Accessibility-Based Constraints on Morphosyntax in Corpora of 54 Languages

Kyle Mahowald (UC Santa Barbara), Isabel Papadimitriou (Stanford University),
Dan Jurafsky (Stanford University), Richard Futrell (UC Irvine)

Introducing new information into a clause is cognitively costly and is often restricted to specific linguistic environments, a factor in many sentence processing models (Givón, 2001; Arnold, et al. 2003; MacDonald, 2013). Here we investigate a particular aspect of these costs: the difference between transitive and intransitive subjects. The theory of Preferred Argument Structure (PAS) predicts that new lexical content is less likely to appear as transitive subjects (A, in morphosyntactic notation) than intransitive subjects (S) and transitive objects (O) (Du Bois, 1987; Du Bois et al., 2003). Specifically, the claim is that the transitive subject (A) is a dispreferred location for new information since the object also often introduces new information, and it is cognitively costly to introduce new information in two core argument slots at the same time. Here, we operationalize these constraints in terms of referential form, which has been argued to correlate with accessibility in production (Ariel, 2001). The hypothesis is that more accessible nominals (null arguments, pronouns, proper nouns) are more likely to occur in A argument positions, whereas less accessible nominals (nominals with determiners, modified nominals) are more likely to occur in S and O positions. We run a reproducible, large-scale, cross-linguistic analysis, to evaluate the extent to which these claims about subjecthood and accessibility constitute a universal feature of language.

Our main experimental contribution consists of using the Universal Dependencies corpus of 54 languages from 11 families to extract and correlate two pieces of information about core verb arguments: (a) the accessibility of the argument , and (b) whether it is A, S or O. We investigate accessibility rather than the new/given information distinction because there are very few corpora across languages annotated specifically for information structure. We use UD annotations to classify core verb arguments into five classes of decreasing accessibility: empty subjects, pronouns, proper nouns, lexical items (with no modification other than a determiner), and modified lexical items. For (b), we use the Universal Dependencies parses to determine whether each argument is an A, S, or O.

We found that accessibility asymmetries between A, S, and O broadly hold across languages. Transitive subjects (A) are the least likely to be lexical, followed by intransitive subjects (S), and transitive objects (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 show a fine-grained breakdown of our results for A and S in Figure 1, comparing how likely it is for arguments in different accessibility classes to appear as A rather than S. Each point in the graph indicates a different language. The downward trend from left to right for all languages shows that more accessible items (empty or pronouns) are more likely to be A, while less accessible items (eg. modified lexical items) are more likely to be S. Assessing significance by fitting a logistic maximal mixed effect model (predicting whether the argument is lexical as a function of argument role) with a random effect for language, we found a significant difference ($\beta = .59$, $p < .0001$) 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.

Overall, we show that, cross-linguistically, less accessible (or newer) information is more likely to appear as S than A, and most likely to appear as O. Moreover, our experimental method is easily reproducible and generalizable to more languages. While previous support for theories such as Preferred Argument Structure relied on small, spoken corpora of a handful of languages, we hope that our analysis can lay the groundwork for supporting the empirical cross-lingual universality of such claims about information processing.

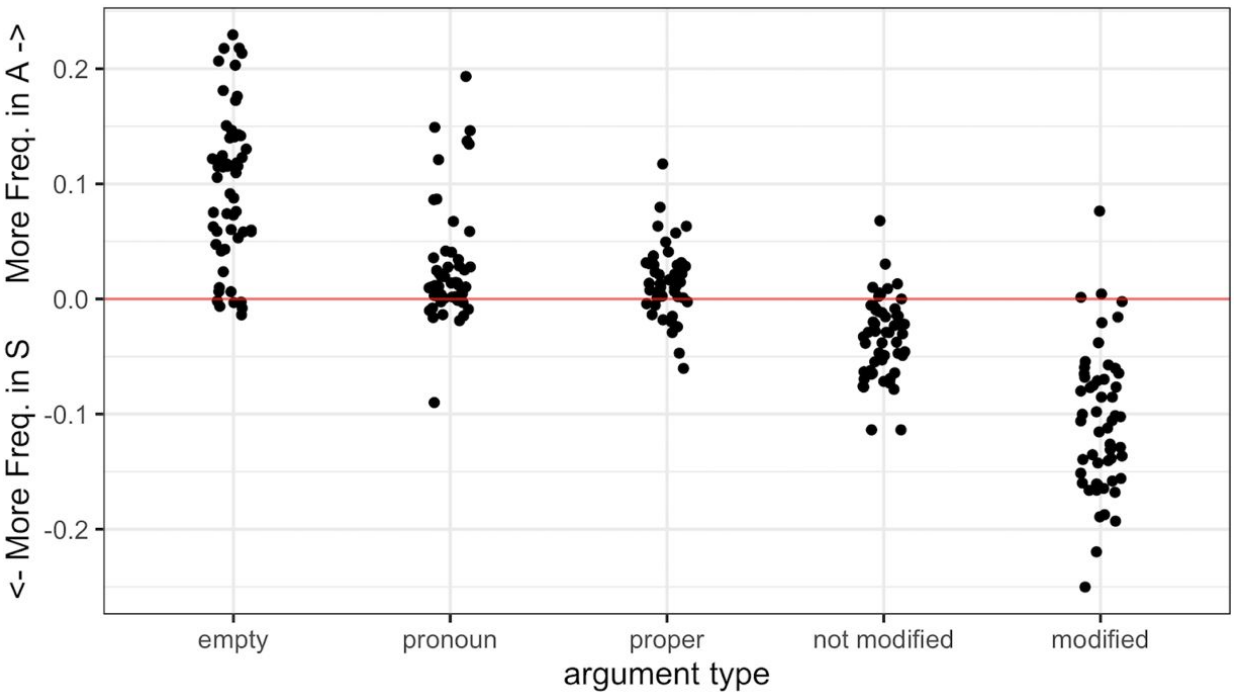

Figure 1 For each argument type, its relative frequency in A relative to S. Each dot is a language. Across languages, empty subjects are more common as transitive subjects, whereas modified subjects are more common in intransitive subjects.

References

Ariel, M. (2001). Accessibility theory: An overview. Text representation: Linguistic and psycholinguistic aspects, 8, 29-87.

Arnold, J. E., Fagnano, M., & Tanenhaus, M. K. (2003). Disfluencies signal theee, um, new information. Journal of Psycholinguistic Research, 32(1), 25-36.

Du Bois, J., Kumpf, L., & Ashby, W. (2003). *Preferred argument structure: Grammar as architecture for function*. (Vol. 14) John Benjamins Publishing.

Du Bois, J. (1987). The discourse basis of ergativity. *Language*, 805–855.

Givón, T. (2001). *Syntax: an introduction*. (Vol. 1) John Benjamins Publishing.

MacDonald, M. (2013). How language production shapes language form and comprehension*Frontiers in psychology, 4*, 226.