LEARNABILITY AND EMERGENCE OF DEPENDENCY STRUCTURES IN AN ARTIFICIAL LANGUAGE

Emily Davis[*1] and Kenny Smith[2]

[*]Corresponding Author: e4davis@ucsd.edu
[1]Department of Linguistics, UC San Diego, La Jolla, CA, USA
[2]Centre for Language Evolution, School of Philosophy, Psychology and Language Sciences, University of Edinburgh, UK

In a pair of artificial language experiments, we investigated the learnability and emergence of different dependency structures: branching, center-embedding, and crossed. In natural languages, branching is the most common dependency structure; center-embedding occurs but is often disfavored, and crossed dependencies are very rare. Experiment 1 addressed learnability, testing comprehension and production on small artificial languages exemplifying each dependency type in noun phrases. As expected, branching dependency grammars were the easiest to learn, but crossed grammars were no harder to learn than center-embedding. Experiment 2 employed iterated learning to examine the emergence and stabilization of consistent grammar using the same type of stimuli as Experiment 1. The initial participant in each chain of transmission was trained on phrases generated by a random grammar, with the language produced by that participant passed to the next participant through an iterated learning process. Branching dependency grammar appeared in most chains within a few generations and remained stable once it appeared, although one chain stabilized on output consistent with a crossed grammar; no chains converged on center-embedding grammars. These findings, along with some previous results, call into question the assumption that crossed dependencies are more cognitively complex than center-embedding, while confirming the role of learnability in the typology of dependency structures.

1. Introduction

In a pair of artificial language experiments, we investigated the learnability of different dependency structures: branching, center-embedding, and crossed. Long-distance dependencies between words (e.g. nouns and verbs, or nouns and adpositions) are an essential aspect of human language, and can be arranged in different ways. In (1), the dependency between subject and verb is indicated by the subscript (examples based on Vosse and Kempen 1991):

1.
   a. (Right-) branching dependencies in English:
   … when John$_1$ saw$_1$ Peter$_2$ walk$_2$

b. Center-embedded dependencies in German:
… als $Johan_1$ $Peter_2$ $laufen_2$ sah_1

c. Crossed dependencies in Dutch:
… toen Jan_1 $Peter_2$ sag_1 $lopen_2$

Most natural languages predominantly use some combination of branching and center-embedding dependency structures; crossed dependencies are rare, but some examples are attested, e.g. in Dutch subordinate clauses (Bresnan et al. 1982). 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 shaping natural language structures (e.g. Culbertson 2012, Fedzechkina 2018), and the regularization and stabilization of irregular input (e.g. Hudson Kam & Newport 2005). In the present study we applied these methods to dependencies. In two experiments, we investigated the learnability of different dependency types within the same paradigm, using meaningful sequences (descriptions for visual scenes) that constituted a small artificial language with several nouns and spatial adpositions.

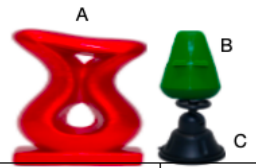
2. Experiment 1: Learnability

Experiment 1 concerned the learnability of the three types of dependency structures, as well as head-dependent order (head-initial or head-final), using noun phrases with locative adpositional phrases for configurations of novel objects.

2.1. Methodology

We created simple artificial languages representing the six possible combinations of dependency type and order (Figure 1). The associated visual stimuli were composed of images from the NOUN database as components (Horst 2015). Both experiments were conducted online through Mechanical Turk. Participants (n = 120, 20 per condition, all self-reported English speakers) were trained on one of these languages by a mix of passive training trials (viewing individual objects or configurations of objects with an appropriate description; Figure 2a) and comprehension trials (being presented with a description and attempting to identify the corresponding scene from an array of possibilities, with corrective feedback provided; Figure 2b-d).

Over the course of the training procedure participants progressed from scenes involving single objects to three-object configurations. In the final stage of the experiment, participants were asked to produce descriptions of scenes by clicking on labeled buttons (Figure 2e); these final test items included new configurations not seen in training (including some scenes featuring four objects), in order to assess generalization.



	Head Initial	Head Final
Branching	*kilkul moy vanva rae tovo* A left.of B atop C	*tovo rae vanva moy kilkul* C atop B left.of A
Center-Embedded	*kilkul vanva tovo rae moy* A B C atop left.of	*moy rae tovo vanva kilkul* left.of atop C B A
Crossed	*kilkul vanva tovo moy rae* A B C left.of atop	*rae moy tovo vanva kilkul* atop left.of C B A

Figure 1. Sample array with description in each of the six syntactic orders in Experiment 1. Objects are labeled with letters for clarity here and in other example arrays; no such labels were used on stimuli for the actual experiment.

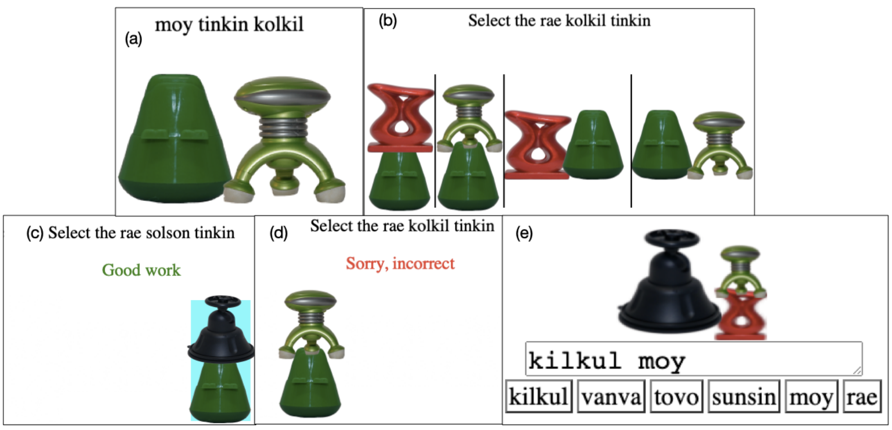


Figure 2: Passive training (a), comprehension (b), feedback on correct (c) and incorrect responses to comprehension (d) trials, and production trial with partially entered description (e) from Experiment 1.

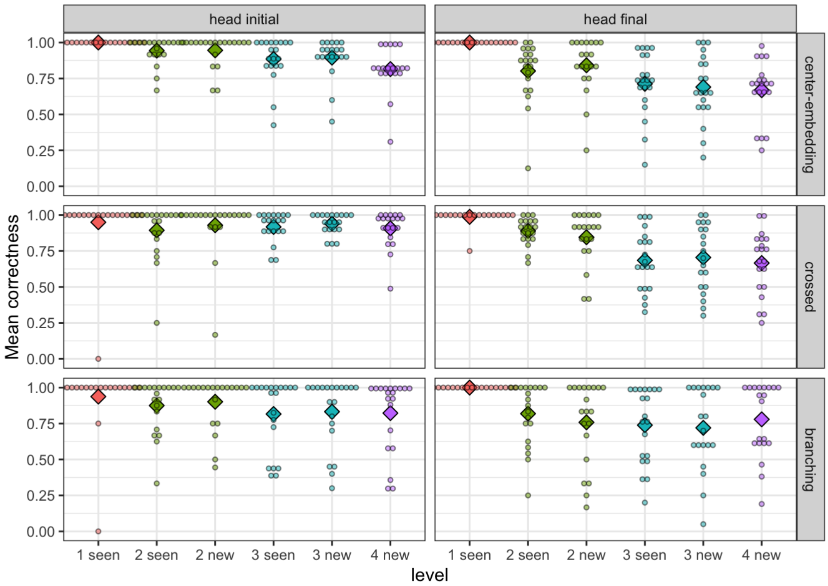


Figure 3: Learning accuracy (1 - string edit distance between trained and produced descriptions) by level (number of objects in the scene) and grammar type. Each point is one participant's performance at that level; large diamonds represent mean performance across all participants. "Seen" indicates items seen in training; "new" are novel test items. Note that participants were not tested on single-item scenes, and were not trained on 4-item scenes.

2.2. Analysis and results

Participant-created labels were analyzed for consistency with the training data (evaluated by calculating string edit distance between the trained and produced label for each scene) and for adherence to one of many possible grammars (by generating grammars with all possible combinations of dependency order and headedness, and finding the grammar which generated the description which best matched the participant's description, as assessed by edit distance). Figure 3 shows accuracy by condition. As expected, branching grammars were the easiest to learn: 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 ($B = -0.02$, $SE = 0.001$, $p = .04$), and participants trained on branching grammars generally produced output consistent with a branching grammar. Learning accuracy for crossed grammars was not significantly different from center-embedding grammars ($B = -0.01$, $SE = 0.02$, $p = .51$). However, 6 participants of the 40 trained on center-embedded grammar produced labels consistent with a crossed grammar, suggesting that they had reanalyzed the input in conformity with a new grammar. The converse (participants trained on crossed grammar generating center-embedded strings) did not occur. There was also a significantly greater decline in accuracy with level in head-final than in head-initial conditions ($B = -0.03$, $SE = 0.01$, $p < .001$).

3. Experiment 2: Emergence of grammar

Experiment 2 set out to examine the emergence and stabilization of a consistent grammar in noun phrases, and furthermore to observe whether the results reflected real language typology. We hypothesized that branching syntax would be predominant in the results, due to its greater learnability and comprehensibility, whereas center-embedding would be uncommon and crossed grammar might appear rarely, in keeping with previous experimental results (e.g. Öttl et al. 2015) and with the learnability results of Experiment 1. To this end we employed the method of *iterated learning* (Kirby et al. 2008), using stimuli of the same type as Experiment 1. Iterated learning is a process where the response of one participant (in our case, the descriptions produced for scenes) becomes the training input for the next participant in a chain of transmission. The first participants are trained on experimenter-designed items (termed *generation 0*); each participant's response to the input that he or she is exposed to is referred to as a *generation*, and the "genealogy" of all participant responses in the same line of descent constitutes a *chain*.

3.1. Methodology

This study was conducted online with English-speaking participants. Generation 0 participants were trained on an artificial language with no consistent structure; each description contained the nouns and adpositions of the artificial language in a randomized order (for example, an array with object B atop object A could bear the description (in English) "A B atop," "atop B A," "B atop A" and so on). Participants were trained on these descriptions using the same procedure as in Experiment 1, and as in Experiment 1, in the final production stage, participants were asked to label further scenes, including both trained and novel items. The labels for trained items created in this stage were passed on as input to the next participant in the chain, using training and comprehension trials in the same format as in Experiment 1. Consequently, at each generation, participants were trained on description for the same scenes, and then asked to label novel scenes; these novel labels made it possible to observe how the grammar was being generalized and whether it was stabilizing. We ran twelve chains, each of five generations (n = 60).

3.2. Analysis and results

Each string generated by a participant was compared to corresponding strings (i.e. referring to the same configuration of objects) generated by all 6 possible grammars (branching, crossed, center-embedding; head-initial or head-final), and the best match determined in the same manner as with the Experiment 1 results. We then calculated the entropy of the resulting grammar distribution (the distribution of best fit grammars for all strings produced by a given participant).

Grammar entropy of 0 bits would indicate a participant whose productions best matched those of a single grammar; entropy of 1 bit would indicate that a participant's productions were split over 2 grammars.

Grammars clearly stabilized over generations: a significant decrease in entropy occurred from 0 to 5 as measured by linear regression ($B = -.207$, $SE = .033$, $p < .001$), consistent with an increase in uniformity as irregularities were eliminated and a consistent grammar came to predominate. Figure 4 shows the best fit grammar for every production for every participant. While head-initial branching grammars (i.e. English-like grammars) dominate, one chain converges on a head-final branching grammar (chain A) and one chain (chain K) appears to converge on a crossed-dependency grammar. Center-embedding patterns did not emerge.

4. General discussion and conclusion

The results of these experiments are consistent with long-standing psycholinguistic findings (e.g. Bach et al. 1982) concerning the difficulty of center-embedded dependencies compared to branching structures. The status of crossed dependencies, however, remains somewhat anomalous: this form of dependency appears to be at least as easy to learn as center-embedding and possibly easier, but is very rare in natural languages. In our experiments, (strings consistent with) crossed dependency grammar emerged both as reanalyses of center-embedded input, and as a result of iterated learning (in one chain) from initially unstructured input.

In sum, these results provide additional evidence for the somewhat unexpected observation 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. The emergence of a crossed-dependency grammar in one iterated learning chain is also interesting, especially in light of the fact that center-embedding grammar was not observed to emerge in any chain. This strikingly conflicts with the observed patterns in natural languages. These results are also significant in light of the fact that this experiment featured meaningful semantic strings, demonstrating that the anomaly of crossed and center-embedded dependencies can arise in a meaningful language-like system as well as in non-semantic symbol sequences (e.g. DeVries et al 2008, Uddén et al 2012). Why natural language typology, where grammars allowing center-embedding are common, seems to be inconsistent with learnability remains to be explained.

Acknowledgments

This research received funding from the European Research Council under the European Union's Horizon 2020 research and innovation program (Grant 681942, held by K. Smith).

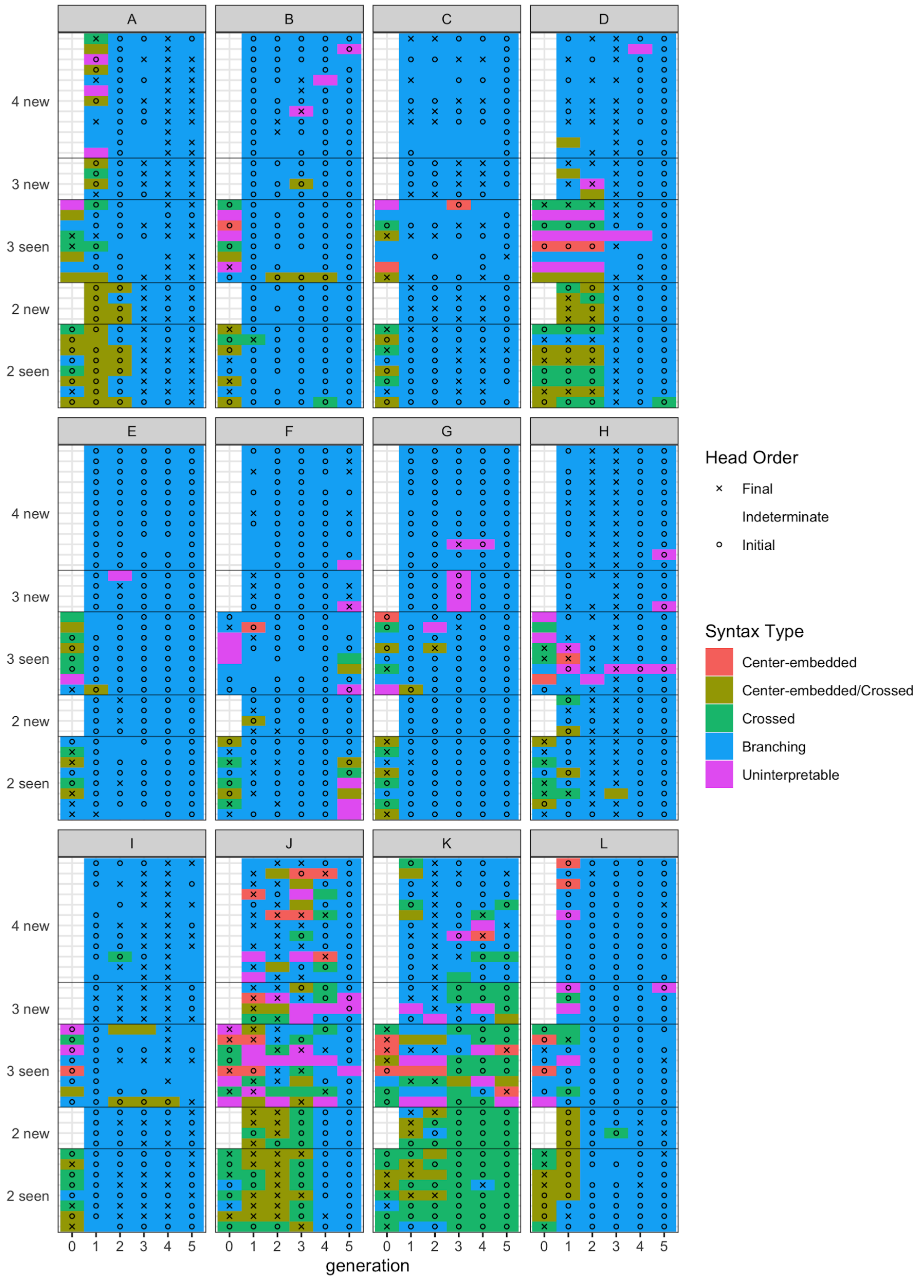


Figure 4: grammar by string, chain, and generation, Experiment 2. Each facet (A to L) is one chain; generations (participants) are columns and rows are individual scenes; the color of a given cell represents dependency type and O/X symbols represent head order, as listed in the sidebar.

References

Bach, E., Brown, C. & Marslen-Wilson, W. (1986). Crossed and nested dependencies in German and Dutch: a psycholinguistic study. *Language and Cognitive Processes, 1*, 249-262.

Bresnan, J., Kaplan, R. M., Peters, S., & Zaenen, A. (1982). Cross-serial dependencies in Dutch. In W. Savitch et al. (Eds.) *The formal complexity of natural language* (pp. 286-319). Springer, Dordrecht.

Culbertson, J., Smolensky, P., & Legendre, G. (2012). Learning biases predict a word order universal. *Cognition, 122*(3), 306-329.

de Vries, M., Monaghan, P., Knecht, S., & Zwitserlood, P. (2008). Syntactic structure and artificial grammar learning: The learnability of embedded hierarchical structures. *Cognition, 106*, 763-774

Fedzechkina, M., Chu, B., & Florian Jaeger, T. (2018). Human information processing shapes language change. *Psychological Science, 29*(1), 72-82.

Hawkins, J. (2004). *Efficiency and complexity in grammars.* Oxford University Press.

Horst, J. S., & Hout, M. C. (2015). The Novel Object and Unusual Name (NOUN) Database: A collection of novel images for use in experimental research. *Behavior Research Methods, 48*, 1393-1409.

Hudson Kam, C. L., & Newport, E. L. (2005). Regularizing unpredictable variation: The roles of adult and child learners in language formation and change. *Language Learning and Development, 1*, 151–195.

Kirby, S., Cornish, H., & Smith, K. (2008). Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. *Proceedings of the National Academy of Sciences, 105*(31), 10681-10686.

Kirby, S., Griffiths, T., & Smith, K. (2014). Iterated learning and the evolution of language. *Current Opinion in Neurobiology, 28,* 108-114.

Kuno, S. (1974). The position of relative clauses and conjunctions. *Linguistic Inquiry, 5*(1), 117-136.

Öttl, B., Jäger, G., & Kaup, B. (2015). Does formal complexity reflect cognitive complexity? Investigating aspects of the Chomsky hierarchy in an artificial language learning study. *PloS one, 10*(4), e0123059.

Uddén J, Ingvar M, Hagoort P, Petersson KM. (2012). Implicit acquisition of grammars with crossed and nested non-adjacent dependencies:Investigating the push-down stack model. *Cognitive Science, 36* ,1078–101.

Vosse, T. & Kempen, G. (1991). A hybrid model of human sentence processing: parsing right branching, center-embedded and cross-serial dependencies. In M. Tomita and M. Kempen (Eds.) *Proceedings of the Second International Workshop on Parsing Technologies*.