POS TAG PERPLEXITY AS A MEASURE OF SYNTACTIC COMPLEXITY

Kilu von Prince*[1,2] and Vera Demberg[2]

*Corresponding Author: kilu.von.prince@hu-berlin.de
[1]Humboldt-Universität, Berlin, Germany
[2]Universität des Saarlandes, Saarbrücken, Germany

Comparing languages of the world with respect to their complexity is a longstanding open question in linguistics. We here focus on syntactic complexity, a concept that has been particularly hard to address due to the lack of readily available syntactically annotated corpora and the intricacies of syntactic theories. We propose to use a simple information-theoretic measure, perplexity, on the POS tag sequence of texts. Perplexity captures how predictable POS tags are on average given their recent co-texts. Calculating perplexity based on POS tag sequences helps us to abstract away from morphological or lexical features of the language, in order to get at the predictability of word order. In this paper, we compare POS tag perplexity to other recently proposed measures of syntactic complexity, and evaluate measures by correlating them with expert-proposed scores of syntactic flexibility (Bakker 1998).

1. Introduction

The question of how and why languages may differ in terms of their overall or partial complexity is one of the oldest and most hotly debated issues in typology (Nichols, 1992; Trudgill, 2011; McWhorter, 2001; Sampson, 2009; Joseph & Newmeyer, 2012). Since Juola (1998), several attempts have been made to assess the complexity of languages based on texts rather than typological features (Juola, 2008; Futrell, Mahowald, & Gibson, 2015; Ehret & Szmrecsanyi, 2016; Bentz, 2016; Koplenig, Meyer, Wolfer, & Müller-Spitzer, 2017). In this paper, we will assess the viability of using trigram perplexity at the POS level for assessing cross-linguistic variation between non-parallel corpora. This measure is defined as below:

(1) Trigram perplexity:
$$2^{-\frac{1}{N}\sum_{n=1}^{N} P(pos_n|pos_{n-2},pos_{n-1}) \log_2 P(pos_n|pos_{n-2},pos_{n-1})}$$

The reason for working on POS tag sequences instead of words directly is to avoid possible confounds due to writing systems, lexical richness or choice, and, to some extent, compensate for the fact that the data we are working with are not parallel texts.

2. Related Work

We compare the estimates of our method to various previously proposed measures, which we will briefly introduce here.
Expert-ratings of syntactic complexity. Bakker (1998) rated syntactic flexibility, consistency and consequence of languages based on twelve binary grammatical features as described in descriptive accounts and expert questionnaires on each language. Bakker's syntactic flexibility measure, which was also used in prior evaluations such as Ehret and Szmrecsanyi (2016) seemed like the most representative source of expert syntactic complexity ratings against which automatic measures can be evaluated.
Zip compression as an approximation to Kolmogorov complexity. Zip compression has been known to approximate Kolmogorov complexity and has previously been used as a measure of linguistic complexity (Juola, 1998, 2008; Ehret & Szmrecsanyi, 2016). We calculated zip compression for each of the corpora, to achieve best possible comparability with our proposed POS tag perplexity metric.
Ehret and Szmrecsanyi (2016). The authors used a parallel corpus consisting of translations of Alice in Wonderland into nine lanuages, compiled by Annemarie Verkerk, and non-parallel newspaper corpora from the same languages. To measure syntactic complexity, they masked syntactic regularities by randomly deleting 10% of all word tokens. They then measured the difference between the zip-compressed original text and the zip-compressed masked version.
Koplenig et al. (2017) used the massive Parallel Bible Corpus Mayer and Cysouw (2014), with translation into almost 1200 languages. They also created syntactically masked versions of each text by scrambling sentence-internal word order. They then measured an approximation to entropy for the masked and unmasked versions and calculated the difference between them as a measure of syntactic complexity.

3. Methods

We first analyzed the distribution of POS tags across corpora to ensure comparability. While some of those differences may reflect genuine cross-linguistic variation that speaks to differences in the size of syntactic inventories, we need to keep in mind that some variation may have been caused by language-external factors instead. For example, the Chinese corpus does not use the tag for subordinating conjunctions, even though it has very straightforward candidates for this category, such as yīnwéi, "because", which is tagged as an adposition instead. The Arabic corpus contains a much larger set of Other tags compared to the other corpora, which may be an indication of the limits of the applicability of a universal tag set. Such inconsistencies are a potential cause for concern for future fine-grained comparative work, especially if the list of languages in the set expands to include more non-European languages.

To assess possible effects of POS tag distributions across languages, we also measured unigram perplexity $2^{-\sum_{pos \in \mathrm{POS}} P(pos) \log_2 P(pos)}$; the probability of a POS tag pos was estimated in terms of its frequency in the corpus. Unigram perplexity over POS tags hence quantifies the differences in entropy of the POS tag inventory of a language. In order to separate out trigram perplexity from unigram perplexity, we propose an additional measure: trigram perplexity divided by unigram perplexity to quantify the predictability of syntactic categories given previous context compared to POS tag frequencies. This measure gives us a sense of the predictability of word order that is independent from how big and balanced the inventory of POS tags is.

For calculating perplexity and zip compression, we extracted POS tags from each of the corpora, split the files into chunks of 42k tags – the size of the smallest corpus. This allowed us to also assess the effect of corpus size on complexity scores and also allowed us to calculate variance for different subsets of texts for the same language. We found that estimates were generally reliable; our results below report perplexities for the complete dataset for each language, as our experiments showed that estimates on 42k subcorpora correlated at Spearman's rho 0.98 with estimates from the full corpora. This result demonstrates that working on POS sequences avoids having to deal with data sparsity issues.

4. Results

Figure 1 shows that our perplexity measures are correlated with the syntactic flexibility values proposed by Bakker (1998) ($\rho = .45$, $p < 0.05$). The statistical analysis also shows that our measures predict human ratings by Bakker more reliably than previously proposed measures (Juola, 1998; Ehret & Szmrecsanyi, 2016; Koplenig et al., 2017). Figure 1 visually illustrates the correlation between our Trigram/Unigram measure and Bakker flexibility scores.

Among the languages for which there are no Bakker scores, our perplexity measures would predict that Hebrew, Afrikaans, Hindi, Urdu and Arabic are among the syntactically more complex languages (if complexity means lack of flexibility). For Vietnamese, Chinese, Persian, Hungarian, Ukrainian and Czech, the classification as mid range or low syntactically complex languages depends on whether unigram perplexity is taken into account: Ukrainian and Czech have high unigram perplexity, and would hence be classified as highly flexible languages in the trigram measure, but

Table 1. Spearman's correlation between various automatic measures of syntactic complexity and Bakker (1998) flexibility scores. Column dir indicates whether the correlation with Bakker scores is expected to be positive or negative. The right-hand part of the table compares only the set of seven languages used both as part of Ehret & Szmrecsanyi (2016) and as part of the datasets provided for the present workshop.

measure	dir	corr	pval	# lang	corr	pval	# lang
avgzip	neg	-0.32	0.11	26	-0.71	0.07	7
Koplenig et al. '17	neg	-0.36	0.08	24	-0.29	0.53	7
E&S'16: Parallel Alice	neg				-0.71	0.07	7
E&S'16: News	neg				-0.43	0.33	7
unigram perplexity	NA	0.12	0.57	26	-0.09	0.85	7
trigram perplexity	pos	0.45	0.02	26	0.87	0.01	7
trigram/unigram	pos	0.44	0.02	26	0.85	0.01	7

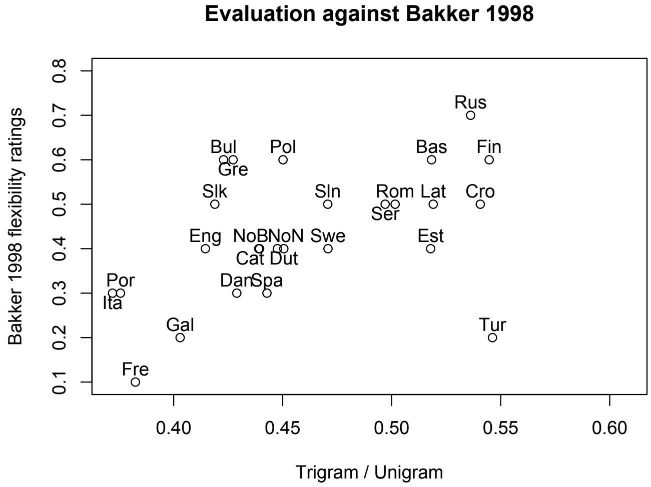

Figure 1. Correlations of trigram perplexities divided by unigram perplexities with flexibility values in Bakker 1998.

as medium complexity languages in the trigram/unigram measure. On the other hand, Persian, Vietnamese and Chinese have low unigram perplexities and hence are only classified as highly flexible languages in the trigram/unigram measure.

5. Discussion

In sum, our results show that surprisal values of POS tags, even in relatively small, non-parallel corpora can be a meaningful measure of syntactic complexity and perform better than similar methods at the word level. Our hypothesis is that measures at the word level may be compromised by unrelated factors, such as the rate of homographs in a given corpus. By focusing on POS levels, these factors can be avoided.

While we did find a significant overall correlation of our methods with Bakker (1998), some languages show a much better fit than others. In particular, Turkish, Bulgarian and Greek are outside the expected range. At this point, we do not have a perfect explanation for these mismatches. Unfortunately, the ratings in Bakker (1998) are not entirely transparent: it is unclear exactly which feature combination is assigned to each language. Despite the very low score of 0.2 (on a scale from zero to one) from Bakker (1998), Turkish could be expected to receive a relatively high score, since it is well-known to be a free-word-order language. In this case, it may therefore be that the corpus-based measure gives a more accurate estimate of the actual flexibility the language has. Bulgarian and Greek are also known for their relative freedom of word order, in line with Bakker's high scores, so their comparatively low values of POS trigram perplexity are rather unexpected. It might be that these differences between description-based assessments and corpus-based measures speak to actual differences between theoretical possibilities and their implementation in language use. It is however also possible that these mismatches are due to imperfections in the POS tagger (which may affect some languages more than others) or a non-representative selection of syntactic features in Bakker (1998).

References

Bakker, D. (1998). Flexibility and consistency in word order patterns in the languages of europe. In A. Siewierska (Ed.), Constituent order in the languages of Europe: Empirical approaches to language typology (Vol. 20, p. 383-420). Mouton De Gruyter.

Bentz, C. (2016). The low-complexity-belt: evidence for large-scale language contact in human prehistory? In S. Roberts, C. Cuskley, L. McCrohon, L. Barceló-Coblijn, O. Feher, & V. T. (Eds.), The evolution of language: Proceedings of the 11th international conference.

Ehret, K., & Szmrecsanyi, B. (2016). An information-theoretic appraoch to assess linguistic complexity. In R. Baechler & G. Seiler (Eds.), Complexity, isolation and variation (p. 71-94). Berlin, New York: de Gruyter.

Futrell, R., Mahowald, K., & Gibson, E. (2015). Large-scale evidence of dependency length minimization in 37 languages. Proceedings of the National Academy of Sciences, 112(33), 10336-10341.

Joseph, J. E., & Newmeyer, F. J. (2012). All languages are equally complex: The rise and fall of a consensus. Historiographia Linguistica, 39, 341-368.

Juola, P. (1998). Measuring linguistic complexity: The morphological tier.

Journal of Quantitative Linguistics, 5, 206-250.

Juola, P. (2008). Assessing linguistic complexity. In M. Miestamo, K. Sinnemäki, & F. Karlsson (Eds.), Language complexity – typology, contact, change (Vol. 94, p. 89-108). Amsterdam, Philadelphia: John Benjamins Publishing Company.

Koplenig, A., Meyer, P., Wolfer, S., & Müller-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).

Mayer, T., & Cysouw, M. (2014). Creating a massively parallel bible corpus. In K. Choukri, D. T., L. H., B. Maegaard, & J. Mariani (Eds.), Proceedings of the ninth international conference on language resources and evaluation (p. 26-31). Reykjavik, Iceland: European Language Resources Association (ELRA).

McWhorter, J. H. (2001). The world's simplest grammars are creole grammars. Linguistic Typology, 5, 125-166.

Nichols, J. (1992). Linguistic diversity in space and time. Chicago and London: University of Chicago Press.

Sampson, G. (2009). A linguistic axiom challenged. In G. Sampson, D. Gil, & P. Trudgill (Eds.), Language complexity as an evolving variable (p. 1-18). Oxford, New York: Oxford University Press.

Trudgill, P. (2011). Sociolinguistic typology: Social determinants of linguistic complexity. Oxford University Press.