Word order correlations from a quantitative perspective

Matías Guzmán Naranjo - Laura Becker

Background Cross-linguistic tendencies of VO and OV languages have received a lot of attention since the seminal work by Dryer (1991). However, there have been few quantitative studies (Liu, 2010; Chen and Gerdes, 2017), and these only focus on one or a few languages (Celano, 2014). Most studies have insisted on finding a representative word order for every language, and finding correlations with other orders based on it. How to determine the basic word order for any language is a problem on itself, but even if one accepts this premise, it means that stricter OV order languages like Japanese are treated like flexible OV order languages like German. This means that despite of the strong correlations that have been found so far, we still have a poor understanding of how quantitative word order tendencies behave cross-linguistically. We want to explore the question of whether the fact that some languages allow for flexible orders between the verb and its arguments may have some influence on whether the language allows flexible or fixed orders between nouns and their dependents, and whether these tendencies are cross-linguistic or language specific.

Methodology Following recent studies (Croft et al., 2017), we will present a method based on dependency treebanks to investigate this issue. Using Universal Dependencies Treebanks (Nivre et al., 2016), we extracted for 60 languages, belonging to 14 different families, all verb-dependent relations and noun-dependent relations and their corresponding linear order. For verbs, we extracted dependents with the functions of: adverbial clauses, adverbial modifiers, subject, object and oblique object; while for nouns: adjectival clauses, adverbial clauses, adjectival modifiers, appositional modifiers, case markers, compound dependents, conjuncts, determiners, nominal modifiers, and numerals.

After normalization in parts per million, we obtain the number of times an object follows the verb and precedes it for every language, and similarly for all other factors. We will present several evaluation methods for this data, but the key findings can be seen by using a mixed-effect model. We do this by fitting a model where the predicted variable is `factor_follows - factor_precedes`, and the independent variables are all orderings with respect to the verb. We included language family as a random effect to control for family biases.

(Selected) Results Model performance is shown in Table 1. This table shows the marginal R2 (variance explained by fixed effects), and the conditional R2 (variance explained by all effects). We can observe several tendencies. First, not all word orders are equally predictable. The position of the numeral modifier with respect to the noun is more predictable than the position of an adverbial clause, or the dependent of a compound. Both these factors in turn are well predicted from the family, while the order of the noun with respect to an appositional modifier is not predictable from the family, and only slightly predictable overall. The difference in R2 between using all predictors, verb and subject, and verb and object clearly show that it is not only the latter that has an influence, but rather that all selected predictors seem to have an effect.

	All predictors		Only V and O		Only V and S	
	R2 marg	R2 cond	R2 marg	R2 cond	R2 marg	R2 cond
adjectival clause	0.459	0.860	0.349	0.832	0.133	0.664
adverbial clause	0.091	0.956	0.018	0.933	0.006	0.940
adjectival modifier	0.367	0.746	0.269	0.653	0.135	0.420
appositional modifier	0.306	0.310	0.135	0.141	0.051	0.101
case marker	0.629	0.880	0.410	0.744	0.233	0.575
compound dependent	0.135	0.934	0.022	0.899	0.098	0.914
conjunct	0.689	0.744	0.659	0.742	0.394	0.413
determiner	0.456	0.757	0.186	0.421	0.102	0.419
nominal modifier	0.479	0.894	0.310	0.592	0.148	0.469
numeral	0.703	0.703	0.415	0.697	0.223	0.692

Table 1: Model performance for each independent variable.

References

Dryer, Matthew S (1991). "SVO Languages and the OV: VO Typology". In: *Journal of Linguistics* 27.2, pp. 443–482.

Liu, Haitao (2010). "Dependency Direction as a Means of Word-Order Typology: A Method Based on Dependency Treebanks". In: *Lingua* 120.6, pp. 1567–1578.

Celano, Giuseppe GA (2014). "A Computational Study on Preverbal and Postverbal Accusative Object Nouns and Pronouns in Ancient Greek". In: *The Prague Bulletin of Mathematical Linguistics* 101.1, pp. 97–110.

Nivre, Joakim et al. (2016). "Universal Dependencies v1: A Multilingual Treebank Collection." In: *LREC*.

Chen, Xinying and Kim Gerdes (2017). "Classifying Languages by Dependency Structure. Typologies of Delexicalized Universal Dependency Treebanks". In: *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017), September 18-20, 2017, Università Di Pisa, Italy*. 139. Linköping University Electronic Press, pp. 54–63.

Croft, William et al. (2017). "Linguistic Typology Meets Universal Dependencies." In: *TLT*, pp. 63–75.