Preferred Argument Structure is preferred in a typologically diverse corpus of 54 languages

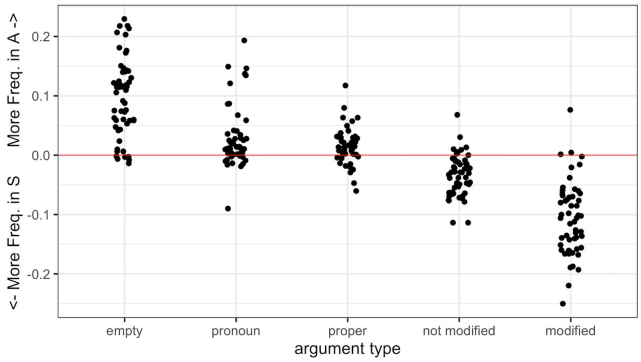
The theory of Preferred Argument Structure (PAS) states that transitive subjects (A, in morphosyntactic notation) are less likely slots for new lexical content than intransitive subjects (S) and transitive objects (O) (Du Bois, 1987; Du Bois et al., 2003). One explanation is that it is cognitively taxing to include too many core arguments in the same clause, and so new content should be introduced either in object position or in the subject position of an intransitive clause, but not in the subject position of a transitive clause.

While this phenomenon has been explored in various spoken-language corpora, there is debate as to its empirical structure and underpinnings. For instance, Everett (2009) argues that it follows from pragmatic constraints for animate agents. Because subjects precede objects in most languages, the subject/object asymmetry posited by PAS is also consistent with psycholinguistic theories that given information precedes new information (Givon, 1984) and that easy material should come first (MacDonald, 2013). To develop the empirical landscape of PAS cross-linguistically and evaluate whether it is indeed a universal feature of language, we conducted a large-scale analysis using Universal Dependencies treebanks (a typologically rich treebank of dependency-parsed trees from 54 languages (11 families) for which we could obtain 1000 sentences fitting our criteria). Previous quantitative analysis of PAS has considered only a handful of languages.

We hypothesize that, across languages, there will be less lexical content in A, relative to S or O. Because we work with full parses, we do not limit ourselves to a binary distinction between lexical and non-lexical arguments but can use an accessibility continuum, as shown below. We provide large-scale quantitative evidence for PAS and illuminate its underpinnings.

Method

We parsed Universal Dependencies 2.5 treebanks using the `conllu` library. We classified the core arguments for each root verb (excluding copulas) into 5 categories of decreasing accessibility: empty, pronoun, proper noun, lexical item with no modification other than a determiner, modified lexical item. The empty category and the pronoun category are considered "non-lexical" in our binary analysis and the rest is lexical.



Results

We broadly found the "avoid lexical A" pattern across languages, with S intermediate between A and O. For 93% of languages in our sample, O was more likely to contain a lexical argument than S, and S was more likely to contain a lexical argument than A. We also can use finer grained distinctions than lexical/non-lexical. The plot to the right shows, for each language, the difference between the proportion of the time A falls into a particular category, as compared to S. The downward trend from left to right shows that, in the majority of languages, A is more likely to be empty or a pronoun than S is (i.e., in those columns most points fall above 0). On the other extreme (right of the graph), we see that S is more likely to be modified than A is (most points below 0 in the modified column). Proper nouns and unmodified lexical items are intermediate.

Assessing significance by fitting a logistic maximal mixed effect model with a random effect for language, we found a significant difference (β=.59, $p < .0001$) between A and S in probability of containing a lexical argument. O was even more likely to consist of a lexical argument, and more likely to have that argument modified.

While we found a gradient of A < S < O in probability of containing a lexical argument, a second possible prediction of PAS does not have clear support in our sample. PAS's prescription to avoid more than one lexical core argument implies that, if the subject is lexical, the object is less likely to be lexical. But, in transitive sentences, we did not observe a significant change in the likelihood that the object would be lexical by conditioning on the status of the subject. Further exploration suggests that this may be due to genre differences. That is, some genres tend to use more pronouns than others and so, in those genres, we observe more pronouns in both argument positions for transitive sentences. More work is needed to understand this possibility further.