Infinite use of finite means? Evaluating the generalization of center embedding learned from an artificial grammar

R. Thomas McCoy,[1] Jennifer Culbertson,[2] Paul Smolensky,[3,1] and Géraldine Legendre[1]
tom.mccoy@jhu.edu, jennifer.culbertson@ed.ac.uk, psmo@microsoft.com, legendre@jhu.edu
[1]Department of Cognitive Science, Johns Hopkins University
[2]Centre for Language Evolution, University of Edinburgh
[3]Microsoft Research AI, Redmond, WA USA

Abstract

Human language is often assumed to make "infinite use of finite means"—that is, to generate an infinite number of possible utterances from a finite number of building blocks. From an acquisition perspective, this assumed property of language is interesting because learners must acquire their languages from a finite number of examples. To acquire an infinite language, learners must therefore generalize beyond the finite bounds of the linguistic data they have observed. In this work, we use an artificial language learning experiment to investigate whether people generalize in this way. We train participants on sequences from a simple grammar featuring center embedding, where the training sequences have at most two levels of embedding, and then evaluate whether participants accept sequences of a greater depth of embedding. We find that, when participants learn the pattern for sequences of the sizes they have observed, they also extrapolate it to sequences with a greater depth of embedding. These results support the hypothesis that the learning biases of humans favor languages with an infinite generative capacity.

Keywords: language acquisition; extrapolation; inductive biases; center embedding; artificial language learning

Introduction

During language acquisition, a learner's set of input sentences must have some maximum length, yet the languages acquired are often taken to be unbounded; language is often claimed to make "infinite use of finite means" (Chomsky, 1965, quoting von Humboldt, 1836). However, this view is not uncontroversial. It has been contested on logical grounds (Pullum & Scholz, 2010; Tiede & Stout, 2010), based on corpus data (Karlsson, 2010), and for particular languages (Everett, 2005). Further, even if we assume that learners do acquire an unbounded language, there are multiple possible explanations for *why* they might do so. One possibility is that language-external factors encourage unboundedness. For instance, using a form of semantic bootstrapping (Pinker, 1984, pg. 87), learners might generalize from *the child's mother* to the larger phrase *the child's mother's mother* based on the world knowledge that mothers have mothers of their own. Other aspects of experience that might promote unboundedness include nursery rhymes which gradually build recursive structures (e.g., "This is the House that Jack Built"; de Villiers & de Villiers, 2014) and sentences that are contextually implied to be infinitely long: *The meeting ran on and on and on and...* (Ziff, 1974). An alternative explanation is that unboundedness arises from some preference on the part of the learner—an *inductive bias*—that favors unbounded languages over bounded ones.[1] This explanation predicts that, for example, even without semantic grounding, people will generalize syntactic patterns beyond the finite bounds of their input.

To test this prediction, we use an artificial language learning paradigm, in which we train and test participants on a miniature language that has no semantics. We train participants on (bounded) center-embedded pairs of words, such as *A1 A2 A3 B3 B2 B1*, where there are two categories of words (category *A* and category *B*) and each word has a symmetrically opposite word that it depends on (e.g., *A2* and *B2* depend on each other: which *B*-word *B2* can be depends on which *A*-word *A2* is). How learners acquire such a grammar has been the focus of much past work with human learners (e.g., Perruchet & Rey, 2005; Hochmann, Azadpour, & Mehler, 2008; Poletiek et al., 2018) and connectionist models (e.g., Christiansen & Chater, 1999; Kirov & Frank, 2012; Lakretz, Dehaene, & King, 2020), as center embedding is often (albeit controversially) claimed to be a key type of structure in human languages and perhaps even only learnable by humans (Hauser, Chomsky, & Fitch, 2002).

Critically, it is unclear from past work whether people who learn center-embedded patterns also generalize them to greater sequence lengths than were seen during training. In naturally-occurring text and speech, even though deep embedding is fairly common for tail recursion, having more than one level of center embedding is extremely rare (Karlsson, 2010).[2] Moreover, deep center embedding poses substantial processing difficulties (Gibson & Thomas, 1999) which have led some to conclude that human language does not permit unbounded center embedding (Reich, 1969; Christiansen, 1992). Others counter by invoking the competence/performance distinction to argue that center embedding is not bounded in speakers' competence but only appears bounded due to memory constraints (Miller & Chomsky, 1963). In artificial language learning, Gentner, Fenn,

[1]If people have such an inductive bias, an additional question is what the nature of this bias is. For example, Perfors, Tenenbaum, Gibson, and Regier (2010) show that an inductive bias for simplicity can sometimes favor unbounded languages. See the Discussion.

[2]The presence of deep tail recursion in natural corpora is why we used center embedding in our experiment even though tail recursion is a simpler source of unboundedness. If we had used tail recursion, participants might have accepted deep embedding purely due to transfer from prior linguistic experience, rather than extrapolation from the experimental training set.

Margoliash, and Nusbaum (2006) found evidence that songbirds extrapolate center embedding to novel lengths, but did not test humans. Fitch and Hauser (2004, supplement) tested such extrapolation in humans, but later work that controlled for several confounds concluded that participants had learned a non-linguistic heuristic rather than the intended grammatical pattern (Perruchet & Rey, 2005; Hochmann et al., 2008). Poletiek (2002) also investigated human extrapolation, but in one experiment did not get clear evidence of learning even for the sequence lengths participants had seen, and in another only found generalization to novel lengths when the instructions indicated that sequences could be longer than the ones shown during training. Similarly, in Cho, Szkudlarek, and Tabor (2016), participants were given feedback after each test item, and such feedback also gave a direct signal that long sequences were acceptable. See the online supplement for a thorough review of prior work.[3]

To test whether people generalize center embedding to novel lengths, we use an extrapolation paradigm (Wilson, 2006; Culbertson & Adger, 2014): We train participants on a dataset that is ambiguous between two grammars of interest, and then test them on examples that disambiguate these possibilities, thus revealing learners' biases. In our case, the two grammars of interest are one that is bounded at the greatest depth of center embedding seen during training, and another that is not bounded at this level. We evaluate whether participants interpolate and extrapolate the pattern they are taught. By *interpolate*, we mean that they will have learned the intended pattern for (seen and unseen) sequences of lengths less than or equal to the maximum length they have seen. By *extrapolate* we mean that they will extend this pattern to allow sequences of a length greater than they have seen.

If participants have learned the bounded grammar, they should interpolate but not extrapolate; if they have learned the unbounded grammar, they should both interpolate and extrapolate. Importantly, some participants might fail to interpolate, making their behavior not consistent with either grammar. As is typical in work using the extrapolation paradigm, such participants are considered irrelevant: if they have not learned the relevant pattern in the training data, they cannot extrapolate it. For our core analyses, therefore, we ask whether participants who successfully interpolate also extrapolate.

To anticipate our results: We find that, when participants successfully interpolate the grammatical pattern we teach them, they also robustly extrapolate that pattern to a greater sequence length. This result supports the hypothesis that people have a learning bias which favors unbounded grammatical patterns over bounded ones.

Methods

Except where noted, all methods and analyses were preregistered.[4] Due to space constraints, not all preregistered analyses appear in the paper, but they are available in the online supplement. A demo of the experiment is also online.[5]

Participants

103 adult participants were recruited on Amazon Mechanical Turk.[6] We restricted the participant pool to those with a 95% approval rate and over 5000 approved Human Intelligence Tasks (HITs), under the Mechanical Turk blog's recommendations for improving the quality of participants.[7] Informed consent was obtained prior to the experiment. Participants took approximately 18 minutes and were paid $4.00 USD.

Materials

Our materials were based on the grammar in Figure 2. Generating sentences from this grammar involves center embedding, the process of embedding one structure in the center of another structure of the same type (in our case, S). The sequences generated by the grammar have the form A^nB^n, meaning n words from category A followed by n words from category B. There are nested dependencies between the A and B elements: which B word can appear in a given position is dictated by which A word appears in the symmetrically opposite position. Such a sequence might have the form $A_1A_2B_2B_1$, where A_1 and B_1 depend on each other, and A_2 and B_2 depend on each other.

All words in the grammar are single syllables, following most artificial language learning work on center embedding (e.g., Fitch & Hauser, 2004). The words in category A have the vowel *i*, while those in category B have the vowel *o*. Each A word has exactly one B word that can appear in the symmetrically opposite position; specifically, this B word is the one that has the same syllable structure as the A word. For example, *gri* is always matched with *klo* because both have consonant-consonant-vowel syllable structure. An example sequence generated by the grammar is *gri djirn vi fo cholm klo*, whose derivation is in Figure 3. That example has two **levels of embedding** because it contains the sequence *vi fo* embedded inside the sequence *djirn cholm*, in turn embedded inside the sequence *gri klo*.

In our extrapolation design, the training set contained 114 grammatical sequences, using 0, 1, or 2 levels of embedding (see Figure 1 for a breakdown of the training set). Further levels of embedding were withheld, so the training set is ambiguous as to whether embedding deeper than 2 levels is permitted. The test set contained 24 grammatical sequences and 24 ungrammatical sequences with 0, 1, 2, or 3 levels of embedding. Critically, the test examples with 3 levels of embedding will indicate whether participants extrapolated to sequences longer than those seen during training. All training and test examples obeyed the constraint that no word could appear twice within a given sequence, to prevent participants from looking for spurious patterns in such repetitions. No

[3] https://github.com/tommccoy1/center_embedding_extrapolation

[4] https://osf.io/dft6r

[5] http://rtmccoy.com/center_embedding.html

[6] https://www.mturk.com/

[7] https://blog.mturk.com/improving-quality-with-qualifications-tips-for-api-requesters-87eff638f1d1

Levels of embedding	Count in training set	Count in test set	Grammatical example	Ungrammatical example
0	54	6	djirn cholm	djirn **klo**
1	36	10	zin vi fo som	zin vi fo **plom**
2	24	16	i djirn vi fo cholm o	i djirn vi fo **o cholm**
3	0	16	zin id brin gri klo plom ot som	zin id brin gri klo **ot plom** som

Figure 1: Composition of the training and test sets. In the training set, all examples are grammatical. In the test set, half of the examples for each depth of embedding are grammatical, and the other half are ungrammatical. The bolding of ungrammatical examples was not present in the experiment. The counts in the training set use the length distribution given by a simple probabilistic version of our grammar in which each sequence size has 1.5 times the probability of the size one greater than it, but with the progression truncated after two levels of embedding.

S → i S o	S → gri S klo
S → vi S fo	S → brin S plom
S → id S ot	S → djirn S cholm
S → zin S som	S → ε

Figure 2: The grammar. ε indicates the empty string.

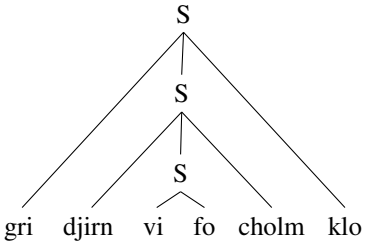

Figure 3: A tree generated by the grammar in Figure 2 (omitting the final null S), yielding the sequence *gri djirn vi fo cholm klo*. This sequence has two levels of embedding: an S embedded inside an S embedded inside another S.

sequence was used more than once across the training and test set, except for the sequences with 0 levels of embedding, since there were too few of those to avoid repetition.

Grammatical examples were generated randomly from the grammar. We used two methods to generate ungrammatical sequences. For ungrammatical sequences with 2 or 3 levels of embedding, we used the **swap method**: Generate a grammatical sequence, then select two words from the second half of the sequence and swap them to break the sequence's nested dependency structure. Neither of the selected words could be part of the innermost pair of words. The swap method ensured that the ungrammatical sequences preserved the following properties:

(1) The number of *A* words is equal to the number of *B* words.

(2) Every pair of consecutive words can grammatically appear in sequences generated by the grammar.

(3) Each word's partner from the other *A* or *B* class is also present (albeit potentially in the wrong place).

Preserving these properties ensures that participants must have acquired the grammar's nested dependency structure in order to differentiate grammatical and ungrammatical sequences. They could not succeed by simply counting *A* and *B* words (ruled out by property 1), observing only local transitions between words (ruled out by property 2), or treating the sequences as unordered sets (ruled out by property 3).

To generate ungrammatical sequences with 0 or 1 levels of embedding, we used the **point mutation** method: change the last word in the sequence to a different *B* word, breaking the dependency between the sequence's first word and last word. These examples lacked property (3), and the ones with 0 levels of embedding further lacked property (2); it is impossible to generate ungrammatical sequences with 0 or 1 levels of embedding that have all 3 properties. Therefore, we excluded these test examples from our primary analyses (although for completeness we report results on all test examples).

Both the training set and the test set were generated randomly for each participant.

Procedure

Training phase: Participants were told that they would see sequences that were sentences in an alien language. The 114 sequences in the training set were presented in random order. For each sequence, a fixation cross was presented for 1 second, and then the sequence was presented one word at a time. As each word appeared, the participant had to press a button corresponding to that word (Figure 4, left). These buttons were arranged in a way that was intended to help highlight the dependencies between words. If a participant pressed the wrong button, an error message appeared and the sequence started over from the beginning.

Testing phase: Participants were told they must judge whether new sequences are possible sentences in this language. The 48 test sequences were then presented in random order. Each entire sequence was presented at once to mitigate the memory limitations that arise with processing center embedding (e.g., Gibson & Thomas, 1999), and participants had to click a button indicating whether the sequence was a valid sentence in the alien language (Figure 4, right).

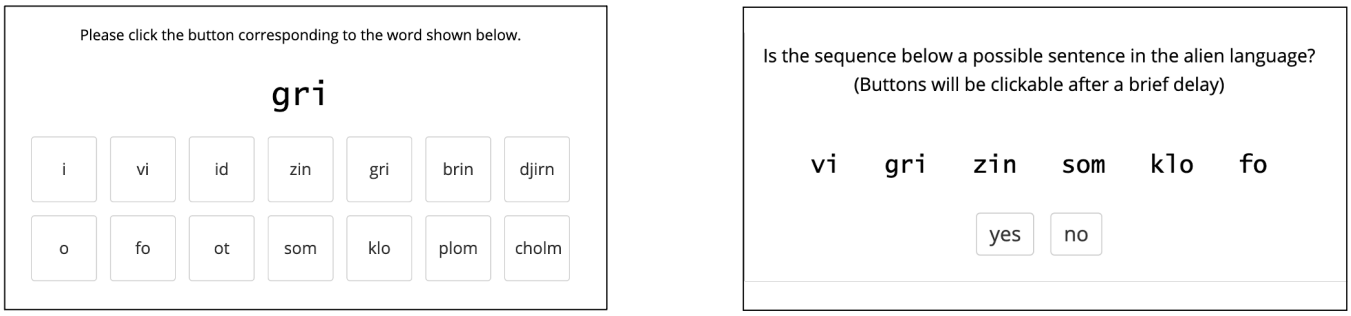

Figure 4: Experimental interface. Left: example training screen; right: example testing screen.

We asked for absolute judgments rather than relative judgments (e.g., selecting which of two sentences is better) because only absolute judgments can establish if participants had extrapolated the pattern: even if participants did not extrapolate, they might still find grammatical extrapolation examples to be less bad than ungrammatical extrapolation examples. Thus, with relative judgments, participants could show similar behavior whether they had extrapolated or not, whereas absolute judgments would differentiate these two types of participants.

To discourage participants from rapidly clicking through the test without looking at the sequences, there was a brief delay before the response buttons could be clicked. In addition, we paid a bonus ($1.00) to participants scoring $\geq 75\%$ on items with 0, 1, or 2 levels of embedding.

Results

We divide the test set into three parts: examples with 0 or 1 levels of embedding; examples with 2 levels of embedding (the *interpolation subset*); and examples with 3 levels of embedding (the *extrapolation subset*). The preregistered statistical analyses below (`https://osf.io/dft6r`) support the following hypotheses, qualitatively suggested by Figures 5 and 6: on all three test subsets, average performance is above chance (Figure 5, top); further, interpolation accuracy and extrapolation accuracy are strongly positively correlated (Figure 5, bottom; Figure 6).

All participants: Comparisons to chance

We first test whether participants indeed scored significantly above chance on the three test subsets. For each of these subsets, we ran an intercept-only mixed-effects logistic regression with by-item and by-participant random intercepts. The binary response variable was a 1 if the participant responded correctly or 0 otherwise. These analyses showed that participants scored significantly above chance on the 0 or 1 levels of embedding trials (mean = 0.61; $p < 0.001$), the interpolation subset (mean = 0.61; $p < 0.001$), and the extrapolation subset (mean = 0.59; $p < 0.001$).

Interpolation success implies extrapolation success

To analyze the relationship between interpolation accuracy and extrapolation accuracy, we performed three analyses.

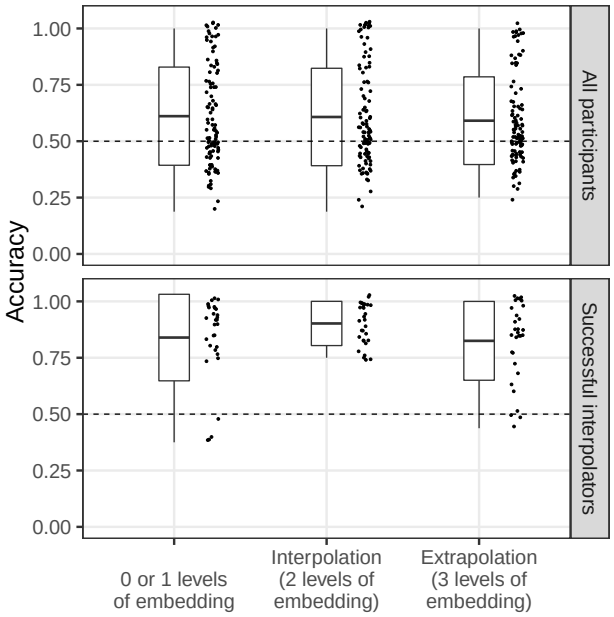

Figure 5: Accuracy summary. Top plot includes all participants, bottom plot contains only participants who scored 75% or above on the interpolation subset. Dots are individual participants (with x and y jitter). Boxplots show the mean, one standard deviation above or below the mean, and the range.

First, we ran a correlation test which revealed a strong positive correlation between interpolation accuracy and extrapolation accuracy (Pearson's correlation coefficient: 0.82; $p < 0.001$). This shows that higher accuracy at identifying grammatical sequences with 2 levels of embedding (the greatest depth seen during training) is associated with higher accuracy at identifying grammatical sequences with 3 levels of embedding (not seen during training).

Second, we investigated the performance of the subset of participants whose interpolation accuracy was higher than chance. We did this because our hypothesis is about how participants will generalize the pattern *that they have learned* to a novel length. This hypothesis is thus best evaluated by looking at participants who have actually learned the pattern for the lengths they have observed. Our preregistered crite-

rion for successful interpolation was 75% or above on the interpolation test subset: this is the minimum score x such that achieving a score of x or above has a probability less than 0.05 under a binomial model with $p(success) = 0.5$ (i.e., the probability of success that participants would have by chance if guessing). 30 participants met this criterion. To see whether these successful interpolators also extrapolated the language to 3 levels of embedding, we ran an intercept-only mixed-effects logistic regression with by-participant and by-item random intercepts. This regression had a singular fit, so (following our preregistration) we backed off by removing the by-participant random intercept. The resulting model showed that these participants scored significantly above chance on the extrapolation subset (mean = 0.83, $p < 0.001$).

It is especially noteworthy that extrapolation accuracy was high on the grammatical extrapolation trials (mean = 0.87). This provides particularly strong evidence that participants have extrapolated: The accuracy on these trials would be 0.00 if participants had learned a grammar bounded at two levels of embedding, or 0.50 if participants had guessed randomly on extrapolation. Less importantly, extrapolation accuracy was also high on the ungrammatical trials (mean = 0.78); i.e., participants correctly rejected ungrammatical sequences, as predicted under the bounded or unbounded grammar.

As a final way to evaluate whether successful interpolation implied extrapolation, we conducted a non-preregistered (post-hoc) analysis of the performance of individual participants (presented in the online supplement). This analysis reveals no clear examples of individuals who acquired the grammar in a bounded way, while providing strong evidence that some individuals have extrapolated the grammar.[8]

Discussion

In this experiment, we tested the hypothesis that learners are biased in favor of inferring unbounded structures in language. We predicted that participants learning a grammar with nested dependencies would extrapolate to a level of embedding not present in their training data. As in previous artificial language experiments on the learning of center embedding, this task was difficult for participants. On average, however, our participants displayed successful learning, albeit with a small effect size (average accuracy of 61%, where chance is 50%). Crucially, individuals who successfully learned this pattern also robustly extended the pattern to larger sequences, with an average accuracy of 83% on the extrapolation cases. This result is consistent with the hypothesis that people have a learning bias which favors extrapolation of grammatical patterns.

Why did participants do so well? Even ignoring extrapolation, merely finding above-chance interpolation of center embedding is noteworthy. In prior work, several apparent cases of success have later been cast into doubt because the relevant test sets had not observed the 3 properties identified as crucial above: property (1) (Hochmann et al., 2008; De Vries, Monaghan, Knecht, & Zwitserlood, 2008), property (2) (Perruchet & Rey, 2005; De Vries et al., 2008), or property (3) (De Vries et al., 2008).[9] The prior experiment most similar to our setup is the "random" condition in Experiment 2 from Poletiek et al. (2018), in which the average accuracy was 0.51 (which was not significantly above chance).

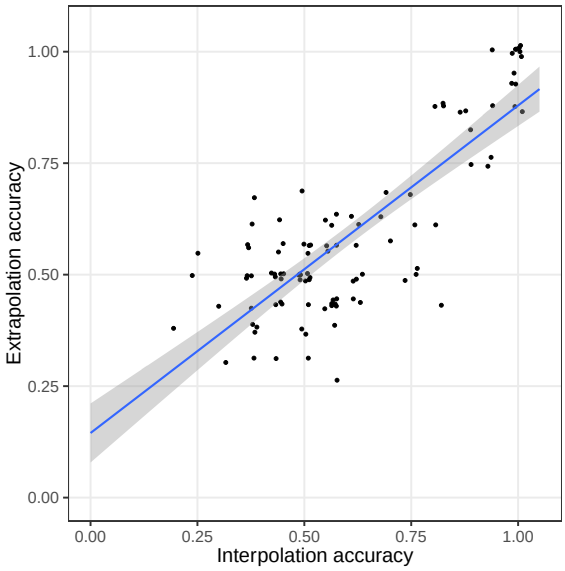

Figure 6: Extrapolation vs. interpolation accuracy. Dots are individual participants (with x and y jitter). The blue line is a regression line with a 95% confidence interval in gray.

Two manipulations that have improved learning of center embedding are the use of a long training phase spread over multiple days (Uddén et al., 2009) and the use of a so-called *starting small* set-up, in which training items are ordered from smallest to largest (Conway, Ellefson, & Christiansen, 2003; Poletiek et al., 2018). However, we found successful learning while controlling for the properties listed above, and without either of these manipulations.

We suspect that two novel components of our design contributed to successful learning. First, we believe that our use of syllable structure (both phonological and orthographic) as a cue to dependencies makes these dependencies more salient than they are in past work: most past work either used no

[8] Other participants learned neither the bounded grammar nor the unbounded grammar; most of these participants appear to have been guessing randomly, though there was also a sizable proportion who labeled all test items as grammatical.

[9] Three previous papers have found successful learning while also ruling out the heuristics that the three properties avoid. These papers achieved this by requiring participants to *generate* full or partial sequences, rather than *discriminating* between grammatical and ungrammatical sequences. Specifically, Rey, Perruchet, and Fagot (2012) and Ferrigno, Cheyette, Piantadosi, and Cantlon (2020) required participants to rearrange a provided set of units to form a full sequence, and Jiang et al. (2018) required participants to complete a partial sequence. We did not use these paradigms because they only show relative preferences between potential sequences, whereas our question required absolute judgments of acceptability. That is, when a participant generates a sequence, it is unclear if the participant believes that the sequence is grammatical, or if it is the least bad option from a set of options that are all ungrammatical.

phonological cues to the dependencies (e.g., Conway et al., 2003) or used only the place of articulation of a word's onset (e.g., Poletiek et al., 2018), which we believe is likely less salient than our property of syllable structure (which was also in most cases accompanied by place-of-articulation cues).

Second, the button arrangement used during training may have provided helpful spatial or motor cues. We note, however, that participants could not succeed at the test if all they learned was a certain motor pattern applicable to the buttons, because the buttons were not present during the test phase.

Unbounded generalization? Our results show that participants generalized the grammar one level of embedding deeper than they had witnessed. Does this mean that they have learned an unbounded grammar, or simply a grammar that is bounded at a level higher than the one they have observed?

One way to think about the difference between a grammar with bounded center embedding and a grammar with unbounded center embedding is that the former would likely need to include one component for every level of embedding. For instance, the language $\{A^nB^n,\ 0 \leq n \leq 3\}$ (without A-B dependencies) could be expressed with the context-free grammar in (4), which has one *rule* per sequence size, or with the context-sensitive grammar in (5), which has one *context* per sequence size ('#' marks edges):

(4) $S \rightarrow \varepsilon$; $S \rightarrow AB$; $S \rightarrow AABB$; $S \rightarrow AAABBB$

(5) $S \rightarrow ASB \;/\; [\ \#\_\# \mid \#A\_B\# \mid \#AA\_BB\#\ ]$; $S \rightarrow \varepsilon$

(6) $S \rightarrow ASB$; $S \rightarrow \varepsilon$

If participants have in fact acquired a bounded grammar along the lines of (4) or (5), then in order to generalize to unseen levels of embedding, they would have needed to posit a specific part of the grammar for that specific level of embedding without ever having seen a sequence that used that part of the grammar. While that is in principle possible, it seems less likely than that they have acquired a grammar with a recursive rule that generates any level of embedding, as in (6).

Nature of the inductive bias: Our results show that people have an inductive bias that leads them to extrapolate a center-embedded pattern that they have learned. What is the nature of this bias? We are aware of two possibilities. The more obvious possibility is a bias which favors unbounded over bounded nesting. The other possibility is a bias for simplicity (Perfors et al., 2010): in many cases, including ours, an unbounded grammar—e.g., (6)—provides a simpler explanation of the training data than a bounded grammar does—e.g., (4)—under a Bayesian definition of simplicity that factors in the size of the grammar (the prior) and the probability that the grammar assigns to the training corpus (the likelihood). A learner could therefore prefer the unbounded grammar solely because of a general bias for simplicity, rather than a bias for unboundedness. The current study cannot differentiate these biases, but it verifies a crucial behavioral prediction made by both of them, namely that people will generalize center embedding beyond the bounds they have observed, even without real-world grounding that could encourage unboundedness. This fact is not clear from existing natural language acquisition data, so establishing it is an important first step in investigating these biases. Now that we have verified the behavior that must be explained, follow-ups are in progress to tease apart the possible explanations for that behavior.

Ecological validity: By design our artificial language is much simpler than natural language, and participants learn it in a way that is in some sense unnatural. However, the main strength of artificial language learning paradigms is that they enable us to carefully control the input to learning in a way that is impossible when studying natural language acquisition. In particular, here we can ensure that there is no direct evidence for depths of embedding greater than 2. That said, there may be interesting ways in which enriching the input might affect our results. For example, future work could test whether learning behavior changes when the stimuli are semantically meaningful.

There remains the concern that laboratory language learning experiments might not tap into the learning mechanisms relevant for natural language acquisition. For example, previous research on center embedding has in some cases shown that participants use heuristics (Perruchet & Rey, 2005). While we have designed our stimuli to make those heuristics unhelpful, it is still worth noting that here, as elsewhere, converging evidence is needed to convincingly determine what biases learners bring to language acquisition. In past work, ALL has corroborated or enhanced insights from natural language acquisition (Wonnacott, Newport, & Tanenhaus, 2008), language typology (Culbertson, Smolensky, & Legendre, 2012), and computational modeling (Schuler, Yang, & Newport, 2016), so we conclude that ALL can—and does—play an important role in piecing together our understanding of learning biases. See Culbertson and Schuler (2019) and Morgan and Newport (1981) for further discussion of what ALL can tell us about language acquisition.

Conclusion

In this study, we used an artificial language learning paradigm to show that, when participants learned a center-embedded pattern from sequences containing at most 2 levels of embedding, they extrapolated it to a greater depth of embedding. Interestingly, we found successful learning of the intended grammar with a simple design (i.e., without manipulations like starting small or using a multi-day training period that were necessary in previous studies) and while controlling for common confounds present in previous work. Our results are consistent with the hypothesis that people have a bias for generalizing syntactic patterns to greater sizes than they have observed. Such a bias would support long-standing claims that human languages "make infinite use of finite means."

Acknowledgments

For helpful comments, we are grateful to Grusha Prasad, Najoung Kim, Tal Linzen, Robert Frank, the JHU Neurosymbolic Computation Lab, and the NYU CAP Lab. This research was supported by NSF GRFP No. 1746891.

References

Cho, P. W., Szkudlarek, E., & Tabor, W. (2016). Discovery of a recursive principle: An artificial grammar investigation of human learning of a counting recursion language. *Frontiers in Psychology*.

Chomsky, N. (1965). *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press.

Christiansen, M. (1992). The (non) necessity of recursion in natural language processing. In *Proc. CogSci.*

Christiansen, M., & Chater, N. (1999). Toward a connectionist model of recursion in human linguistic performance. *CogSci*.

Conway, C., Ellefson, M., & Christiansen, M. (2003). When less is less and when less is more: Starting small with staged input. In *Proc. CogSci.*

Culbertson, J., & Adger, D. (2014). Language learners privilege structured meaning over surface frequency. *PNAS*.

Culbertson, J., & Schuler, K. (2019). Artificial language learning in children. *Annual Review of Linguistics*.

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

de Villiers, J. G., & de Villiers, P. A. (2014). The role of language in theory of mind development. *TLD*.

De Vries, M., Monaghan, P., Knecht, S., & Zwitserlood, P. (2008). Syntactic structure & artificial grammar learning: The learnability of embedded hierarchical structures. *Cognition*.

Everett, D. (2005). Cultural constraints on grammar and cognition in Pirahã: Another look at the design features of human language. *Current anthropology*.

Ferrigno, S., Cheyette, S., Piantadosi, S., & Cantlon, J. (2020). Recursive sequence generation in monkeys, children, US adults, and native Amazonians. *Sci. Advances*.

Fitch, T., & Hauser, M. (2004). Computational constraints on syntactic processing in a nonhuman primate. *Science*.

Gentner, T., Fenn, K., Margoliash, D., & Nusbaum, H. (2006). Recursive syntactic pattern learning by songbirds. *Nature*.

Gibson, E., & Thomas, J. (1999). Memory limitations & structural forgetting: The perception of complex ungrammatical sentences as grammatical. *Lang. & Cog. Proc*.

Hauser, M., Chomsky, N., & Fitch, T. (2002). The faculty of language: what is it, who has it, and how did it evolve? *Science*.

Hochmann, J.-R., Azadpour, M., & Mehler, J. (2008). Do humans really learn A^nB^n artificial grammars from exemplars? *CogSci*.

Jiang, X., Long, T., Cao, W., Li, J., Dehaene, S., & Wang, L. (2018). Production of supra-regular spatial sequences by macaque monkeys. *Current Biology*.

Karlsson, F. (2010). Syntactic recursion and iteration. *Recursion and Human Language*.

Kirov, C., & Frank, R. (2012). Processing of nested and cross-serial dependencies: an automaton perspective on SRN behaviour. *Connection Science*.

Lakretz, Y., Dehaene, S., & King, J.-R. (2020). What limits our capacity to process nested long-range dependencies in sentence comprehension? *Entropy*.

Miller, G., & Chomsky, N. (1963). Finitary models of language users. In *Handbook of Mathematical Psychology.*

Morgan, J. L., & Newport, E. L. (1981). The role of constituent structure in the induction of an artificial language. *Journal of Verbal Learning and Verbal Behavior*.

Perfors, A., Tenenbaum, J., Gibson, E., & Regier, T. (2010). How recursive is language? A Bayesian exploration. *Recursion and Human Language*.

Perruchet, P., & Rey, A. (2005). Does the mastery of center-embedded linguistic structures distinguish humans from nonhuman primates? *Psychonomic Bulletin & Review*.

Pinker, S. (1984). *Language Learnability and Language Development*. Harvard University Press.

Poletiek, F. (2002). Implicit learning of a recursive rule in an artificial grammar. *Acta Psychologica*.

Poletiek, F., Conway, C., Ellefson, M., Lai, J., Bocanegra, B., & Christiansen, M. (2018). Under what conditions can recursion be learned? Effects of starting small in artificial grammar learning of center-embedded structure. *CogSci*.

Pullum, G. K., & Scholz, B. C. (2010). Recursion and the infinitude claim. *Recursion in Human Language*.

Reich, P. (1969). The finiteness of natural language. *Lang*.

Rey, A., Perruchet, P., & Fagot, J. (2012). Centre-embedded structures are a by-product of associative learning and working memory constraints: Evidence from baboons (papio papio). *Cognition*.

Schuler, K., Yang, C., & Newport, E. (2016). Testing the Tolerance Principle: Children form productive rules when it is more computationally efficient to do so. In *Proc. CogSci.*

Tiede, H.-J., & Stout, L. (2010). Recursion, infinity and modeling. *Recursion and Human Language*.

Uddén, J., Araujo, S., Forkstam, C., Ingvar, M., Hagoort, P., & Petersson, K. (2009). A matter of time: Implicit acquisition of recursive sequence structures. In *Proc. CogSci.*

von Humboldt, W. (1836). *Über die Verschiedenheit des Menschlichen Sprachbaues*.

Wilson, C. (2006). Learning phonology with substantive bias: An experimental and computational study of velar palatalization. *CogSci*.

Wonnacott, E., Newport, E. L., & Tanenhaus, M. K. (2008). Acquiring and processing verb argument structure: Distributional learning in a miniature language. *Cog. Psych*.

Ziff, P. (1974). The number of English sentences. *Foundations of Language*.