EXPLOITING UNIVERSAL DEPENDENCIES TREEBANKS FOR MEASURING MORPHOSYNTACTIC COMPLEXITY

Çağrı Çöltekin[*1] and Taraka Rama[2]

[*]Corresponding Author: ccoltekin@sfs.uni-tuebingen.de
[1]Department of Linguistics, University of Tübingen, Country
[2]Department of Informatics, University Oslo, Norway

We present six different measures of morphosyntactic complexity, calculated on 37 Universal Dependencies treebanks. We define the measures (some of which are not published in the earlier literature), present the results, and discuss relationships between the measures.

1. Introduction

There has been recent interest in quantifying linguistic complexity (Juola, 1998; Dahl, 2004; Newmeyer & Preston, 2014; Bentz, Alikaniotis, Cysouw, & Ferrer-i Cancho, 2017; Koplenig, Meyer, Wolfer, & Mueller-Spitzer, 2017; Stump, 2017). Besides the theoretical interest, quantifying complexity of languages or subsystems of languages is also important for first and second language acquisition research. In this paper, we present a number of morphosyntactic measures, some proposed in earlier literature, and some novel to the best of our knowledge.

The Measuring Linguistic Complexity (MLC) shared task aims to bring together different measures of linguistic complexity, encouraging the use of Universal Dependencies (UD) treebanks (Nivre et al., 2016). The UD project defines a unified tagset, and the UD treebanks already include a large number of languages.[1] The multi-lingual focus of the UD project requires paying attention to linguistic typology (Croft, Nordquist, Looney, & Regan, 2017), and the treebanks, in return, constitute a promising resource for the typological (and in general multi-lingual) research. Not surprisingly, the MLC shared task offers a subset of the UD treebanks as the data set for measuring complexity of (subsystems of) languages.

In this paper, we present a number of quantitative measures of morphosyntactic complexity, namely, *type/token ratio* (TTR, e.g., Kettunen, 2014); *mean size of paradigm* (MSP Xanthos et al., 2011); *entropy of morphological-feature distribution*; *entropy of morphological-feature distribution conditioned on the word*

[1]Current UD release (v2.1) includes over 100 treebanks covering 64 languages. The candidate treebanks for the upcoming release includes treebanks for 16 other languages.

forms; *entropy of word-form distribution conditioned on morphological features*; *and part-of-speech tag n-gram perplexity*, calculated on the MLC selection of the 37 UD treebanks.

2. Measures

We report five measures (TTR, MSP, and variants of morphological feature entropy) for measuring morphological complexity, and one, POS tag n-gram perplexity, for measuring syntactic complexity. Except the first two (TTR and MSP), the measures discussed here are all suitable for richly-annotated corpora, and to our knowledge not used in this form in the previous literature.

2.1. *Type/token ratio (TTR)*

The TTR is a time-tested metric for measuring linguistic complexity. When used as a measure of complexity of a language, high TTR indicates rich morphology. Since the TTR depends on corpus length, it is a common practice to calculate the TTR using a fixed window size (Kettunen, 2014). We calculate the TTR on a fixed-length random sample, and take average over multiple samples. The sampling procedure is described in Section 3.

2.2. *Mean size of paradigm (MSP)*

Xanthos et al. (2011) propose the MSP as a measure of morphological complexity, and show its relation with the acquisition of morphology by young learners. The MSP is simply the number of word-form types divided by the number of lemma types. The MSP also depends on the text size. Hence, similar to Xanthos et al. (2011), we use a sampling-based approach (as in the TTR calculation).

2.3. *Morphological feature entropy (MFE)*

Any corpus that annotates words (or tokens) with a set of labels defines a categorical distribution. With MFE (defined in Equation 1), we estimate the categorical distribution of morphological features from the treebank, and calculate its entropy.

$$\mathrm{MFE} = -\sum_{\mathrm{f}} \mathrm{p}(\mathrm{f}) \log_2 \mathrm{p}(\mathrm{f}) \tag{1}$$

where f ranges over all observed feature-value pairs (e.g., `Case=Acc`) in the treebank. The probabilities are estimated with the maximum likelihood estimation (MLE) over all tokens (not types).

Intuitively, the entropy of this distribution indicates the richness of the morphological features encoded in the language. Everything being equal, a language with a larger morphological feature inventory will have higher MFE. However, the shape of the distribution also matters. A distribution that tends towards the uniform distribution, where all labels are equally likely, will also have higher entropy

compared to distributions that favor only a few high-probability (or frequent) features. Since the MFE does not depend on corpus size, we report values that are calculated over the complete available corpus.[2] This measure is similar to the *enumerative complexity* as defined by Ackerman and Malouf (2013).

2.4. *Conditional feature entropy*

Another aspect or dimension of morphological complexity is about transparency of a morphological system. Arguably, if we can predict morphological features from surface forms, and surface forms from morphological features, the language exhibits less complexity – e.g., when viewed from a learner's perspective.

As a first approximation for measuring transparency of the morphological system, we calculate two average conditional feature entropy values. The conditional entropy of a distribution Y given another distribution X is defined as

$$H(Y|X) = \sum_{x \in X, y \in Y} p(x, y) \log_2 p(y|x) \ .$$

The first measure we present, $\mathrm{CFE}_{w|m}$, is simply the conditional entropy of word forms given morphological features, $H(w \mid m)$, and the second measure, $\mathrm{CFE}_{m|w}$, is the conditional entropy of features given word forms, $H(m \mid w)$. It should be noted that these measures do not only measure the complexity of the morphological system but also measure the lexical complexity or ambiguity.

The conditional entropy measures we use are similar to *integrative complexity* defined by Ackerman and Malouf (2013). However, our measures reflect actual usage as reflected by the morphologically annotated corpora at hand, as opposed to the paradigm tables extracted from descriptive grammars.

2.5. *POS tag n-gram perplexity (POSP)*

As a measure of predictability of strictness of word order, we also compute the average perplexity of the UD POS tag n-grams. The perplexity is a popular measure of unpredictability in computational linguistics literature. It is defined as $2^{H(X)}$, where $H(X)$ is the entropy of a probability distribution X (of POS tag sequences in our case). The intuitive interpretation of POSP is the average number of possible POS tags after each position in the corpus. Intuitively, the languages with more strict word order is expected to have lower entropy (hence lower POSP). The POSP should correlate with the morphological complexity, particularly MFE, since rich morphology is typically associated with flexibility in the word order.

In this paper, we only present results of bigram perplexity. However, this can easily be extended to use higher order n-grams, or using entropy rate (Kontoyiannis, Algoet, Suhov, & Wyner, 1998; Gao, Kontoyiannis, & Bienenstock, 2008) for estimating the entropy of the POS tag sequence.

[2] However, the estimation of the underlying distribution will be better with larger corpora.

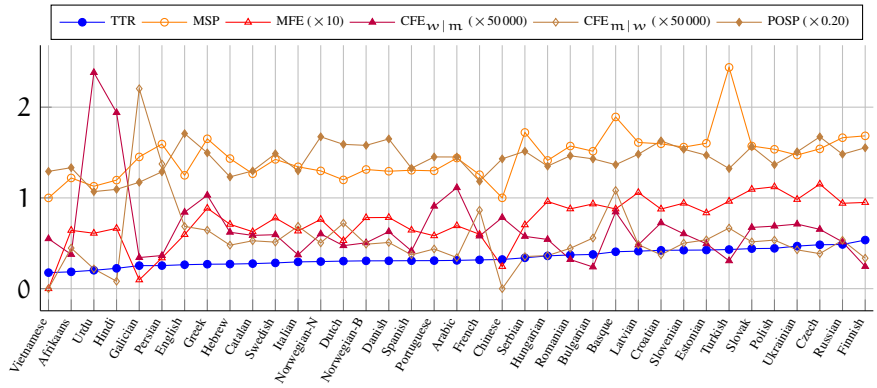

Figure 1. The values of the complexity measures. The measures are linearly scaled to fit into the same y-axis range, the languages are sorted in order of increasing TTR.

3. Data and experimental setup

The data set contains 37 treebanks from Universal Dependencies (UD) project, from 36 languages.[3] Although all treebanks conform to UD v2 annotation scheme, the sizes of the treebanks and some aspects of annotations vary considerably. The smallest treebank (Hungarian) has 1 801 sentences and 42 032 tokens, and the largest (Czech) consists of 87 914 sentences and 1 506 484 tokens. All treebanks, except Galician, includes morphological feature annotations. The usage of UD POS tag inventory is relatively stable across languages. The number of POS tags used vary between 14 and 18. The morphological features and relation types used in different treebanks are more varied, ranging between 2 to 29 and 25 to 55 for morphological feature labels and dependency labels, respectively.

As noted above, some of our measures depend on text size. To be able to get comparable measures, we calculate TTR and MSP from 20 000 tokens sampled randomly. The numbers we report are the mean of 1 000 random samples.[4]

4. Results and Discussion

We present values of all measures discussed in Figure 1. The correlation between the languages are reported in Table 1. The overall results agree with our expectations and the earlier literature. The languages known to be more morphologically complex, are placed on the upper end of the scale with respect to measures that indicate enumerative morphological complexity. However, we also observe that

[3]Norwegian is represented by two treebanks, with different, but closely related dialects that also follow different orthographic conventions.

[4]The source code used for calculating the measures is publicly available at `https://github.com/coltekin/mlc2018`.

Table 1. Correlations between all measures. The values presented in the upper triangle matrix are Pearson's correlation coefficient, while Spearman's rank correlation is listed in the lower triangle.

	TTR	MSP	MFE	$\text{CFE}_{w\|\text{m}}$	$\text{CFE}_{\text{m}\|w}$	POSP
TTR		0.617 3	0.740 2	–0.333 5	–0.064 5	0.460 1
MSP	0.651 5		0.556 9	–0.308 9	0.276 4	0.142 0
MFE	0.776 4	0.658 4		–0.063 1	–0.237 1	0.418 5
$\text{CFE}_{w\|\text{m}}$	–0.100 8	–0.235 4	–0.027 3		–0.315 6	–0.337 7
$\text{CFE}_{\text{m}\|w}$	–0.027 3	0.254 5	–0.029 9	–0.292 3		–0.175 3
POSP	0.422 2	0.236 8	0.402 3	0.122 3	–0.022 3	

there is a modest but negative correlation between the enumerative complexity and integrative complexity measures used in this study. Furthermore, the (enumerative) morphological complexity, as expected, is also moderately correlated with flexibility of the word-order of the language measured by POSP.

The results also show some curious differences, e.g., Chinese showing moderately high TTR, despite lower MSP and MFE. Some of these, e.g., unexpectedly low MFE for Galician, however, is due to lack of annotations in the particular treebank. POSP seems to correlate with morphological complexity measures, indicating that POS tag sequences are less predictable in morphologically rich languages. However, some observations in Figure 1 needs further investigations. For example, the fact that Germanic languages, including English, showing rather high POSP, and despite being morphologically complex, Turkish showing showing a low POSP. Some of these differences may be due to the fact that our measurements are based on bigrams, hence being sensitive word order flexibility in local contexts, e.g., noun phrases, rather than flexibility at the level of the clause.

There are two major differences between the current study (also many others in this volume) and most earlier corpus- and grammar-based work on quantifying linguistic complexity. First, we make use of rich linguistic annotations, which offer many novel ways to measure linguistic complexity. Second, unlike many earlier studies, our material is not a (translated) parallel corpus collection. This allows measuring the complexity on a more 'natural' linguistic data, however, it also requires measures that indicate the differences between the languages, rather than other dimensions such as domain, genre or style. Compared to works that are based on descriptive grammars, working with relatively small corpora may result in missing some (rare) linguistic constructions. In this respect, larger (automatically annotated) data sets can be useful, or recent grammar-book treebanks (Çöltekin, 2015; Rama & Vajjala, 2017) may offer an interesting middle ground.

Although the measures and the results presented here needs further investigation and refinements that are beyond the scope of this short paper, the results are encouraging about using richly and uniformly annotated corpora, such as UD treebanks, for investigating many aspects of linguistic complexity.

References

Ackerman, F., & Malouf, R. (2013). Morphological organization: The low conditional entropy conjecture. *Language*, *89*(3), 429–464.

Bentz, C., Alikaniotis, D., Cysouw, M., & Ferrer-i Cancho, R. (2017). The entropy of words—learnability and expressivity across more than 1000 languages. *Entropy*, *19*(6), 275.

Çöltekin, Ç. (2015). A grammar-book treebank of Turkish. In M. Dickinson, E. Hinrichs, A. Patejuk, & A. Przepiórkowski (Eds.), *Proceedings of the 14th workshop on treebanks and linguistic theories (tlt 14)* (pp. 35–49). Warsaw, Poland.

Croft, W., Nordquist, D. I., Looney, K., & Regan, M. (2017). Linguistic typology meets universal dependencies. In *Proceedings of the 15th workshop on treebanks and linguistic theories (tlt15)* (pp. 63–75). Bloomington, IN, USA.

Dahl, Ö. (2004). *The growth and maintenance of linguistic complexity*. John Benjamins.

Gao, Y., Kontoyiannis, I., & Bienenstock, E. (2008). Estimating the entropy of binary time series: Methodology, some theory and a simulation study. *Entropy*, *10*(2), 71–99.

Juola, P. (1998). Measuring linguistic complexity: The morphological tier. *Journal of Quantitative Linguistics*, *5*(3), 206–213. doi: 10.1080/09296179808590128

Kettunen, K. (2014). Can type-token ratio be used to show morphological complexity of languages? *Journal of Quantitative Linguistics*, *21*(3), 223-245. doi: 10.1080/09296174.2014.911506

Kontoyiannis, I., Algoet, P. H., Suhov, Y. M., & Wyner, A. J. (1998). Nonparametric entropy estimation for stationary processes and random fields, with applications to english text. *IEEE Transactions on Information Theory*, *44*(3), 1319–1327.

Koplenig, A., Meyer, P., Wolfer, S., & Mueller-Spitzer, C. (2017). The statistical trade-off between word order and word structure–large-scale evidence for the principle of least effort. *PloS one*, *12*(3), e0173614.

Newmeyer, F. J., & Preston, L. B. (2014). *Measuring grammatical complexity*. Oxford University Press.

Nivre, J., de Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajič, J., Manning, C., ... Zeman, D. (2016). Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the tenth international conference on language resources and evaluation (lrec'16)* (pp. 23–28). Portorož, Slovenia.

Rama, T., & Vajjala, S. (2017). A Telugu treebank based on a grammar book. In *Proceedings of the 16th international workshop on treebanks and linguistic theories* (pp. 119–128). Prague, Czechia.

Stump, G. (2017). The nature and dimensions of complexity in morphology. *Annual Review of Linguistics*, *3*, 65–83.

Xanthos, A., Laaha, S., Gillis, S., Stephany, U., Aksu-Koç, A., Christofidou, A., … Dressler, W. U. (2011). On the role of morphological richness in the early development of noun and verb inflection. *First Language*, *31*(4), 461-479. doi: 10.1177/0142723711409976