Quantifying parsing complexity as a function of grammar complexity

Jonathan Brennan (Children's Hospital of Philadelphia), Sarah VanWagenen (Haifa & University of California, Los Angeles), & Edward P. Stabler (University of California, Los Angeles)

brennanj3@email.chop.edu

Syntactic complexity; Computational modeling; Narrative; English

Introduction Recent research has used measures of complexity generated by computational models to predict human processing effort as measured in behavioral and neurophysiological experiments, ([2,4,5,6,9] among others). For ostensibly practical reasons, these models often rely on simplifying grammatical assumptions. For example, many are defined over structures from the Penn Treebank II which, while an impressive annotation scheme, is still far from the grammars that syntacticians propose to account for patterns in natural language. Though there is emerging consensus that natural languages are at least mildly context sensitive ([12,14]), the dominant, if implicit assumption in both computational and psycholinguistics is that simplified context free grammars are adequate approximations for theories of processing (but cf. [7]). However, while significant research has examined the degree to which different parsing strategies match human parsing performance (e.g., [1,10,13]), the effect of grammar (even within a particular class) on estimates of processing complexity has not been systematically investigated.

Methods We directly compared the impact of grammar on estimated processing complexity for a 12 minute English narrative. We used two grammars: One generated by an automated parser ([3]) trained on Wall Street Journal corpora annotated according to the Penn Treebank 2 schema (hereafter Penn), and one manually constructed by a trained linguist in the Minimalist Grammar framework ([11]) and represented as an X-bar schema (hereafter Xbar). The XBar grammar includes head movement (making it non-context free), A and A-bar movement, left and right adjunction, vp and pp shells, and Kaynian relative clauses ([8]). The impact of grammar was contrasted against the impact of parsing strategy using two stack-based strategies: Top-Down (TD) and Bottom-Up (BU). A TD strategy builds the root node first and is fully incremental and predictive whereas the BU strategy makes no predictions at all. Rather, it waits for all daughters to be complete before projecting a mother node. For each lexical item n, we calculated the number of nodes built between items n and $n - 1$ as our complexity measure.

Results Across parsing strategies, the incremental complexity of the Penn grammar is not systematically correlated with the complexity of the Xbar grammar ($r_{TD} = 0.14$, $r_{BU} = -0.19$), suggesting that grammatical assumptions have a large impact on parsing complexity. Furthermore, we observe an interaction between parsing strategy and grammar choice such that TD and BU measures over the Xbar grammar are highly correlated ($r_{Xbar} = 0.64$), i.e., make similar predictions for human processing effort, but make very different predictions over the Penn grammar ($r_{Penn} = 0.12$).

Conclusion The degree to which parsing strategies make different predictions for human processing effort is in a large part a function of grammar choice. Indeed, in the fragment of English we considered, the impact of grammar choice matches or exceeds the impact of parsing strategy on estimates of processing complexity. This finding shows that even for a single, fixed corpus and a fixed parsing strategy, the choice of grammar is extremely important: a treebank grammar predicts different, and under some circumstances opposite, parse complexities from a standard Xbar grammar.

References

[1] S Abney and M Johnson. *J Psycholing Res*, 20(3):233–250, 1991. [2] A Bachrach. PhD thesis, MIT, 2008. [3] DM Bikel. In *Proceedings of the 2nd HLT Conference*, 2002. [4] JR Brennan. PhD thesis, NYU, 2010. [5] V Demberg-Winterfors. PhD thesis, University of Edinburgh, 2010. [6] JT Hale. PhD thesis, Johns Hopkins University, 2003. [7] JT Hale. *Cognitive Science*, 30(4), 2006. [8] RS Kayne. *The Antisymmetry of Syntax*. MIT Press, 1994. [9] R Levy. PhD thesis, Stanford University, 2005. [10] P Resnik. In *Proceedings of the 14th COLING Conference*, pages 191–197, 1992. [11] EP Stabler, *Derivational Minimalism*, pages 68–95. Springer, 1997. [12] EP Stabler. *Oxford Handbook of Linguistic Minimalism*, pages 616–641. OUP, 2010. [13] P Sturt and V Lombardo. *Cognitive Science*, 29(2):291–305, 2005. [14] K Vijay-Shanker, DJ Weir, and AK Joshi. In *Proceedings of the 25th Annual Meeting of the ACL*, pages 104–111, 1987.