	12	Changed to AUX
hrv has POS X (corresponds most often to ADP in srp, though sometimes to PROPN)	253	Changed to ADP
hrv has POS SYM (used for combinations like *20%*, which in srp are treated as separate tokens: 20 as NUM; % as PUNCT)	117	Changed to NUM
hrv has feature "Gender[psor]" (values: "Fem", "Masc,Neut")	342	Feature removed
hrv has feature "Number[psor]" (values: "Plur", "Sing")	797	Feature removed
hrv has feature "Polarity" (values: "Neg", "Pos")	1161	Feature removed
hrv has feature "Voice" (values: "Act", "Pass")	7594	Feature removed
Feature "Mood" can have value "Cnd" in hrv	772	Value removed
Feature "Mood" can have value "Ind" in hrv	18153	Value removed
Feature "PronType" can have value "Int,Rel" in hrv	3899	Value changed to "Int"
Feature "PronType" can have value "Neg" in hrv	138	Value changed to "Ind"
Feature "Tense" can have value "Imp" in hrv	2	None
Feature "VerbForm" can have value "Conv" in hrv	155	Value removed
Feature "VerbForm" can have value "Fin" in hrv	19143	Value removed
hrv has relation "advmod:emph"	43	Changed to "advmod"
hrv has relation "aux:pass"	998	Changed to "aux"
hrv has relation "csubj:pass"	61	Changed to "csubj"
hrv has relation "dislocated"	8	None
hrv has relation "expl"	12	None
hrv has relation "expl:pv"	2161	Changed to "compound"
hrv has relation "flat:foreign"	115	Changed to "flat"
hrv has relation "nsubj:pass"	1037	Changed to "nsubj"
srp has relation "nummod:gov"	611	Changed to "nummod"
srp has relation "det:numgov"	107	Changed to "det"

Table 3: Harmonization of the Croatian (hrv) and Serbian (srp) treebanks.

4 Harmonization and parallelism

The Norwegian-Bokmål and Norwegian-Nynorsk treebanks are of approximately the same size (310K resp. 301K tokens) and are not parallel. They were, however, converted by the same team from the same resource (Øvrelid and Hohle, 2016). The annotation is very similar, but Norwegian-Bokmål has some additional features. We harmonize the annotation by eliminating the prominent discrepancies (see Table 2). We ignore the discrepancies that concern very few instances and thus are unlikely to affect our results.

The Croatian treebank (Agić and Ljubešić, 2015) has richer annotation than the Serbian one (though Serbian has some features that Croatian is missing) and is much bigger (197K resp. 87K tokens); the Serbian treebank is parallel to a subcorpus of the Croatian treebank (Samardžić et al., 2017). We created three extra versions of the Croatian treebank: Croatian-parallel (the parallel subcorpus with no changes to the annotation); Croatian-harmonized (the whole corpus with the annotation harmonized as described in Table 3);

Measure	Harmonization	Parallelism	Both
CR_TTR	0.000	**-0.887**	**-0.890**
CR_MSP	0.005	**-0.877**	**-0.885**
CR_MFE	**-0.648**	-0.271	**-0.924**
CR_CFEwm	**-0.333**	-0.500	**-0.667**
CR_POSP	**-0.988**	-0.505	**-0.646**
Eh_Synt	0.005	**-0.888**	**-0.872**
Eh_Morph	0.191	0.117	**-0.751**
PD_POS_tri	-0.227	**-0.812**	**-0.985**
PD_POS_tri_uni	0.348	**-0.904**	**-0.574**
Ro_Dep	-0.514	-0.114	-0.605

Table 4: Effects of treebank manipulation on the difference between Croatian and Serbian. Numbers show relative changes of the original difference after the respective manipulation. Bold indicates cases when the new difference lies below the defined threshold, i.e. when the measure passes the validity test.

Croatian-parallel-harmonized (the parallel subcorpus with the annotation harmonized as described in Table 3) and one extra version of the Serbian treebank: Serbian-harmonized.

It should be noted that our harmonization (for both language pairs) is based on comparing the stats.xml file included in the UD releases and the papers describing the treebanks (Øvrelid and Hohle, 2016; Agić and Ljubešić, 2015; Samardžić et al., 2017). If there are any subtle differences that do not transpire from these files and papers (e.g. different lemmatization principles), they are not eliminated by our simple conversion.

Using the harmonized version of Norwegian-Bokmål does not affect the difference for CR_POSP (which is unsurprising, given that the harmonization changed only feature annotation, to which this measure is not sensitive).

For Croatian, we report the effect of the three manipulations in Table 4. Using Croatian-parallel solves the problems with CR_TTR, CR_MSP, EH_SYNT, PD_POS_TRI, PD_POS_TRI_UNI. Using Croatian-harmonized and Serbian-harmonized has an almost inverse effect. It solves the problems with CR_MFE, CR_CFEWM, CR_POSP, but not with any other measures. It does strongly diminish the difference for RO_DEP, though. Finally, using Croatian-parallel-harmonized and Serbian-harmonized turns out to be most efficient. It solves the problems with all the measures apart from RO_DEP, but the difference does become smaller also for this measure. Note that this measure had the biggest original difference (see Section 3.2).

Some numbers are positive, which indicates that the difference increases after the harmonization. Small changes of this kind (e.g. for CR_MSP, EH_SYNT) are most likely random, since many measures are using some kind of random sampling and never yield exactly the same value. The behaviour of EH_MORPH also suggests that the changes are random (this measure cannot be affected by harmonization, so Croatian-harmonized and Croatian-parallel-harmonized should yield similar results). The most surprising result, however, is the big increase of PD_POS_TRI_UNI after harmonization. A possible reason is imperfect harmonization of POS annotation, which introduced additional variability into POS trigrams. Note, however, that the difference for CR_POSP, which is similar to PD_POS_TRI_UNI, was reduced almost to zero by the same manipulation.

It can be argued that these comparisons are not entirely fair. By removing the unreasonable discrepancies between the languages we are focusing on, but not doing that for all language pairs, we may have introduced a certain bias. Nonetheless, our results should still indicate whether the harmonization and parallelization diminish the differences (though they might overestimate their positive effect).

5 Discussion

As mentioned in Section 1, some notion of complexity is often used in linguistic theories and analyses, both as an explanandum and an explanans. A useful visualization of many theories that involve the notion of complexity can be obtained, for instance, through The Causal Hypotheses in Evolutionary Linguistics Database (Roberts, 2018). Obviously, we want to be able to

understand such key theoretical notions well and quantify them, if they are quantifiable. To what extent are we able to do this for notions of complexity?

In this paper, we leave aside the question of how well we understand what complexity "really" is and focus on how good we are at quantifying it using corpus-based measures (it should be noted that other types of complexity measures exist, e.g. grammar-based measures, with their own strengths and weaknesses).

Our non-robustness metric shows to what extent a given measure or a given treebank can be trusted. Most often, two equal treebank halves yield virtually the same results. For some treebanks and measures, on the other hand, the proportion of cases in which the differences are significant (and large) is relatively high. Interestingly, measures of syntactic complexity seem to be on average less robust in this sense than measures of morphological complexity. This might indicate that language-internal variation of syntactic complexity is greater than language-internal variation of morphological complexity, and larger corpora are necessary for its reliable estimation. In particular, syntactic complexity may be more sensitive to genres, and heterogeneity of genres across and within corpora may affect robustness. It is hardly possible to test this hypothesis with UD 2.1, since detailed genre metadata are not easily available for most treebanks. Yet another possible explanation is that there is generally less agreement between different conceptualizations of what "syntax" is than what "morphology" is.

Our validity metric shows that closely related languages which should yield minimally divergent results can, in fact, diverge considerably. However, this effect can be diminished by using parallel treebanks and harmonizing the UD annotation. The latter result has practical implications for the UD project. While Universal Dependencies are meant to be universal, in practice language-specific solutions are allowed on all levels. This policy has obvious advantages, but as we show, it can inhibit cross-linguistic comparisons. The differences in Table 2 and Table 3 strongly affect some of our measures, but they do not reflect any real structural differences between languages, merely different decisions adopted by treebank developers. For quantitative typologists, it would be desirable to have a truly harmonized (or at least easily harmonizable) version of UD.

The observation that non-parallelism of treebanks also influences the results has further implications for a corpus-based typology. Since obtaining parallel treebanks even for all current UD languages is hardly feasible, register and genre variation are important confounds to be aware of. Nonetheless, the Norwegian treebanks, while non-parallel, did not pose any problems for most of the measures. Thus, we can hope that if the corpora are sufficiently large and well-balanced, quantitative measures of typological parameters will still yield reliable results despite the non-parallelism. In general, our results allow for some optimism with regards to quantitative typology in general and using UD in particular. However, both measures and resources have to be evaluated and tested before they are used as basis for theoretical claims, especially regarding the interpretability of the computational results.

References

Agić, Željko and Nikola Ljubešić. 2015. Universal Dependencies for Croatian (that Work for Serbian, too). In *Proceedings of the 5th Workshop on Balto-Slavic Natural Language Processing*. Association for Computational Linguistics, pages 1-8. http://www.aclweb.org/anthology/W15-5301

Aleksandrs Berdicevskis and Christian Bentz. 2018. *Proceedings of the First Shared Task on Measuring Language Complexity*.

Boneau, Alan. 1960. The effects of violations of assumptions underlying the *t* test. *Psychological Bulletin* 57(1): 49-64. https://doi.org/10.1037/h0041412

Dominique Brunato, Felice Dell'Orleta, Giulia Venturi, Thomas François, Philippe Blache. 2016. *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity*. Association for Computational Linguistics. http://www.aclweb.org/anthology/W16-4100

Çağrı Çöltekin and Taraka Rama. 2018. Exploiting universal dependencies treebanks for measuring morphosyntactic complexity. In *Proceedings of the First Shared Task on Measuring Language Complexity*, pages 1-8.

Östen Dahl. 2004. *The growth and maintenance of linguistic complexity*. John Benjamins, Amsterdam, The Netherlands.

Rick Dale and Gary Lupyan. 2012. Understanding the origins of morphological diversity: The linguistic niche hypothesis. *Advances in Complex Systems* 15(3): 1150017 https://doi.org/10.1142/S0219525911500172.

Katharina Ehret. 2017. An information-theoretic approach to language complexity: variation in naturalistic corpora. Ph.D. thesis, University of Freiburg. https://doi.org/10.6094/UNIFR/12243.

Katharina Ehret. 2018. Kolmogorov complexity as a universal measure of language complexity. In *Proceedings of the First Shared Task on Measuring Language Complexity*, pages 8-14.

Jan Terje Faarlund, Svein Lie and Kjell Ivar Vannebo. 1997. *Norsk referansegrammatik*, Universitetsforlaget, Oslo, Norway.

Carla Hudson Kam and Elissa Newport. 2005. Regularizing unpredictable variation: The roles of adult and child learners in language formation and change. *Language Learning and Development* 1(2):151-195. https://doi.org/10.1080/15475441.2005.9684215.

Vera Kempe and Patricia Brooks. 2018. Linking Adult Second Language Learning and Diachronic Change: A Cautionary Note. *Frontiers in Psychology.* https://doi.org/10.3389/fpsyg.2018.00480.

Adam Kilgarriff. 2005. Language is never, ever, ever random. *Corpus Linguistics and Linguistic Theory* 1–2:263-275. https://doi.org/10.1515/cllt.2005.1.2.263.

John McWhorter. 2001. The world's simplest grammars are creole grammars. *Linguistic Typology* 5(2-3):125-166. https://doi.org/10.1515/lity.2001.001.

Joakim Nivre, Johan Hall, Jens Nilsson, Atanas Chanev, Gülşen Eryigit, Sandra Kübler, Svetoslav Marinov and Erwin Marsi. 2007. MaltParser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering* 13(2):95–135. https://doi.org/10.1017/S1351324906004505.

Joakim Nivre, Agić Željko, Lars Ahrenberg et al. 2017. Universal Dependencies 2.1, LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University, http://hdl.handle.net/11234/1-2515.

Amy Perfors. 2012. When do memory limitations lead to regularization? An experimental and computational investigation. *Journal of Memory and Language* 67: 486-506. https://doi.org/10.1016/j.jml.2012.07.009.

Øvrelid, Lilja and Petter Hohle. 2016. Universal Dependencies for Norwegian. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016).* European Language Resources Association, pages 1579-1585.

Sean Roberts. 2018. Chield: causal hypotheses in evolutionary linguistics database. In *The Evolution of Language: Proceedings of the 12th International Conference (EVOLANGXII)*. https://doi.org/10.12775/3991-1.099.

Kilu von Prince and Vera Demberg. 2018. POS tag perplexity as a measure of syntactic complexity. In In *Proceedings of the First Shared Task on Measuring Language Complexity*, pages 20-25.

Daniel Ross. 2018. Details matter: Problems and possibilities for measuring cross-linguistic complexity. In *Proceedings of the First Shared Task on Measuring Language Complexity*, pages 26-31.

Samardžić, Tanja, Mirjana Starović, Agić Željko and Nikola Ljubešić. 2017. Universal Dependencies for Serbian in Comparison with Croatian and Other Slavic Languages. In *Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing*. Association for Computational Linguistics, pages 39-44. http://www.aclweb.org/anthology/W17-1407

Roland Sussex and Paul Cubberley. 2006. *The Slavic languages*. Cambridge University Press, Cambridge, UK.

Bill Thompson and Gary Lupyan. 2018. Morphosemantic complexity. In *Proceedings of the First Shared Task on Measuring Language Complexity*, pages 32-37.

Peter Trudgill. 2011. *Sociolinguistic typology: social determinants of linguistic complexity*. Oxford University Press, Oxford, UK.

Alison Wray and George Grace. 2007. The consequences of talking to strangers: Evolutionary corollaries of socio-cultural influences on linguistic form. *Lingua* 117(3):543-578. https://doi.org/10.1016/j.lingua.2005.05.005.

Chunxiao Yan and Sylvain Kahane. 2018. Syntactic complexity combining dependency length and dependency flux weight. In *Proceedings of the First Shared Task on Measuring Language Complexity*, pages 38-43.

A Languages ranked by complexity (descending order)

Language A	CR_TTR	CR_MSP	CR_MFE	CR_CFEwm	CR_CFEmw	CR_POSP	Eh_Synt	Eh_Morph	YK_avrCW_AT	YK_maxCW_AT	Ro_Dep	PD_POS_tri	PD_POS_tri_uni	TL_SemDist	TL_SemVar
afr	35	31	26	30	22	26	2	36	7	23	15	29	32	33	33
arb	19	18	23	3	31	20	22	8	3	3	12	31	16	2	2
eus	12	2	14	6	2	23	20	25	16	25	8	13	9	16	16
bul	13	16	11	36	9	22	17	17	33	33	33	24	29	19	19
cat	28	28	28	19	13	30	4	30	10	5	20	28	26	29	29
cmn	17	35	35	8	35	21	32	1	4	6	18	10	3	35	35
hrv	10	9	15	9	27	5	19	22	21	28	2	5	6	15	15
ces	3	14	1	13	26	3	26	12	14	1	9	3	12	17	17
dan	22	27	17	14	16	4	28	7	15	25	22	19	27	28	27
nld	24	32	33	28	4	6	23	18	11	14	3	16	21	31	31
eng	31	30	31	7	5	1	14	15	30	21	1	8	31	34	34
est	8	8	16	26	10	17	36	4	27	23	27	4	10	6	6
fin	1	4	8	35	32	9	31	13	23	4	10	7	5	5	5
fra	18	29	30	20	3	34	10	21	23	11	24	32	34	27	28
ell	30	6	12	4	8	13	5	35	12	19	29	27	28	11	12
heb	29	19	21	15	21	33	34	2	29	29	5	34	33	1	1
hin	33	33	24	2	34	35	7	33	5	18	36	35	20	32	32
hun	15	21	7	23	29	25	9	29	6	16	11	23	11	18	18
ita	26	22	27	31	5	29	11	27	16	6	31	33	36	23	23
lav	11	7	4	27	20	15	21	16	26	27	7	6	8	7	7
nob	23	23	18	25	19	7	25	14	32	29	26	15	25	26	25
nno	25	26	20	16	17	2	18	20	31	20	24	18	23	24	24
pes	32	10	34	32	1	32	13	6	1	6	28	25	2	3	4
pol	5	15	2	11	11	24	35	5	35	34	32	22	22	12	10
por	20	25	32	5	24	19	15	24	13	17	23	30	35	25	26
ron	14	12	13	33	23	18	16	23	16	12	4	14	13	20	20
rus	2	5	10	24	11	16	27	19	28	9	13	2	7	10	11
srp	16	3	22	21	30	11	6	34	22	32	17	20	15	9	9
slk	6	11	3	12	14	8	29	3	36	36	19	9	30	8	8
slv	9	13	9	16	18	10	30	10	25	31	35	12	19	14	13
spa	21	24	25	29	28	27	8	28	9	13	16	26	24	21	22
swe	27	20	19	18	14	14	12	32	20	2	21	21	18	22	21
tur	7	1	6	34	7	28	24	9	8	21	6	11	4	4	3
ukr	4	17	5	10	25	12	33	11	19	9	14	1	14	13	14
urd	34	34	29	1	33	36	1	31	2	15	30	36	17	30	30
vie	36	36	36	22	35	31	3	26	34	35	33	17	1	36	36

B Supplementary material

Data, detailed results and scripts that are necessary to reproduce the findings can be found at https://sites.google.com/view/sasha-berdicevskis/home/resources/sm-for-udw-2018