A usage-based approach to word order variation in Baltic, Slavic, Germanic and Romance

Basic word order is often assumed to bear signals of areal convergence and/or genealogical distance, but the extent to which word order actually varies within languages, areas and genealogical groupings is not usually taken into account. In a bottom-up approach, we assess the syntactic variation between 19 languages of four Indo-European branches (Baltic, Slavic, Germanic, Romance) in a set of 979 declarative sentences from a parallel corpus (*ParTy* corpus, Levshina 2017). Each sentence is categorized as either verb-initial, verb-medial or verb-final. For each language pair the log-transformed adjusted mutual information score is calculated, which ranks languages as similar to each other if the linearization in one language is predictable by the other language. With fuzzy clustering and variation partitioning, we evaluate the impact of areal and phylogenetic relations on the (dis)similarities of the languages.

Fuzzy clustering yields two clusters as the best solution for our data. The first cluster consists of Germanic languages and French (membership coefficients between 0.66 and 0.78). The second cluster contains Baltic, Slavic (except Russian) and Romance languages, with membership coefficients ranging from 0.61 to 0.74. Russian is almost equiprobably assigned to either of the two clusters with membership coefficients close to 0.5 (fig. 1).

The impact of phylogenetic and geographical constraints on the syntactic variation are estimated by variation partitioning (Legendre 2008). Geographical distances are calculated based on the language polygons provided by Ethnologue (Simons and Fennig 2018), with 1000 samples of random points drawn. For phylogenetic distances, we randomly samples 1000 trees from the posterior sample in Chang et al. (2015). After reducing the dimensions so that at least 80% of the variance is retained, the variation partitioning was calculated with a partial redundancy analysis for each predictor while controlling for the other predictor. The variation explained in both models is then partitioned between fractions which are exclusively attributed to each predictor and a third, shared fraction of variation.

A mean 60.17% of variation can be explained in total (mean adjusted r^2 = 0.6, standard deviation = 0.03). Phylogeny alone accounts for a median 0.22 fraction of variation (mad = 0.12) of variation, geography for a mean 0.03 (sd = 0.03). The fraction of variance shared by both predictors has a median of 0.35 (mad = 0.12). The fractions of phylogenetic and shared variation sum up to a mean of 0.58 (sd = 0.003) throughout all samples. In 763 out of 1000 samples, the fraction of variation attributed to geography exclusively is greater than zero (fig. 2). Thus, despite the substantial overlap with the phylogenetic signal, the spatial structure offers additional explanation for syntactic variation.

The results of both methods support the assumption of a strong phylogenetic signal in word order. In contrast to established linguistic areas like Standard Average European, our analysis supports only the assumption of areal convergence for the Germanic languages and French, which shows a low level of similarity to other Romance languages. Apart from that, no effect of areal convergence can be detected that is not consistent with a phylogenetic grouping.

References

Chang, Will, Chundra Cathcart, David Hall, and Andrew Garrett. 2015. "Ancestry-Constrained Phylogenetic Analysis Supports the Indo-European Steppe Hypothesis." *Language* 91 (1): 194–244.

Legendre, Pierre. 2008. "Studying Beta Diversity: Ecological Variation Partitioning by Multiple Regression and Canonical Analysis." *Journal of Plant Ecology* 1 (1): 3–8. https://doi.org/10.1093/jpe/rtm001.

Levshina, Natalia. 2017. "Online Film Subtitles as a Corpus: An *N*-Gram Approach." *Corpora* 12 (3): 311–38.

Simons, Gary F., and Charles D. Fennig. 2018. *Ethnologue: Languages of the World.* 21st ed. Dallas, Texas: SIL International. http://www.ethnologue.com.

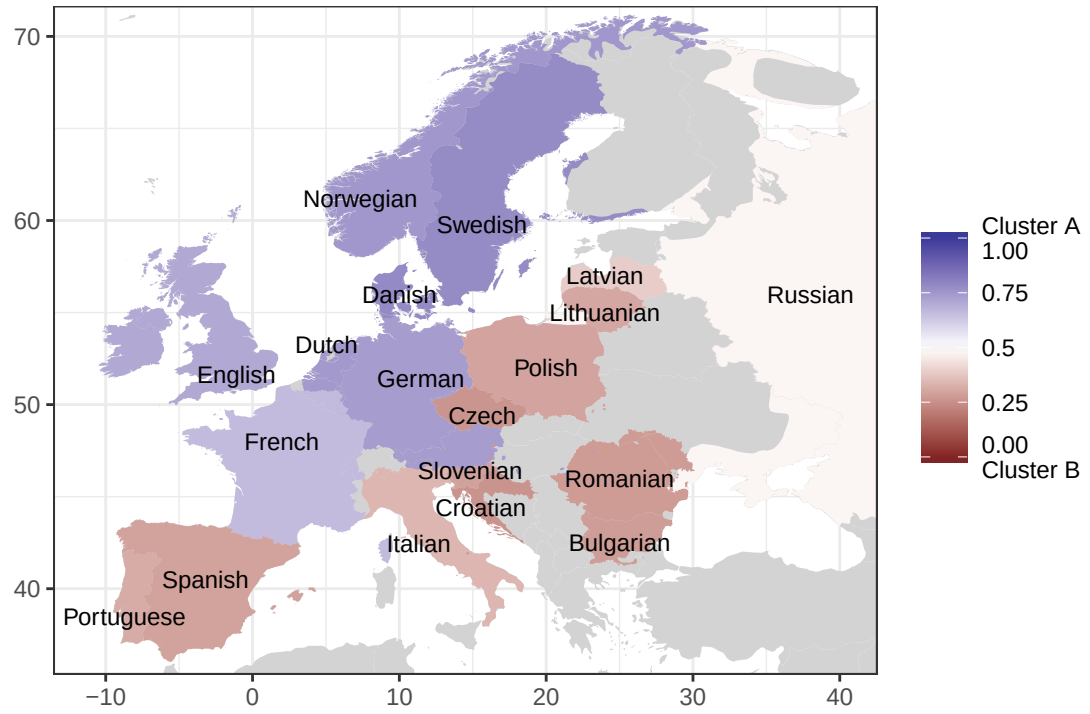

Figure 1: Map of fuzzy clustering coefficients. Language polygons according to Ethnologue (2018). Darker shades of red and blue indicate a high certainity of being assigned to one of the clusters. Lighter areas indicate equiprobable assignment to either of the two clusters. Gray areas indicate missing data.

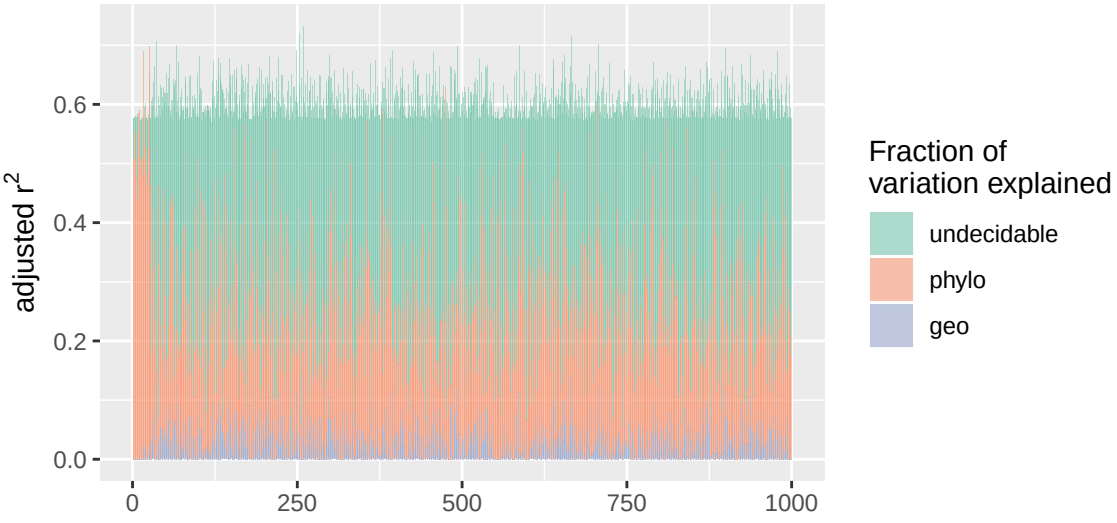

Figure 2: Stacked bar plot of adjusted r^2 values for each fraction of variation. Each bar represents a sample of random points and a posterior tree.