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 Appendix ("Thorough review of related word") for a more extensive discussion of prior work.

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.[3] Due to space constraints, not all preregistered analyses appear in the paper, but they are available in the Appendix. A demo of the experiment as it appeared to participants is also available online.[4]

Participants

103 adult participants were recruited on Amazon Mechanical Turk.[5] 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.[6] 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://osf.io/dft6r

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

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

[6] 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.

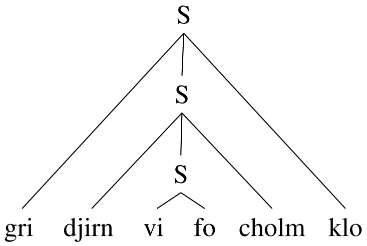


Figure 3: A tree generated by the grammar in Figure 2 (omitting the final null S), yielding the sequence *gri djirn vi fo cholm klo*. This sequence has two levels of embedding: an S embedded inside an S embedded inside another S.

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).

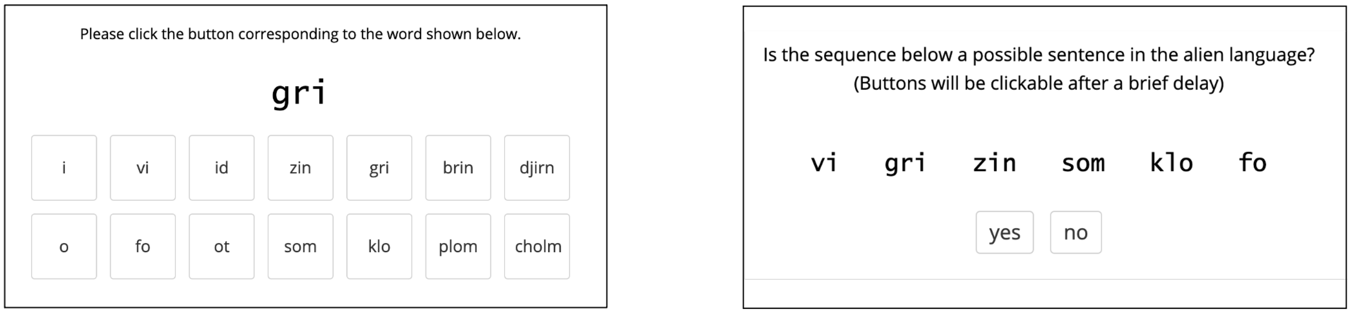
To generate ungrammatical sequences with 0 or 1 levels of embedding, we used the **point mutation** method: change the last word in the sequence to a different *B* word, breaking the dependency between the sequence's first word and last word. These examples lacked property (3), and the ones with 0 levels 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).



Please click the button corresponding to the word shown below.

gri

Is the sequence below a possible sentence in the alien language?
(Buttons will be clickable after a brief delay)

vi gri zin som klo fo

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.

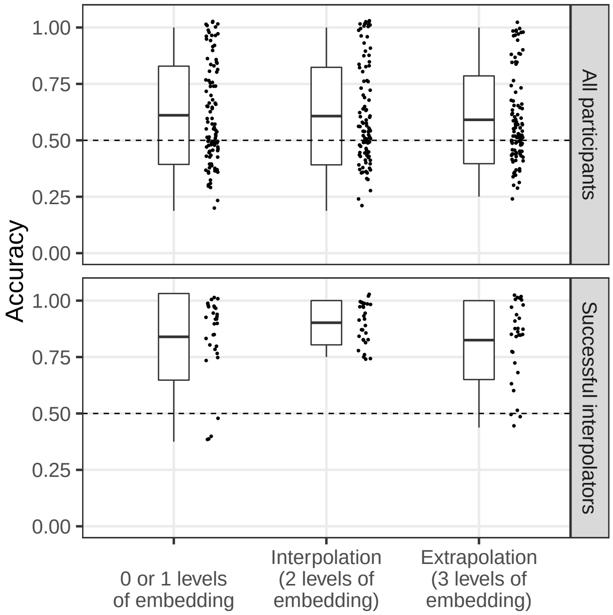


Figure 5: Accuracy summary. Top plot includes all participants, bottom plot contains only participants who scored 75% or above on the interpolation subset. Dots are individual participants (with x and y jitter). Boxplots show the mean, one standard deviation above or below the mean, and the range.

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 criterion 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 Appendix). 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.[7]

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).[8] 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).

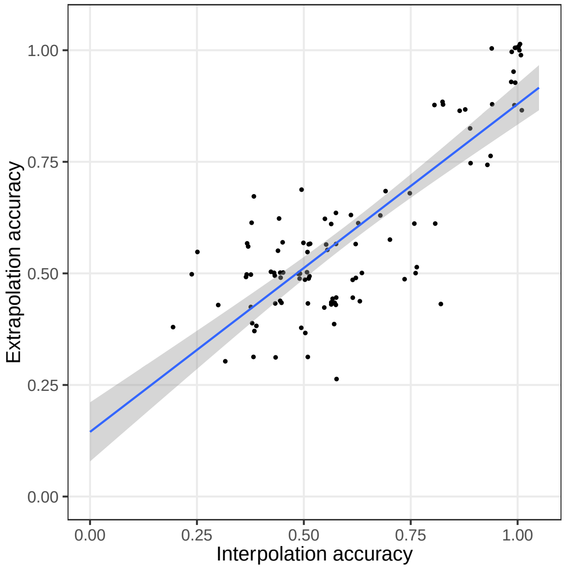


Figure 6: Extrapolation vs. interpolation accuracy. Dots are individual participants (with x and y jitter). The blue line is a regression line with a 95% confidence interval in gray.

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

[7] 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.

[8] 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 $\{\mathrm{A}^n\mathrm{B}^n,\ 0 \le n \le 3\}$ (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) $\mathrm{S} \rightarrow \varepsilon;\quad \mathrm{S} \rightarrow \mathrm{AB};\quad \mathrm{S} \rightarrow \mathrm{AABB};\quad \mathrm{S} \rightarrow \mathrm{AAABBB}$

(5) $\mathrm{S} \rightarrow \mathrm{ASB} \;\Big/\; [\ \#\_\# \mid \#\mathrm{A}\_\mathrm{B}\# \mid \#\mathrm{AA}\_\mathrm{BB}\#\ ];\quad \mathrm{S} \rightarrow \varepsilon$

(6) $\mathrm{S} \rightarrow \mathrm{ASB};\quad \mathrm{S} \rightarrow \varepsilon$

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

Abe, K., & Watanabe, D. (2011). Songbirds possess the spontaneous ability to discriminate syntactic rules. *Nature Neuroscience*.

Alamia, A., Gauducheau, V., Paisios, D., & VanRullen, R. (2019). Which neural network architecture matches human behavior in artificial grammar learning? *arXiv preprint arXiv:1902.04861*.

Alexandre, J. (2010). Modeling implicit and explicit processes in recursive sequence structure learning. In *Proc. CogSci.*

Bahlmann, J., Gunter, T., & Friederici, A. (2006). Hierarchical and linear sequence processing: An electrophysiological exploration of two different grammar types. *Journal of Cognitive Neuroscience*.

Bahlmann, J., Schubotz, R., & Friederici, A. (2008). Hierarchical artificial grammar processing engages Broca's area. *Neuroimage*.

Bahlmann, J., Schubotz, R., Mueller, J., Koester, D., & Friederici, A. (2009). Neural circuits of hierarchical visuospatial sequence processing. *Brain Research*.

Cho, P. W., Szkudlarek, E., Kukona, A., & Tabor, W. (2011). An artificial grammar investigation into the mental encoding of syntactic structure. In *Proc. CogSci.*

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.

Chouinard, M., & Clark, E. (2003). Adult reformulations of child errors as negative evidence. *Journal of Child Language*.

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., & de Villiers, P. (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*.

De Vries, M., Petersson, K., Geukes, S., Zwitserlood, P., & Christiansen, M. (2012). Processing multiple non-adjacent dependencies: Evidence from sequence learning. *Philosophical Transactions of the Royal Society B: Biological Sciences*.

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

Fedor, A., Varga, M., & Szathmáry, E. (2012). Semantics boosts syntax in artificial grammar learning tasks with recursion. *Journal of Experimental Psychology: Learning, Memory, and Cognition*.

Fenn, K., Brawn, T., Gentner, T., Margoliash, D., & Nusbaum, H. (2007). Complex acoustic pattern learning in songbirds and humans. In *Proc. CogSci.*

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*.

Friederici, A., Bahlmann, J., Heim, S., Schubotz, R., & Anwander, A. (2006). The brain differentiates human and non-human grammars: functional localization and structural connectivity. *PNAS*.

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, W. 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^{n}B^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*.

Lai, J., de Jong, C., Gao, D., Huang, R., Krahmer, E., & Sprenger, J. (2016). The influence of language-specific auditory cues on the learnability of center-embedded recursion. In *Cogsci.*

Lai, J., & Poletiek, F. (2010). The impact of starting small on the learnability of recursion. In *Proc. CogSci.*

Lai, J., & Poletiek, F. (2011). The impact of adjacent-dependencies and staged-input on the learnability of center-embedded hierarchical structures. *Cognition*.

Lai, J., & Poletiek, F. (2013). How “small” is “starting small” for learning hierarchical centre-embedded structures? *Journal of Cognitive Psychology*.

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

Malassis, R., Dehaene, S., & Fagot, J. (2020). Baboons (papio papio) process a context-free but not a context-sensitive grammar. *Scientific reports*.

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

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

Mueller, J., Bahlmann, J., & Friederici, A. (2010). Learnability of embedded syntactic structures depends on prosodic cues. *CogSci*.

Ojima, S., & Okanoya, K. (2020). Children’s learning of a semantics-free artificial grammar with center embedding. *Biolinguistics*.

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*.

Perfors, A., Ransom, K., & Navarro, D. (2014). People ignore token frequency when deciding how widely to generalize. In *Proc. CogSci.*

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*.

Ravignani, A., Westphal-Fitch, G., Aust, U., Schlumpp, M., & Fitch, W. T. (2015). More than one way to see it: Individual heuristics in avian visual computation. *Cognition*.

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*.

Rohrmeier, M., & Cross, I. (2009). Tacit tonality-implicit learning of context-free harmonic structure. In *ESCOM.*

Rohrmeier, M., Fu, Q., & Dienes, Z. (2012). Implicit learning of recursive context-free grammars. *PloS One*.

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.*

Shin, W.-J., & Eberhard, K. M. (2015). Learning a center-embeddding rule in an artificial grammar learning task. In *Cogsci.*

Stobbe, N., Westphal-Fitch, G., Aust, U., & Fitch, W. T. (2012). Visual artificial grammar learning: comparative research on humans, kea (Nestor notabilis) and pigeons (Columba livia). *Philosophical Transactions of the Royal Society B: Biological Sciences*.

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.*

Udden, J., Ingvar, M., Hagoort, P., & Petersson, K. M. (2012). Implicit acquisition of grammars with crossed and nested non-adjacent dependencies: Investigating the push-down stack model. *CogSci*.

Van Heijningen, C., De Visser, J., Zuidema, W., & Ten Cate, C. (2009). Simple rules can explain discrimination of putative recursive syntactic structures by a songbird species. *PNAS*.

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

Westphal-Fitch, G., Giustolisi, B., Cecchetto, C., Martin, J., & Fitch, W. T. (2018). Artificial grammar learning capabilities in an abstract visual task match requirements for linguistic syntax. *Frontiers in psychology*.

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

Winkler, M., Mueller, J., Friederici, A., & Männel, C. (2018). Infant cognition includes the potentially human-unique ability to encode embedding. *Science Advances*.

Wonnacott, E., Newport, E., & Tanenhaus, M. (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*.

Zimmerer, V., Cowell, P., & Varley, R. (2011). Individual behavior in learning of an artificial grammar. *Memory & Cognition*.

Zimmerer, V., Cowell, P., & Varley, R. (2014). Artificial grammar learning in individuals with severe aphasia. *Neuropsychologia*.

Zimmerer, V., & Varley, R. (2015). A case of “order insensitivity”? natural and artificial language processing in a man with primary progressive aphasia. *Cortex*.

Zuidema, W. (2013). Context-freeness revisited. In *Proc. CogSci.*

Results of all preregistered analyses

Our study was preregistered at `https://osf.io/dft6r`. Due to space constraints, not all of our preregistered analyses were included in the main paper. Here we present the results of all preregistered analyses, both those included in the paper and those left out of the paper.

Each analysis was run twice: once for all participants, and once for the subset of participants who were deemed to have successfully interpolated. These successful interpolators were identified as those who scored 75% or higher on the interpolation test items. The accuracy level of 75% was chosen as 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 if guessing by chance).

All participants: Comparisons to chance

For each of four subsets of our test set, we ran an intercept-only mixed-effects logistic regression with by-item and by-participant random intercepts, to test whether participants perform significantly above chance on that subset. These four subsets are defined by sequence length: sequences of length 2 or 4 (i.e., the non-critical test items that do not preserve properties (1) through (3)); sequences of length 4; sequences of length 6 (i.e., the interpolation test items); and sequences of length 8 (i.e., the extrapolation test items). The formula for these regressions is:

$$\text{correct} \sim (1|\text{participant}) + (1|\text{item}) + 1 \qquad (1)$$

The dependent variable, *correct*, is a binary variable with a value of 1 if the participant got that item correct, or 0 if the participant got that item incorrect; we define an answer as being correct if it aligns with the predictions of our target grammar, whether those predictions are being made for interpolation or extrapolation. All analyses were run in R using the `glmer` function from the `lme4` package.

For all four of these subsets, participants performed significantly above chance. For the subset of lengths 2 and 4, $p < 0.001$; for the subset of length 4, $p < 0.01$; for the subset of length 6, $p < 0.001$; and for the subset of length 8, $p < 0.001$.

All participants: Correlation between interpolation and extrapolation

We ran a correlation test correlating participants' interpolation accuracy (i.e., accuracy on the test subset of items with length 6) and extrapolation accuracy (i.e., accuracy on the test subset of items with length 8). We find that there is a strong positive correlation between interpolation accuracy and extrapolation accuracy (Pearson's correlation coefficient: 0.82; $p < 0.001$).

All participants: Effect of length

To check whether there was a significant difference in performance based on whether a sequence was longer than the lengths seen during training, we ran a mixed-effects logistic regression on the interpolation and extrapolation subsets pooled together (i.e., the test items with length 6 or length 8), with a fixed effect for whether the sequence is interpolation or extrapolation, as well as by-participant and by-item random intercepts, and a by-participant random slope for whether the sequence is interpolation or extrapolation. This model had a singular fit, so (following our preregistration) we removed the by-participant random slope for whether the sequence is interpolation or extrapolation. This yielded the following formula for the regression:

$$\text{correct} \sim \text{interpolation} + (1|\text{participant}) + (1|\text{item}) + 1 \quad (2)$$

This test found no significant effect of the distinction between interpolation and extrapolation ($p = 0.29$).

Successful interpolators: Comparisons to chance

We repeated the comparisons to chance specified above but with only the subset of participants who were deemed successful interpolators. For both the length 6 subset and the length 8 subset, the model had a singular fit. Therefore, following our preregistration, we removed the by-item random intercept from the length 6 model, while we removed the by-participant random intercept from the length 8 model.

As before, participants performed significantly above chance for all four subsets. For the subset of lengths 2 and 4, $p < 0.001$; for the subset of length 4, $p < 0.001$; for the subset of length 6, $p < 0.001$; and for the subset of length 8, $p < 0.001$. We caution that the results for the length 6 subset are not very meaningful because performance on this subset was used to select the set of successful interpolators.

Successful interpolators: Correlation between interpolation and extrapolation

We repeated the correlation described above but with only the subset of participants who were deemed successful interpolators. As before, we found a strong positive correlation between interpolation accuracy and extrapolation accuracy (Pearson's correlation coefficient: 0.81; $p < 0.001$.

Successful interpolators: Effect of length

We repeated the test specified above to test whether there was a significant difference in performance based on whether a sequence was longer than those seen during training but with only the subset of participants who were deemed successful interpolators. As before, the model had a singular fit when using the maximal set of effects, so we removed the by-participant random slope for whether the sequence was interpolation or extrapolation. In contrast to the analogous test run for all participants, this test found that there was a significant effect for whether the sequence was in the interpolation or extrapolation subset ($p < 0.001$). However, we caution that this effect may not be meaningful: The successful interpolators were identified as those with high interpolation accuracies, without any restrictions placed on their extrapolation accuracies. Thus, the participants were constrained to have high interpolation accuracies but could potentially have lower extrapolation accuracies, making it plausible that this significant effect arises purely because of the selection bias created by the exclusion criterion.

Analysis of individual participants

As a final way to see whether successful interpolation implies extrapolation, we conducted a non-preregistered analysis of the performance of individual participants. Our goal was to identify individuals who had successfully interpolated and extrapolated, as well as individuals who had successfully interpolated but who rejected all extrapolation sequences. We operationalized these two types of generalization as follows: A participant was deemed to have successfully interpolated and extrapolated if they scored 75% or above on each part of the test set (the items with 0 or 1 levels of embedding, the interpolation subset, and the extrapolation subset). A participant was deemed to have successfully interpolated but to have rejected all extrapolation cases if they scored 75% or above on the fillers and interpolation subset but rejected at least 75% of the extrapolation items. The threshold was 75% because this is the level of success needed to have a probability of less than 5% for participants who were guessing randomly, under a binomial model.

Framed in this way, 23 participants interpolated and extrapolated, while only 1 participant interpolated but rejected the extrapolation examples. However, even the 1 participant who met our criterion to be categorized as rejecting the extrapolation examples may not have actually learned a bounded grammar: this participant still labeled 4 of the 16 extrapolation examples as grammatical, so they were not very reliable at rejecting the extrapolation examples. Even more illuminating is this participant's free response answer describing what they had learned (at the end of the testing phase, participants were given the option to provide a free-response answer to the prompt *If you would like, please describe how you decided which sequences were possible sentences in the alien language*, and this participant provided a response): "it took me a while to see the pattern, but it looked like certain pairs of words should appear in an equally distant location from the center. For example. D C B A A B C D. From there, it was also necessary to remember the pairs. I am uncertain if order of appearance matters. (plom brin == brin plom?)." This response includes the example *D C B A A B C D*, which has 3 levels of embedding (our extrapolation size), suggesting that they have not in fact concluded that the grammar is bounded at 2 levels of embedding. In sum, our analysis of individual participants reveals no clear examples of individuals who acquired the grammar in a bounded way, while providing strong evidence to conclude that some individuals have extrapolated the grammar.

For the complete results for each participant (accuracy on each type of example, as well as responses to the optional free-response question), see our online results table.[9]

Thorough review of related work

In Tables 1, 2, and 3, we summarize all past works that we are aware of which use artificial language learning to study center embedding. For each case, we report whether the experiment meets each of 9 requirements that are necessary for addressing our question of whether humans generalize center embedding to novel depths. Below we define each of the 9 requirements that are used in these tables. As the tables show, no past work meets all 9 requirements, whereas our current work does meet all of them.

Note that past works had different goals from ours, so the fact that they do not meet these requirements is not necessarily a criticism of those works: these are the requirements for answering our question, but not necessarily for answering the questions addressed by past works. For instance, many works have focused on what it takes for people to learn center-embedded patterns (e.g., Poletiek et al., 2018), rather than focusing on how people generalize such patterns when they are learned, and testing length generalization is not necessary for that question.

Requirement: With humans

Because our question is about how humans generalize, it is necessary to include human participants. Some past works have instead focused exclusively on non-human animals, meaning they do not meet this requirement (Table 1).

Requirement: Tests length generalization

Our question is about generalization to novel sequence lengths, making it necessary to test such generalization. Most past works with humans do not test such generalization, meaning that they do not lend insight into our question (Table 2).

Requirement: Involves absolute grammaticality judgments

Our question is crucially about whether participants believe that long sequences are grammatical. Ascertaining this requires absolute grammaticality judgments (i.e., presenting a single sequence and asking if it is grammatical). Some studies instead use a forced choice between two options (Mueller, Bahlmann, & Friederici, 2010; Ravignani, Westphal-Fitch, Aust, Schlumpp, & Fitch, 2015; Rohrmeier & Cross, 2009; Rohrmeier, Fu, & Dienes, 2012; Stobbe, Westphal-Fitch, Aust, & Fitch, 2012), the completion of a partial sequence (Cho, Szkudlarek, Kukona, & Tabor, 2011; Cho et al., 2016; De Vries, Petersson, Geukes, Zwitserlood, & Christiansen, 2012; Jiang et al., 2018; Malassis, Dehaene, & Fagot, 2020; Rey et al., 2012; Shin & Eberhard, 2015), the formation of a sequence by rearranging scrambled atomic units (Ferrigno et al., 2020), or the use of reaction times or other aspects

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

of processing as evidence for differences between stimulus types (Alexandre, 2010; Abe & Watanabe, 2011; Winkler, Mueller, Friederici, & Männel, 2018). Such approaches are valid for indicating participants' preferences, but they cannot reveal grammaticality judgments: it is possible that all possible choices are grammatical (or all are ungrammatical) and that the participant's choice is simply the least bad of a set of ungrammatical options (or the best of a set of grammatical options), whether that choice be a forced choice between two sequences or the choice of what sequence to generate. In addition, measures of processing cannot replace grammaticality judgment, because sometimes grammatical sentences can be hard to process (e.g., garden path sentences) while ungrammatical sentences can be easy to process (e.g., grammaticality illusions).

Requirement: No negative evidence

For two reasons, we consider it important not to provide any negative evidence (i.e., direct indications that certain sequences are ungrammatical) in order to address our question. First, negative evidence is generally believed not to play a major role in language acquisition (though see Chouinard & Clark, 2003), so we avoid negative evidence to minimize the differences between our setup and natural language acquisition. Second, if negative evidence is provided, there is a risk that participants will learn what sorts of errors they should watch for as indications of ungrammaticality, and then only reject items that have those particular errors. In our case, if negative evidence were provided, it could never be with the novel-length sequences (as the training must remain ambiguous as to whether those are grammatical), so it would instead have to be provided for some particular type of grammaticality violation, such as violations of the sequence's dependency structure. Participants who otherwise might have rejected long sequences might then instead learn that they should only be using dependency violations as a basis for rejection. Many prior studies provide negative evidence via explicit feedback given on participants' grammaticality judgments, making those studies unable to address our main question.

Requirements relating to center embedding

As discussed in the main paper, we enforce several requirements to ensure we can be confident that successful learning is specifically successful learning of center embedding, rather than learning of some other structure or heuristic. The next three requirements all relate to this goal; this goal is also listed in the main paper as property (1), property (2), and property (3).

Requirement: Cannot be solved by counting: In some works, violations are created by having a different number of A's and B's (e.g., Poletiek, 2002; Zimmerer, Cowell, & Varley, 2011). Such a design requires participants to recognize that there must be the same number of A's as B's, but it does not ensure that those A's and B's follow a center-embedded structure. They could instead follow a cross-serial pattern ($A_1A_2A_3B_1B_2B_3$) or have no dependency relationship. To ensure that center embedding is being tested for, it is necessary to ensure that such a counting strategy cannot suffice.

Requirement: Cannot be solved with bigrams: In many works (e.g., Fitch & Hauser, 2004; Gentner et al., 2006; Stobbe et al., 2012), the sequences that participants must recognize as ungrammatical all involve bigrams that cannot appear in any grammatical sequence, such as a B element preceding an A element. Thus, participants can succeed on the test solely by recognizing these ungrammatical local transitions, without having learned a center-embedded pattern. In certain cases, only some test items lead to such bigrams, such as when the symmetrical dependency relations are broken for the innermost A/B pair; in the tables below, such situations are indicated as ✓/✗. For the purposes of the table, sequence boundaries are considered part of a bigram; thus, sequences of all A's or all B's are considered to have bigram violations because of the transition from the start of the sequence to a B, or from an A to the end of the sequence.

Requirement: Cannot be solved without attending to word order When experiments include a center-embedded dependency pattern (as is necessary to test for learning of center embedding), one approach for generating ungrammatical sequences is to replace one element with another element that breaks the dependency structure (e.g., (Bahlmann, Schubotz, & Friederici, 2008; Conway et al., 2003; Poletiek et al., 2018)). Such violations can be detected without recognizing the center-embedded structure but by instead treating the words in the sequence as an unordered set, and checking whether, for each A that is present, its corresponding B is also present. The approach that we and others (e.g., De Vries et al., 2008; Winkler et al., 2018) use to rule out this strategy is to instead create ungrammatical examples by swapping 2 elements from the same half of the sequence, so that the "unordered set" approach would fail to recognize the ungrammatical sequences as ungrammatical.

Requirement: No starting small

Generally speaking, the field has found it challenging to induce learning of center embedding. One manipulation that has been found to substantially assist learning is starting small (Conway et al., 2003; Lai & Poletiek, 2011; Shin & Eberhard, 2015; Poletiek et al., 2018), in which all training sequences with 0 levels of embedding are presented first, followed by all training sequences with 1 level of embedding, followed by all training sequences with 2 levels of embedding. Though this strategy helps markedly with learning, we believe that it is likely to interact with our main question by biasing participants to generalize to novel lengths: Because the training is arranged in a way that successively introduces participants to longer sequences than they have seen before, participants would likely infer from this staged training that they are intended to be generalizing the grammar to longer sequences. Thus, even if participants are not in general biased to extrapolate to novel lengths, starting small might induce them to do so.

Requirement: Successful for seen lengths

Because our question is about how participants generalize what they have learned, the question cannot be meaningfully addressed if participants fail to learn the pattern at all. Thus, it is important to ensure that participants have achieved above-chance learning on the lengths they have seen.

Discussion of particular papers

Though no prior experiments satisfy all of the requirements listed above, a few come close because they test length generalization in humans and only violate a few of our requirements (Table 3). Here we discuss those works in more detail.

Cho et al. (2011, 2016)

Cho et al. (2011, 2016) used a grammar with 2 rules: S → 1 2 3 4 and S → 1 S 2 3 4. Participants were instructed to predict what the next token would be at each point in the sequence where the next token could be deterministically predicted; with this grammar, all of the sequence elements after the first '2' are deterministic. There was no distinct training and testing phase—all inputs were presented in this same way (with participants predicting the next tokens at various points in the sequence)—but the design did restrict the training order in such a way that the greatest depth of embedding did not appear until the last 25% of the experiment. Thus, this final 25% can be viewed as a test portion that involves generalization to a novel length. However, there are several reasons why this study cannot answer our question. First, the experiment only involves grammatical sequences; therefore, by presenting participants with a sequence of a novel length, the design indicates to them that these sequences are grammatical. Second, no explicit grammaticality judgments are collected—so we cannot be certain if participants do accept the novel-length sequences (though, as mentioned in the previous point, it is likely that they do accept them because of the context of the experiment). Finally, because there is only one A element (1) and one B element (2 3 4), we cannot tell if the sequences are being learned as a center-embedded structure or as some other structure that involves a matched number of A's and B's.

Despite these limitations, it is worth reviewing the results of Cho et al. (2011, 2016). These studies find that, when first trained with 0 or 1 level of embedding, participants perform poorly on fully completing sequences with 2 levels of embedding at first, but relatively quickly climb to stronger performance. However, when the training is modified so that the 1-level-of-embedding sequences do not follow the same pattern as the 0-level-of-embedding sequences (i.e., 1 1 4 3 2 4 2 3, instead of the recursively-generated sequence 1 1 2 3 4 2 3 4), participants no longer reach a strong performance on sequences with 2 levels of embedding. The initially low scores suggest participants may not have extrapolated automatically, instead requiring direct training to be taught that the new lengths were grammatical; but the differences between the two conditions suggest that the extrapolation involved application of a recursive rule (instead of merely memorizing the novel-length sequence), as participants only succeeded when a recursive rule was available. Thus, it seems that the participants' training did not induce extrapolation but did facilitate extrapolation via further training. Finally, in another experiment, the training was expanded to include 0-, 1-, or 2-level-of-embedding sequences (all generated with a recursive rule), with the length extrapolation now involving 3 levels of embedding. Interestingly, participants now had a high sequence completion accuracy on even the first instance of the novel length, in stark contrast to the previous experiment where several exposures were required to perform well on the novel length). The authors conclude that there is a shift when training includes 2 levels of embedding: such training now induces participants to extrapolate, whereas including only 1 or 2 levels does not. Nonetheless, as discussed above, we cannot tell if participants believed these novel-length sequences to be grammatical before having the experimental setup provide evidence that they are; but these results are still illuminating, and they suggest that future work within our paradigm could investigate the role of the levels of embedding present during training.

Other sequence-generation works

Besides the work of Cho et al. (2011, 2016), several other works used sequence-generation paradigms to test humans for extrapolation to novel lengths by having them complete a partial sequence (Jiang et al., 2018; Malassis et al., 2020). In both of these works, humans correctly generalized a mirror pattern, and they used a larger alphabet than Cho et al. (2011, 2016), meaning that success required the learning of symmetrical dependencies, rather than just counting. However, these works still do not address our question because the sequence-generation paradigm does not collect absolute judgments of grammaticality, but rather reveals preferences between the set of possible sequences, which may all be viewed as ungrammatical by participants.

Fitch and Hauser (2004)

The influential work of Fitch and Hauser (2004) studied the learning of an A^nB^n pattern (e.g., *AAABBB*) and an $(AB)^n$ pattern (e.g., *ABABAB*) in humans and cotton-top tamarins (a non-human primate). They tested their human participants for generalization to a novel length, and found that humans accepted the novel-length sequence (see their supplement). However, the ungrammatical test items that were used were from the $(AB)^n$ pattern, which can be distinguished from the grammatical A^nB^n pattern merely by the fact that they include a B followed by an A; these transitions are particularly prominent because the A and B syllables were differentiated by using two different speakers, one female and one male. Thus, it is likely that participants succeeded by using this auditory heuristic rather than by learning a center-embedded grammar; Perruchet and Rey (2005), Hochmann et al. (2008), and De Vries et al. (2008) ran replications controlling for this factor, and their results support the conclusion that participants had not learned a center-embedded grammar.

Stobbe et al. (2012)

Like Fitch and Hauser (2004), Stobbe et al. (2012) tested for length generalization but used A^nB^n items as the grammatical sequences and $(AB)^n$ items as ungrammatical sequences, meaning that the task could be solved purely by learning bigram transition probabilities rather than a center-embedded grammar. In addition, Stobbe et al. (2012) used a forced-choice paradigm, meaning that the fact that participants chose the correct sequence does not necessarily mean that they believed this sequence to be ungrammatical; they may have simply believed it to be less bad than the other choice.

Westphal-Fitch, Giustolisi, Cecchetto, Martin, and Fitch (2018)

For Westphal-Fitch et al. (2018), ungrammatical sequences were generated by removing one element from a grammatical sequence. Thus, the task could be solved purely by counting the number of A's and B's, without having learned any center-embedded structure, meaning that the length generalization that they observed may have been purely about counting rather than about extrapolating center embedding.

Zuidema (2013)

Zuidema (2013) used two types of ungrammatical items: items of the form $(AB)^n$, which as discussed above can be solved by only paying attention to bigram transitions; and items of the form A^nB^m, which can be solved by only counting the number of A's and B's (or even by counting the overall number of tokens and noticing that it is odd instead of even). Thus, while Zuidema (2013) tested for length generalization, this may not have required any learning of center-embedded structure.

Perfors, Ransom, and Navarro (2014)

In Perfors et al. (2014), the training set contained sequences of the form $A^nB^mA^n$, where there was only one possibility for A (the syllable *du*), and three possibilities for B. During training, n and m were both at most 4; testing included examples with $n = 5$ and $n = 6$. Participants generally accepted these extra-long test examples. However, the design of the test examples does not ensure that participants learned a center-embedded grammar: the same behavior that was observed would have arisen if participants had learned the regular grammar $A^*B^*A^*$—that is, any number of A syllables followed by any number of B syllables followed by any number of A syllables. In addition, because there was only one option for the A syllable, participants did not necessarily learn a nested dependency structure (as opposed to, e.g., a cross-serial dependency structure, or no dependency structure).

Poletiek (2002)

In both experiments of Poletiek (2002), participants did not extrapolate to a novel length when tested for this; but it is also unclear if they even learned the pattern for the lengths they had seen, as performance was not significantly above chance for items with 1 level of embedding. In the second part of Experiment 2, participants did perform above chance on length extrapolation cases, but only after a new instruction was added which informed participants that some extrapolation cases are grammatical, which means that this extrapolation is most likely due to the experimental design rather than the learners' biases. In addition, for both experiments, ungrammatical items were generating by removing one token from a grammatical sequence, meaning that the task could be solved by counting A and B elements without learning a center-embedded structure.

Shin and Eberhard (2015)

Shin and Eberhard (2015) included 5 conditions and tested for length extrapolation in each. 4 of their 5 conditions used starting small, which—as discussed above—is likely to interact with our hypothesis in ways that make it difficult for a starting-small experiment to illuminate the learners' biases. In the one condition that did not use starting small, no evidence of learning was found.

One starting-small condition found successful learning and extrapolation to novel lengths, but this condition included feedback that gave negative evidence, which (as discussed above) is another factor that prevents these results from speaking to our main question. When this feedback was removed in another condition, no evidence for learning was found.

Another starting-small condition used a sequence-completion paradigm, instead of giving explicit indications of ungrammaticality; this condition led to successful learning and length extrapolation, but the sequence-completion paradigm does not tell us whether participants viewed the length-extrapolated sequences as grammatical. That condition also included feedback on the completions; in another condition where that feedback was removed, participants' performance got substantially worse, though a few still learned successfully.

Our work

None of the prior works discussed above satisfactorily address our question (though we stress again that, in most cases, this is because they were investigating different questions, not because they attempted to address our question and failed). In this work, we created a design that satisfies all 9 criteria so that this question can be investigated without the confounds that prevent prior work from addressing it:

- We evaluate humans.

- We test length generalization.

- We obtain absolute grammaticality judgments of the test items.
- No feedback is given on any grammaticality judgments, so that no negative evidence is provided.
- The test set is carefully designed so that the crucial test items (the interpolation and extrapolation subsets) cannot be solved via bigram probabilities, counting, or ignoring word order (instead requiring attention to the symmetric dependencies).
- We do not use starting small.
- Our participants perform above chance on the sequence lengths they were trained on.

	With humans	Tests length generalization	Involves absolute grammaticality judgments	No negative evidence	Cannot be solved with bigrams	Cannot be solved by counting	Cannot be solved without attending to word order	No starting small	Successful for seen lengths
Abe and Watanabe (2011)	✗	✓	✗	✓	✓	✓	✓	✓	✓
Gentner et al. (2006): *AAAA* and *BBBB* items	✗	✗	✓	✗	✗	✗	✗	✓	✓
Gentner et al. (2006): $(AB)^n$ items	✗	✓	✓	✗	✗	✓	✓	✓	✓
Gentner et al. (2006): *ABBA*, and *BAAB* items	✗	✗	✓	✗	✗	✓	✓	✓	✓
Gentner et al. (2006): A^*B^* items	✗	✓	✓	✗	✓	✗	✓	✓	✓
Ravignani et al. (2015): *Extension* items	✗	✓	✗	✗	✗	✓	✓	✓	✓
Ravignani et al. (2015): *Reversal* and *Permutation* items	✗	✗	✗	✗	✗	✓	✓	✓	✓
Ravignani et al. (2015): *Non-matching* items	✗	✗	✗	✗	✓	✗	✗	✓	✓
Ravignani et al. (2015): *Pure* items	✗	✗	✗	✗	✗	✗	✗	✓	✓
Rey et al. (2012)	✗	✓	✗	✗	✓	✓	✓	✓	✓
Van Heijningen, De Visser, Zuidema, and Ten Cate (2009)	✗	✗	✓	✗	✗	✓	✓	✓	✓

Table 1: Prior works involving A^nB^n artificial language learning, but not with humans. ✓ indicates that the experiment meets the requirement; ✗ indicates that it does not.

START OF TABLE 2; CONTINUED ON NEXT PAGE

	With humans	Tests length generalization	Involves absolute grammaticality judgments	No negative evidence	Cannot be solved with bigrams	Cannot be solved by counting	Cannot be solved without attending to word order	No starting small	Successful for seen lengths
Alamia, Gauducheau, Paisios, and VanRullen (2019)	✓	✗	✓	✗	✓/✗	✓	✓	✓	✓
Alexandre (2010)	✓	✗	✗	✓	✓	✓	✓	✓	✓
Bahlmann, Gunter, and Friederici (2006)	✓	✗	✓	✗	✓/✗	✗	✗	✓	✓
Bahlmann et al. (2008): *Replacement* violations	✓	✗	✓	✗	✓/✗	✗	✗	✗	✓
Bahlmann et al. (2008): *Concatenation* violations	✓	✗	✓	✗	✓/✗	✓	✗	✗	✓
Bahlmann, Schubotz, Mueller, Koester, and Friederici (2009): *Category* violations	✓	✗	✓	✗	✓/✗	✗	✗	✗	✓
Bahlmann et al. (2009): *Index* violations	✓	✗	✓	✗	✓/✗	✓	✗	✗	✓
Conway et al. (2003): *Starting small* conditions	✓	✗	✓	✓	✓/✗	✓	✗	✗	✓/✗
Conway et al. (2003): *Random* conditions	✓	✗	✓	✓	✓/✗	✓	✗	✓	✗
De Vries et al. (2008): *Hier-Viol* conditions	✓	✗	✓	✗	✓/✗	✗	✗	✓	✓
De Vries et al. (2008): *Scram-Viol* conditions	✓	✗	✓	✗	✓/✗	✗	✗	✓	✓

TABLE 2 (CONTINUED)

	With humans	Tests length generalization	Involves absolute grammaticality judgments	No negative evidence	Cannot be solved with bigrams	Cannot be solved by counting	Cannot be solved without attending to word order	No starting small	Successful for seen lengths
De Vries et al. (2008): *Hier-Scram* conditions	✓	✗	✓	✗	✓/✗	✓	✓	✓	✗
De Vries et al. (2012)	✓	✗	✗	✓	✓/✗	✓	✗	✓	✓
Fedor, Varga, and Szathmáry (2012)	✓	✗	✓	✗	✓/✗	✓	✗	✗	✓
Fenn, Brawn, Gentner, Margoliash, and Nusbaum (2007)	✓	✗	✓	✓	✗	✓	✓	✓	✓
Ferrigno et al. (2020)	✓	✗	✗	✓	✓	✓	✓	✓	✓
Friederici, Bahlmann, Heim, Schubotz, and Anwander (2006)	✓	✗	✓	✗	✓/✗	✗	✗	✓	✓
Hochmann et al. (2008): $(AB)^n$ items	✓	✗	✓	✓	✗	✓	✓	✓	✓
Hochmann et al. (2008): A^2B^3 and A^3B^2 items	✓	✗	✓	✓	✓	✗	✗	✓	✗
Lai and Poletiek (2010): *Starting small* condition	✓	✗	✓	✗	✓	✓	✗	✗	✓
Lai and Poletiek (2010): *Random* condition	✓	✗	✓	✗	✓	✓	✗	✓	✗
Lai and Poletiek (2011): *Starting small* condition	✓	✗	✓	✗	✓	✓	✗	✗	✓/✗
Lai and Poletiek (2011): *Random* condition	✓	✗	✓	✗	✓	✓	✗	✓	✗

Table 2 (continued)

	With humans	Tests length generalization	Involves absolute grammaticality judgments	No negative evidence	Cannot be solved with bigrams	Cannot be solved by counting	Cannot be solved without attending to word order	No starting small	Successful for seen lengths
Lai et al. (2016)	✓	✗	✓	✓	✓	✓	✗	✗	✓
Lai and Poletiek (2013)	✓	✗	✓	✓	✓	✓	✗	✗	✓
Mueller et al. (2010)	✓	✗	✗	✓	✓/✗	✓	✓/✗	✓	✓
Ojima and Okanoya (2020): *Single violation* items	✓	✗	✓	✗	✓/✗	✓	✗	✓	✓
Ojima and Okanoya (2020): *Repetition violation* items	✓	✗	✓	✗	✗	✓	✗	✓	✓
Ojima and Okanoya (2020): *Swapped violation* items	✓	✗	✓	✗	✗	✓	✓	✓	✓
Öttl, Jäger, and Kaup (2015)	✓	✗	✓	✓	✗	✓	✓	✓	✓
Perruchet and Rey (2005)	✓	✗	✓	✓	✓/✗	✓	✓	✓	✗
Poletiek et al. (2018): *Starting small* conditions	✓	✗	✓	✓	✓	✓	✗	✗	✓
Poletiek et al. (2018): *Random* conditions	✓	✗	✓	✓	✓	✓	✗	✓	✗
Rohrmeier and Cross (2009): Experiment II	✓	✗	✗	✓	✓/✗	✓	✓	✓	✓
Rohrmeier et al. (2012)	✓	✗	✗	✓	✓/✗	✓/✗	✓/✗	✓	✓
Uddén et al. (2009)	✓	✗	✓	✓	✓/✗	✓	✗	✓	✓
Udden, Ingvar, Hagoort, and Petersson (2012): Experiment 2	✓	✗	✓	✓	✓/✗	✓	✗	✓	✓

Table 2 (continued)

	With humans	Tests length generalization	Involves absolute grammaticality judgments	No negative evidence	Cannot be solved with bigrams	Cannot be solved by counting	Cannot be solved without attending to word order	No starting small	Successful for seen lengths
Winkler et al. (2018)	✓	✗	✗	✓	✓	✓	✓	✓	✓
Zimmerer et al. (2011), Zimmerer, Cowell, and Varley (2014), Zimmerer and Varley (2015): *Type 1* and *Type 2* violations	✓	✗	✓	✓	✓	✗	✗	✓	✓
Zimmerer et al. (2011), Zimmerer et al. (2014), Zimmerer and Varley (2015): *Type 3*, *Type 4*, and *Type 5* violations	✓	✗	✓	✓	✗	✓	✓	✓	✓

Table 2: Prior works involving A^nB^n artificial language learning in humans, but without testing length generalization. ✓ indicates that the experiment meets the requirement; ✗ indicates that it does not.

Start of Table 3; continued on next page

	With humans	Tests length generalization	Involves absolute grammaticality judgments	No negative evidence	Cannot be solved with bigrams	Cannot be solved by counting	Cannot be solved without attending to word order	No starting small	Successful for seen lengths
Cho et al. (2011), Cho et al. (2016)	✓	✓	✗	✓	✓	✗	✗	✓	✓
Fitch and Hauser (2004)	✓	✓	✓	✓	✗	✓	✓	✓	✓
Jiang et al. (2018)	✓	✓	✗	✓	✓	✓	✓	✓	✓
Malassis et al. (2020)	✓	✓	✗	✓	✓	✓	✓	✓	✓
Perfors et al. (2014)	✓	✓	✓	✓	✗	✗	✓	✓	✓
Poletiek (2002)	✓	✓	✓	✓	✓	✗	✗	✓	✗
Shin and Eberhard (2015): Experiment 1, *Incremental* condition	✓	✓	✓	✗	✓	✓	✓	✗	✓
Shin and Eberhard (2015): Experiment 1, *Random* condition	✓	✓	✓	✗	✓	✓	✓	✓	✗
Shin and Eberhard (2015): Experiment 2	✓	✓	✗	✗	✓	✓	✓	✗	✓
Shin and Eberhard (2015): Experiment 3, *Grammaticality judgment* condition	✓	✓	✓	✓	✓	✓	✓	✗	✗
Shin and Eberhard (2015): Experiment 3, *Completion* condition	✓	✓	✗	✓	✓	✓	✓	✗	✓/✗
Stobbe et al. (2012)	✓	✓	✗	✗	✗	✓	✓	✓	✓

Table 3 (continued)

	With humans	Tests length generalization	Involves absolute grammaticality judgments	No negative evidence	Cannot be solved with bigrams	Cannot be solved by counting	Cannot be solved without attending to word order	No starting small	Successful for seen lengths
Westphal-Fitch et al. (2018)	✓	✓	✓	✓	✓	✗	✗	✓	✓
Zuidema (2013): $(AB)^n$ items	✓	✓	✓	✗	✗	✓	✓	✓	✓
Zuidema (2013): A^nB^m items	✓	✓	✓	✗	✓	✗	✗	✓	✓

Table 3: Prior works involving A^nB^n artificial language learning that test length generalization in humans. ✓ indicates that the experiment meets the requirement; ✗ indicates that it does not.