Emergence and learnability of dependency structures in an artificial language

Long-distance dependencies between words (e.g. nouns and verbs, nouns and adpositions) are an essential aspect of human language, and can be arranged in different ways. In the following examples (based on Vosse & Kempen 1991), the dependency between subject and verb is indicated by subscripts:

(1) Branching (right-branching) dependencies in English:
John$_1$ saw$_1$ Peter$_2$ walk$_2$

(2) Center-embedded dependencies in German:
…daß Hans$_1$ Peter$_2$ laufen$_2$ sah$_1$

(3) Crossed dependencies in Dutch:
…dat Jan$_1$ Peter$_2$ sag$_1$ lopen$_2$

Most natural languages predominantly use some combination of branching and center-embedded dependency structures; crossed dependencies are rare in natural languages, but some examples are attested, e.g. in Dutch subordinate clauses (Bresnan et al. 1982) like (3). Center-embedding tends to be less prevalent than branching both within and across languages, likely because of the cognitive difficulties it presents (see e.g. Kuno 1974, Hawkins 2004). Some experimental evidence from natural language processing (Bach et al. 1986) and nonlinguistic sequence learning (Öttl et al. 2015) also suggests that crossed dependencies are easier to process and learn than center-embedded dependencies, which is surprising in light of the marginal status of crossed dependencies in natural languages.

Artificial language learning experiments can shed light on the cognitive factors that shape natural language structures (e.g. Culbertson 2012, Fedzechkina 2018), and the regularization and stabilization of irregular input (e.g. Hudson Kam & Newport 2005). We applied these methods to the three types of dependencies. In two experiments, we investigated the learnability and emergence of all three dependency types within the same paradigm, using meaningful sequences (descriptions for visual configurations of objects) constituting a small artificial language including nouns and spatial adpositions.

Experiment 1 concerned the learnability of the three dependency types. Participants (n = 120, all English speakers) were trained on a simple artificial language with one of these possible structures and either head-initial or head-final order. First they viewed individual objects, and then configurations of these same objects arranged in a particular spatial relation, with a corresponding description. Participants were then asked to label configurations of objects with the vocabulary of that language; these included new configurations not seen in training, to address how well the artificial language had been generalized. Participant-created labels were analyzed for accuracy to training strings, using edit distance as a metric (Figure 1) and for adherence to one of several possible grammars (by generating strings in these grammars, and finding the grammar which generated the description closest in edit distance to the participant's description). As expected, learning accuracy on the mean of center-embedded and crossed dependencies was significantly lower than that for branching structures as shown by a linear mixed effects model ($B = -0.02$, $SE = 0.001$, $p = .04$), for both word orders. Performance on crossed grammars was not significantly different from performance on center-embedding ($B = -0.01$, $SE = 0.02$, $p = .51$). However, 6 participants of the 40 trained on center-embedded grammar produced labels best matching a crossed grammar, suggesting that they had reanalyzed the input to fit a new grammar. Conversely, participants trained on crossed grammar did not generate strings matching center-embedded grammars.

Experiment 2 employed the method of *iterated learning* (Kirby et al. 2008, Beckner et al. 2017) to examine the emergence and stabilization of consistent grammar, using stimuli of the same type as in the first experiment. The first-generation participants were trained on an artificial language with no consistent word order, and asked to label visual stimuli using the vocabulary. Their productions were passed on as input to another group, and so on. Twelve chains of five generations each were run (n = 60, English speakers). In most chains, branching grammar predominantly appeared within a few generations and was stable once it appeared, although one chain stabilized to contain strings which most closely matched a crossed grammar (Figure 2). Strings best matching center-embedding did not emerge in any chain.

The results of these experiments are consistent with long-standing psycholinguistic findings concerning the difficulty of center-embedded dependencies. However, they also support the possibility that the crossed dependency structure rarely seen in natural languages is as learnable as, or even more learnable than, the widely-found center-embedded structure. Of particular interest is that strings consistent with) crossed dependency grammar emerged both as reanalyses of center-embedded input in Experiment 1, and as a result of iterated learning (in one chain) from initially unstructured input in Experiment 2. These findings imply that natural language typology may not be straightforwardly related to learnability.

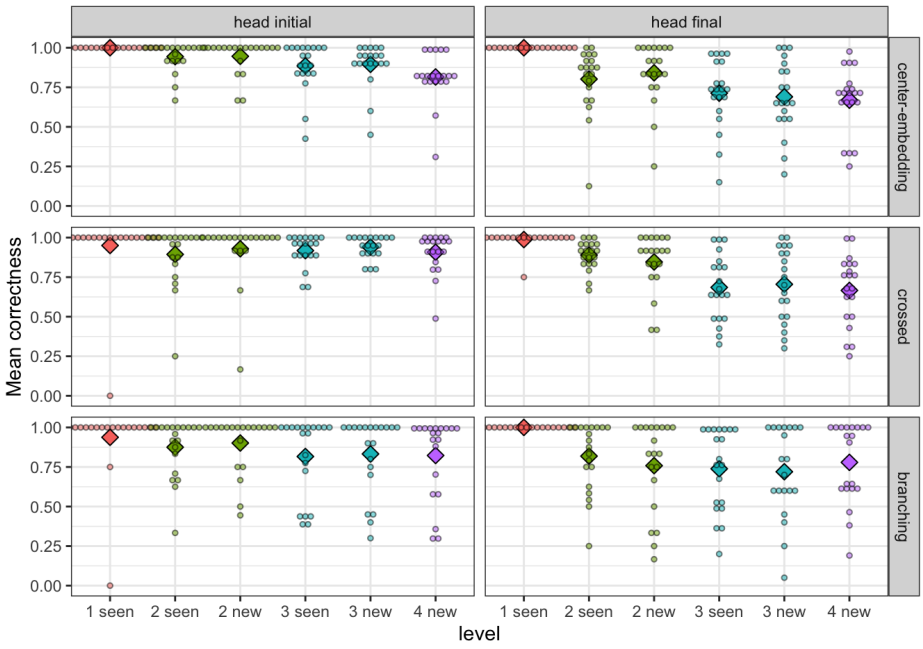

Figure 1: Production accuracy by level (number of objects in the scene) and grammar type. “Seen” indicates items seen in training; “new” are novel test items.

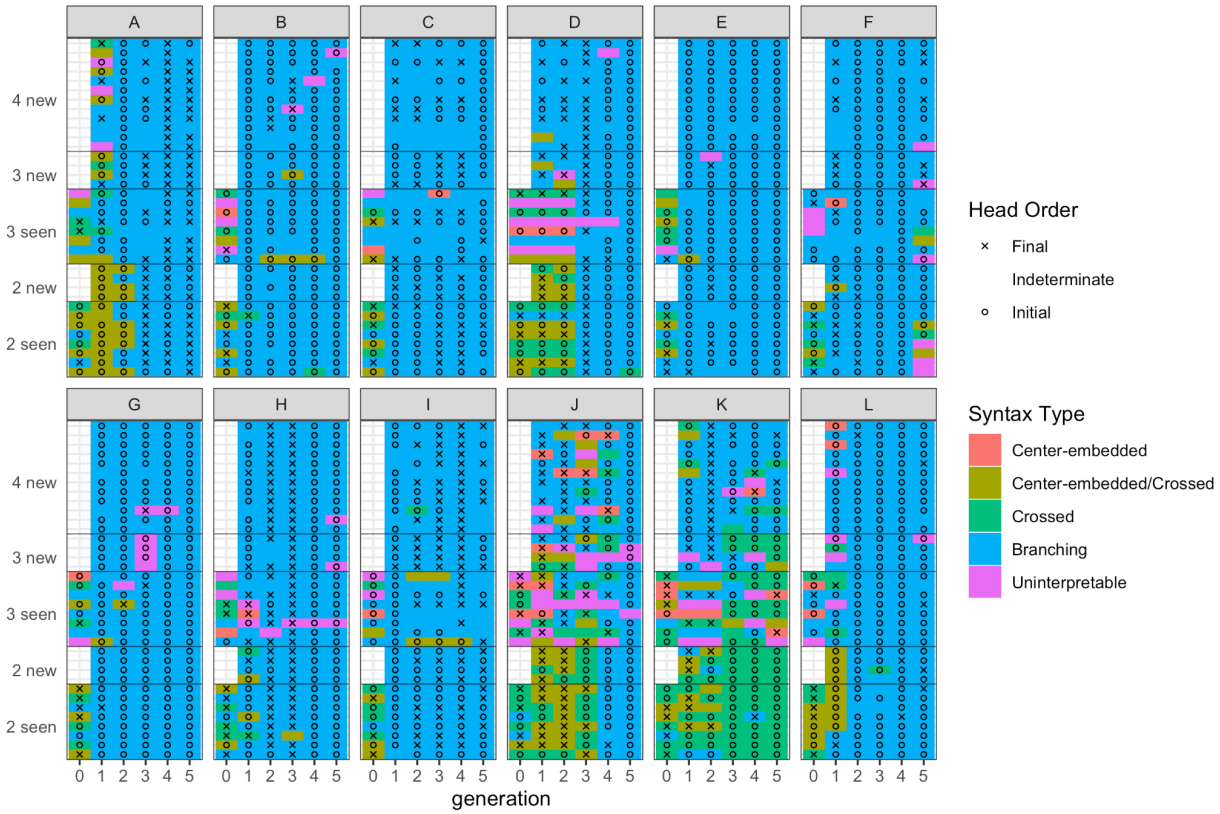

Figure 2: Best-match grammar by string, chain, and generation in Experiment 2. Each panel (A to L) is one chain; generations (participants) are columns and rows represent individual visual stimulus arrays of objects. The color of a given cell represents dependency type and O/X symbols represent head order, as listed in the sidebar. Note the prevalence of branching grammar (blue) in all chains but K, where crossed grammar strings (green) appear early and stabilize in later generations.