A multifactorial approach to crosslinguistic constituent orderings

Zoey Liu (Boston College)

Motivation Hawkins (2014) proposed crosslinguistic syntactic variation is a multifactorial process shaped by processing efficiency, that ordering preferences are driven by several competing and cooperating factors simultaneously. Nevertheless, this proposal still lacks proper quantitative support, as most previous studies have focused on a limited set of factors and languages. Their findings are not directly comparable as most experiments have examined syntactic constructions that do not necessarily allow flexible orderings. These limitations mean that it is not currently clear what the best typological determinants are for syntactic orders across languages.

Current study We aim to bridge this gap with the double adpositional phrase (PP) construction as the test case (Liu, 2020), using multilingual corpora from Universal Dependencies (Zeman et al., 2020). We searched for verb phrases (VP) in which the head verb has two PP dependents occurring on the same side (*She **danced*** [PP1 *with the band*] [PP2 *at the dinner party*]), the order of which allows flexibility in at least some contexts. Preprocessing yielded an initial dataset of 40 languages (33 ended up being Indo-European (IE)). The PP orderings for these languages fall into three different patterns: (1) one for languages with only preverbal PP orders (e.g. Hindi); (2) one for languages with only postverbal PP orders (e.g. Greek); (3) one for languages with both preverbal and postverbal orders (e.g. Czech). A subset of 20 languages was then selected based on data availability, language family and genus coded following The World Atlas of Language Structures (Dryer and Haspelmath 2013), as well as their observed PP ordering pattern in corpora, which included fifteen IE, one Sino-Tibetan (Chinese), one Japanese (Japanese), one Austronesian (Indonesian), and two Afro-Asiatic (Arabic and Hebrew).

Measures We investigated the roles of four theoretically motivated constraints that have been shown to affect syntactic alternations or reflect processing complexity: (1) dependency length, measured as the linear distance between the head verb and the adposition of each PP; (2) semantic closeness, calculated as the semantic similarity between the verb and the lexical head of the PP, using fastText word embeddings (Bojanowski et al., 2017) and cosine similarity; (3) lexical frequency, which was the product of the probability of the adposition and just the lexical head to separate the contribution of phrasal length and frequency; frequency counts were taken from the Python package wordfreq; (4) contextual predictability, which was the product of the conditional probability of the adposition and the lexical head given preceding sentential context; conditional probability was estimated with neural long-short term memory models trained for each language using large-scale texts from Ginters et al. (2017). Specifically, we examined whether there is a typological tendency for the PP that is shorter, or semantically closer to be closer to the verb, and for the more frequent or the more predictable PP to appear first. To better handle issues of missing data, we eventually fit the same model architecture to every language: the order of the two PPs in each VP as the dependent variable, the four factors along with pronominality of each PP as fixed effects and the head verb as a random effect. The predictive power of each factor was evaluated with Bayesian mixed-effects models.

Results Overall, dependency length is the strongest predictor and it is more effective in postverbal than preverbal domains. In certain preverbal cases where dependency length is not effective, semantic closeness and lexical frequency play a weak role. By contrast, contextual predictability does not seem to have a consistent effect across languages.

In each figure, for better representation, statistical significance is indicated by colors: red triangle represents the factor in question has a significant positive effect; green square indicates the factor has a significant negative effect; blue dot means the factor has no effect. 95% confidence intervals for each factor were derived from their respective posterior distributions in the Bayesian regression.

Figure 1: Coefficients for the four factors in languages with only preverbal PP orderings. We included Hindi due to its typologically distinct features, yet without calculating its effect of contextual predictability due to limited training data.

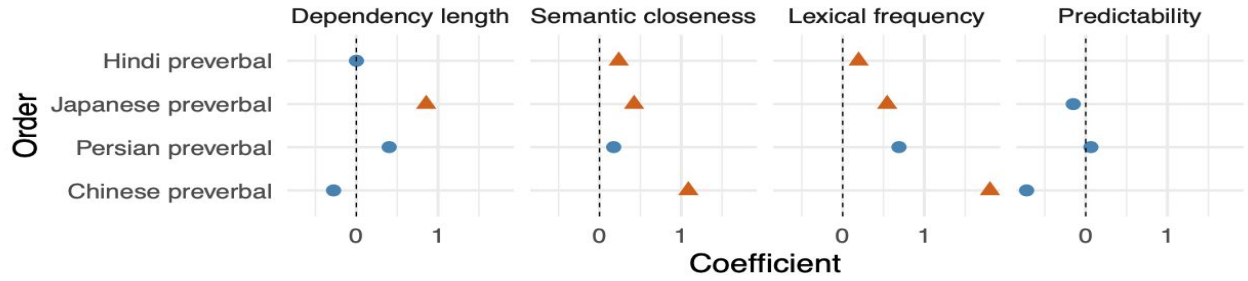

Figure 2: Coefficients for the four factors in languages with only postverbal PP orderings.

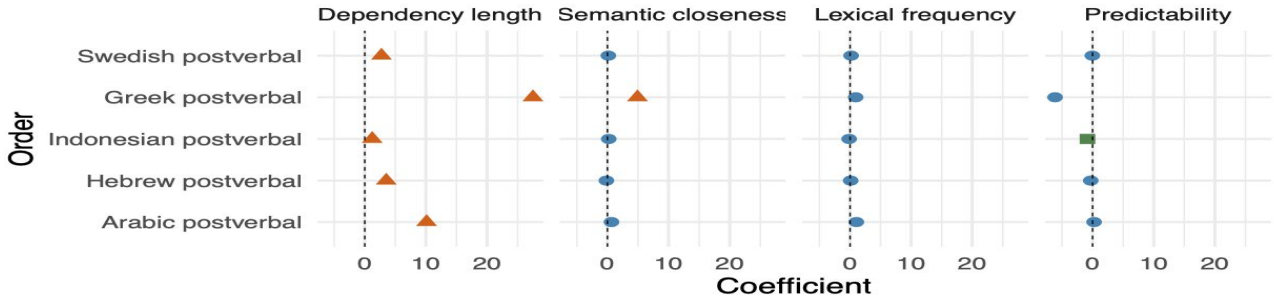

Figure 3: Coefficients for the four factors in languages with both preverbal and postverbal PP orderings.