Explaining morphosyntactic structure with inter-predictability

John Mansfield and Charles Kemp

Where does morphosyntactic structure come from? How do words and morphemes get grouped into constituents, that are spoken or signed in the elegant-yet-messy patterns we call grammar? In this presentation we focus on linear sequences in grammar (as opposed to hierarchical structure), and implement a computational model of how these linear structures might emerge based on relations of INTER-PREDICTABILITY between words or morphemes.

Several recent studies show that some linear orderings exhibit a striking isomorphism with ranked inter-predictability. Elements that are more inter-predictable are closer in linear order. For example, Culbertson and colleagues (2020) focus on a cross-linguistic bias towards particular word orders within noun phrases, whereby some order such as [N Adj Num Dem] and [Dem Adj N Num] are very common, while others such as [Num Adj Dem N] are very rare. They show that this bias aligns strongly with the ranked inter-predictability between nouns and other word types, measured as pointwise mutual information (PMI). More proximate elements such as [N Adj] have higher average PMI than less proximate elements such as [N Dem], and this statistical pattern holds for each of 24 languages surveyed. A similar result has been shown for the linear proximity of agglutinative affixes to stems in six languages (Hahn et al. 2022).

These studies suggest some fundamental relationship between linear proximity and inter-predictability, and in this presentation I explore the nature of this relationship. I test whether inter-predictability could plausibly cause the emergence of linear orders, using NPs as a case study. I implement a simple computational model that starts from unordered symbolic elements of the categories {N, Adj, Num, Dem}, and gradually converges on consistent categorical orderings, driven by inter-predictability. The key mechanism in the model is a version of 'chunking', whereby more inter-predictable combinations of symbols are stored as holistic units (cf. Arnon & Snider 2010; Christiansen & Chater 2015), on the assumption that this produces efficiencies in the storage-and-retrieval system (Wray 2002; Wray 2017). Additionally, a bias towards consistent ordering by grammatical category (e.g. Mansfield et al. 2020; Mansfield et al. 2022), means that specific inter-predictabilities such as {cat, black} influence whole classes of elements such as {N, Adj}.

The model produces NP-internal orderings that closely match those attested in natural languages (Dryer 2018). This suggests at least prima facie plausibility for NP linear ordering being generated largely by inter-predictability, without any recourse to an underlying hierarchical structure. This is in line with some theories that posit 'flat' constituent structures made up of linear sequences (e.g. Simpson & Withgott 1986; Culicover & Jackendoff 2005; Good 2016), as opposed to maximally deep hierarchies with strict binary branching (e.g. Kayne 1994). At the same time, the model is much simpler than the cultural evolutionary processes that characterise language change (e.g. Blythe & Croft 2021), and therefore many questions remain open, including whether this model is compatible with other efficiency principles such as domain minimisation (Hawkins 2004) and uniform information density (Levy & Jaeger 2007).

References

Arnon, Inbal & Snider, Neal. 2010. More than words: Frequency effects for multi-word phrases. *Journal of Memory and Language*. Netherlands: Elsevier Science 62(1). 67–82. (doi:10.1016/j.jml.2009.09.005)

Blythe, Richard A. & Croft, William. 2021. How individuals change language. *PLOS ONE*. Public Library of Science 16(6). e0252582. (doi:10.1371/journal.pone.0252582)

Christiansen, Morten H. & Chater, Nick. 2015. The now-or-never bottleneck: A fundamental constraint on language. *The Behavioral and Brain Sciences* 39. 1–52.

Culbertson, Jennifer & Schouwstra, Marieke & Kirby, Simon. 2020. From the world to word order: Deriving biases in noun phrase order from statistical properties of the world. *Language* 96(3). 696–717.

Culicover, Peter W. & Jackendoff, Ray. 2005. *Simpler syntax*. Oxford: Oxford University Press.

Dryer, Matthew S. 2018. On the order of demonstrative, numeral, adjective, and noun. *Language*. Linguistic Society of America 94(4). 798–833.

Good, Jeff. 2016. *The linguistic typology of templates*. Cambridge: Cambridge University Press.

Hahn, Michael & Mathew, Rebecca & Degen, Judith. 2022. Morpheme ordering across languages reflects optimization for memory efficiency. *Open Mind: Discoveries in Cognitive Science (Accepted)*.

Haspelmath, Martin. 2011. The indeterminacy of word segmentation and the nature of morphology and syntax. *Folia Linguistica* 45(1). 31–80.

Hawkins, John A. 2004. *Efficiency and complexity in grammars*. Oxford: Oxford University Press. (doi:10.1093/acprof:oso/9780199252695.001.0001)

Kayne, Richard S. 1994. *The antisymmetry of syntax*. Cambridge, MA: MIT Press.

Levy, Roger & Jaeger, T. Florian. 2007. Speakers optimize information density through syntactic reduction. In Schölkopf, B. & Platt, J. C. & Hoffman, T. (eds.), *Advances in Neural Information Processing Systems 19*, 849–856. MIT Press. (http://papers.nips.cc/paper/3129-speakers-optimize-information-density-through-syntactic-reduction.pdf) (Accessed October 24, 2019.)

Mansfield, John & Saldana, Carmen & Hurst, Peter & Nordlinger, Rachel & Stoll, Sabine & Bickel, Balthasar & Perfors, Andrew. 2022. Category Clustering and Morphological Learning. *Cognitive Science* 46(2). e13107. (doi:10.1111/cogs.13107)

Mansfield, John Basil & Stoll, Sabine & Bickel, Balthasar. 2020. Category clustering: A probabilistic bias in the morphology of argument marking. *Language* 96(2). 255–293.

Simpson, Jane & Withgott, M. 1986. Pronominal Clitic Clusters and Templates. In Borer, H. (ed.), *The syntax of pronominal clitics* (Syntax and Semantics), vol. 19. New York: Academic Press.

Tallman, Adam J. R. 2021. Constituency and coincidence in Chácobo (Pano). *Studies in Language* 45. 321–383.

Wells, Rulon S. 1947. Immediate constituents. *Language*. Linguistic Society of America 23(2). 81–117. (doi:10.2307/410382)

Wray, Alison. 2002. *Formulaic language and the lexicon*. Cambridge: Cambridge University Press.

Wray, Alison. 2017. Formulaic sequences as a regulatory mechanism for cognitive perturbations during the achievement of social goals. *Topics in Cognitive Science* 9(3). 569–587. (doi:10.1111/tops.12257)