Patterns of non-basic word order in Indo-European

Christian Ebert, Balthasar Bickel, Paul Widmer

The Indo-European family is not consistent with regard to basic word order. While most of the subfamilies situated in Europe have SVO as basic word order, VSO can be found in Celtic and SOV in the Indo-Iranian branch. However, in all languages said to be SVO, deviations from the basic order are possible under certain conditions, and the focus on a basic pattern, however defined, fails to capture much of the actual diversity in corpora and introduces a strong bias into the assumptions we make about the evolution of word order patterns. Here, we take a different approach and explore the phylogenetic distribution of verb positions (verb-initial, verb-medial, vs. verb-final) in main clauses in corpus data from movie subtitles in four Indo-European branches in Europe (Romance, Baltic, Slavic, Germanic).

To investigate phylogenetic distributions, we examine verb-initial, verb-medial and verb-final main clauses in a sample of 21 languages from Tiedemann's OpenSubtitles2018 corpus (Lison & Tiedemann 2016), compute distance matrices from this (F_{st}, D etc.), and explore the tree-likeness we find in the data by calculating δ-scores and Q-residuals (Holland et al. 2002). In addition, we examine distances using NeighborNets (Bryant & Moulton 2004).

As for verb-initial clauses, no overlap with the genealogical classification can be distinguished since the values are distributed in a similar pattern amongst the subbranches (Romance: mean = 0.05, sd = 0.04; Germanic: mean = 0.08, sd = 0.02; Balto-Slavic: mean = 0.07, sd = 0.05), whereas for verb-final clauses, the clustering intersects with genealogical subgroups. The Balto-Slavic group on the one side and the Germanic and Romance groups on the other contrast in their proportions of verb-finality (Romance: mean = 0.05, sd = 0.04; Germanic: mean = 0.02, sd = 0.02; Balto-Slavic: mean = 0.14, sd = 0.06).

δ-scores and Q-residuals suggest that the data can be interpreted as having a moderately tree-like structure ($\delta = 0.24$, $Q = 0.05$), suggesting a weak phylogenetic signal in the data. As for the overall distance structure, for the Romance and Germanic languages there is no observable clustering corresponding to genealogical groups, suggesting effects of sustained contact. Almost all Balto-Slavic languages are grouped together, leaving only Croatian placed in between Germanic languages, a phenomenon that needs further investigation. Slovenian and Romanian appear in what looks like a transition zone between the Balto-Slavic and the Germanic/Romance groups, again suggesting effects of contact. Baltic is surprisingly split, with Lithuanian being close to Russian and Latvian being an outlier to the whole group (with a high proportion of verb-finality). Further analysis is needed to explain this split.

So far, our analysis of corpus data appears to well capture both phylogenetic and contact-induced patterns and paves the way for understanding word order evolution with much higher resolution than the classic basic word order approaches would allow.

References

Bryant D. and Moulton V. 2004: *Neighbor-Net: an agglomerative method for the construction of phylogenetic networks.* In Molecular Biology and Evolution 21, 255–265.

Holland B. R., Huber K. T., Dress A. and Moulton V. 2002: *δ Plots: a tool for analyzing phylogenetic distance data.* In Molecular Biology and Evolution 19, 2051–205.

Lison P. and Tiedemann J., 2016, *OpenSubtitles2016: extracting Large Parallel Corpora from Movie and TV Subtitles.* In Proceedings of the 10th International Conference on Language Resources and Evaluation, 923–929.