

A Universal Prior Distribution over Syntactic Constructions Emerges from Incremental Language Production

Fermín Moscoso del Prado Martín
University of Cambridge

Human language production is long known to proceed incrementally: Speakers do not plan complete utterances before articulation. Instead they construct sentences online while lexical retrieval and syntactic encoding is unfolding dynamically [2, 5]. Contemporary models of production assume that lexical items become available sequentially through lexical selection processes and are progressively integrated into an emerging syntactic “draft” representation. Under this view, syntactic structure is assembled step by step as words are incorporated into a partially constructed dependency graph.

Such incremental construction process has important statistical consequences for the distribution of syntactic constructions. Some recent approaches to language structure argue that the pressures arising from language production constitute the most crucial determinant of linguistic structure itself [6]. In the limiting case in which words enter the developing structure strictly one at a time, newly selected lexical items will preferentially attach to elements that are already structurally central simply because these were introduced earlier in the utterance. Such dynamics would naturally produce a rich-get-richer process, analogous to *preferential attachment* mechanisms widely studied in network science [1]. Earlier lexical items consequently accumulate disproportionately many syntactic dependencies, generating a highly skewed distribution over attachment probabilities and structural configurations. However, language production is also shaped by powerful reuse mechanisms. Speakers systematically recycle recently processed syntactic configurations, both from their own previous utterances and from structures encountered in interlocutors’ speech, through syntactic priming and related adaptation processes [3, 9]. The reuse of partially instantiated structural frames attenuates the purely linear preferential attachment process predicted by strictly sequential assembly. Instead, syntactic growth is expected to follow a *sublinear preferential attachment (sPA)* [4] regime in which structural prominence still increases with prior connectivity, but with diminishing returns due to the recurrent reuse of established constructional patterns.

If syntactic dependency graphs are assumed to grow through sPA, the expected probability of an entire syntactic dependency structure can be derived analytically *a priori*, without requiring any language-specific training data or corpus-based parameter estimation: The probability of a sentence follows directly from the incremental dynamics of structure construction itself, yielding a generative prior distribution over syntactic constructions grounded in psycholinguistic mechanisms of language production. I analyse dependency structures from corpora spanning more than 180 languages represented in the Universal Dependencies project [7, 8]. Across this typologically diverse sample, I find that observed sentence probabilities correlate significantly with the prior probabilities predicted by sPA. The results support that large-scale statistical regularities in syntactic organisation can emerge directly from universal constraints on incremental production and structural reuse. These findings provide evidence for a universal, data-free prior probability distribution over syntactic constructions that applies across languages. More broadly, they support production-based accounts of linguistic structure [6].

References

- [1] Barabási, A.-L., & Albert, R. (1999). Emergence of scaling in random networks. *Science*, 286(5439), 509–512.
- [2] Bock, J. K. (1982). Toward a cognitive psychology of syntax: Information processing contributions to sentence formulation. *Psychological Review*, 89(1), 1–47.
- [3] Bock, J. K. (1986). Syntactic persistence in language production. *Cognitive Psychology*, 18(3), 355–387.
- [4] Krapivsky, P. L., Redner, S., & Leyvraz, F. (2000). Connectivity of growing random networks. *Physical Review Letters*, 85(21), 4629–4632.
- [5] Levelt, W. J. M. (1989). *Speaking: From Intention to Articulation*. Cambridge, MA: MIT Press.
- [6] MacDonald, M. C. (2013). How language production shapes language form and comprehension. *Frontiers in Psychology*, 4, 226.
- [7] de Marneffe, M.-C., Manning, C. D., Nivre, J., & Zeman, D. (2021). Universal Dependencies. *Computational Linguistics*, 47(2), 255–308.
- [8] Nivre, J., de Marneffe, M.-C., Ginter, F., Hajič, J., Manning, C. D., Pyysalo, S., Schuster, S., Tyers, F., & Zeman, D. (2020). Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the 12th Language Resources and Evaluation Conference*, 4034–4043.
- [9] Pickering, M. J., & Ferreira, V. S. (2008). Structural priming: A critical review. *Psychological Bulletin*, 134(3), 427–459.