The learnability and emergence of dependency structures in an artificial language

The learnability and emergence of dependency structures in an artificial language

Abstract:

In a pair of artificial language experiments, we investigated the learnability and emergence of different dependency structures: branching, center-embedding, and crossed. In natural languages, branching is the most common dependency structure; center-embedding occurs but is often disfavored, and crossed dependencies are very rare. Experiment 1 addressed learnability, testing comprehension and production on small artificial languages exemplifying each dependency type in noun phrases. As expected, branching dependency grammars were the easiest to learn, but crossed grammars were not different from center-embedding. Experiment 2 employed iterated learning to examine the emergence and stabilization of consistent grammar using the same type of stimuli as Experiment 1. The initial participant in each chain of transmission was trained on phrases generated by a random grammar, with the language produced by that participant passed to the next participant through an iterated learning process. Branching dependency grammar appeared in most chains within a few generations and remained stable once it appeared, although one chain stabilized on output consistent with a crossed grammar; no chains converged on center-embedding grammars. These findings, along with some previous results, call into question the assumption that crossed dependencies are more cognitively complex than center-embedding, while confirming the role of learnability in the typology of dependency structures.

1. Background: linguistic dependencies

Long-distance dependencies between constituents – e.g. between a subject and verb –are an essential aspect of human language and can be structured in different ways. Dependencies can be branching, center-embedded, or (rarely) crossed, as explained below. In natural languages, branching dependencies are generally, but not always, more prevalent than center-embedded dependencies, with variations in the range of phrase types which allow center-embedding (Hawkins 2004: 128). The third type, crossed or cross-serial, is rare in natural languages and the most computationally complex according to the Chomsky hierarchy (Chomsky 1963, Hunter 2021) – but there is some experimental evidence that this structure is easier to learn and process than center-embedded dependencies (Bach et al. 1986). Artificial language learning experiments (e.g. Culbertson 2012, Fedzechkina 2018) can shed light on the cognitive factors shaping language typology – for example, whether the prevalence of branching over center-embedded dependencies reflects greater learnability[1] of the former relative to the latter. In order to to explore how learnability reflects syntactic typological tendencies for dependency ordering, we ran two experiments, testing the learnability of different dependency types using nested locative NPs in an artificial language (Experiment 1), and the emergence of consistent dependency structure in iterated learning (Experiment 2).

1.1. Three ways of arranging dependencies

[1] Throughout this paper, we use the word *learnability* to mean "how easily a given structure can be learned by humans." This is different from the notion of learnability in mathematical linguistics, which is concerned with which formal classes of languages can be learned by particular learning algorithms (see for example Gold 1967).

There are three logically possible ways of arranging multiple dependencies. Figure 1 shows examples of the three dependency types between a subject noun and its verb: branching (in this case right-branching), center-embedded, and crossed.

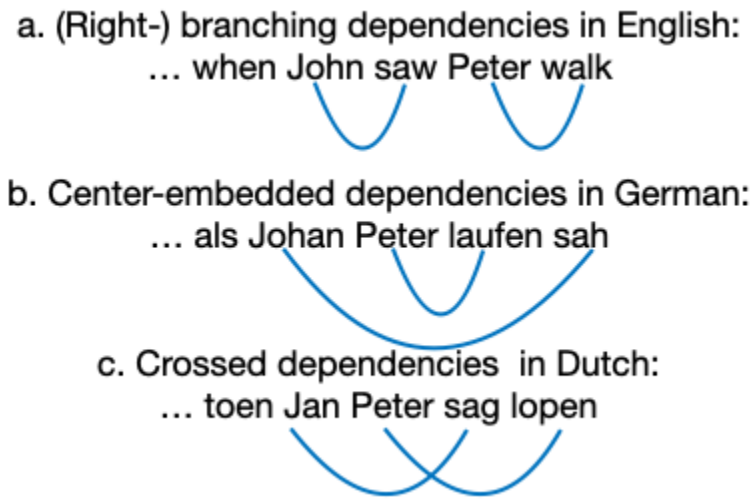


Figure 1: Comparison of dependency types with English, German, and Dutch examples, after Vosse and Kempen 1991. All sentences are glossed “when John saw Peter walk.”

1.1.1. Branching

Branching dependencies can be schematically represented as A1 A2 B1 B2; that is, each dependency (between A1 and A2, between B1 and B2) is resolved before the subsequent one is opened. In right-branching dependencies, the head of the phrase comes before its dependent; in left-branching dependencies the reverse is true. A simple example of English right-branching dependencies is shown in Figure 1a, and the relative clauses in (1) form a more complex example.

1. Jack walked the dog [which worried the cat [which chased the rat [which ate the malt]]].

English possessive NPs are left-branching, with the head noun on the right:

2. Jack's [friend's [cat's [mouse]]]

1.1.2. Center-embedding

Center-embedding refers to a linguistic structure wherein a dependency (B1-B2) is embedded inside another (A1-A2) of the same type, such that the components of A are on either side of B. Thus, the dependency structure would be A1 B1 B2 A2. This differentiates center-embedding from branching. Figure 1b shows an example from German, and 3 shows center-embedded relative clauses in English:

3. A. The rat [the cat chased] ate the malt

   B. The rat [the cat [the dog worried] chased] ate the malt

   C. The rat [the cat [the dog [Jack walked] worried] chased] ate the malt

1.1.3. Crossing

In addition to left and right branching and center-embedding, subject-verb dependencies can also be crossed (Figure 1, example C). In this pattern, the first elements of all dependencies are concatenated, and the second elements are in the same order (A1 B1 A2 B2). This phenomenon is rare in natural languages; few examples are attested, specifically Dutch, Swiss German (Schieber 1985), and Tagalog (MacLachlan and Rambow 2002). A complex example from Dutch is shown in Figure 2, illustrating the crossed dependency pattern that appears in subordinate clauses with infinitives (Bresnan et al. 1982).

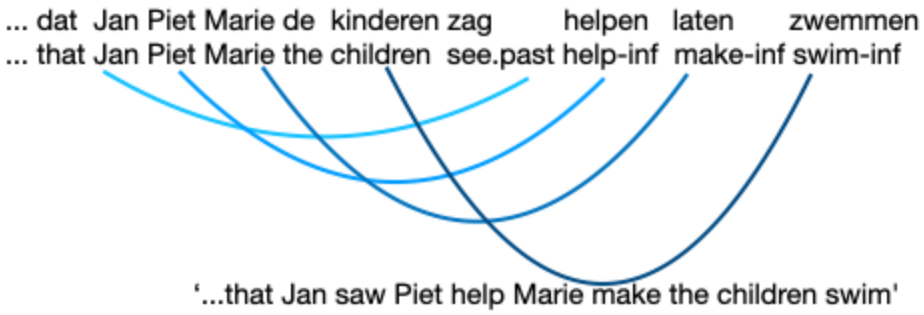


Figure 2: Complex Dutch sentence with crossed dependencies, after Bresnan et al. 1982

In some Dutch dialects, crossed dependencies between infinitive verbs and their subjects are grammatical while equivalent center-embedded ones are ungrammatical, and in others, both are allowed but crossed dependencies are more common (Steedman 1984: 52). Crossed noun-verb dependencies also occur in Swiss German (Huybregts 1984), and in the Austronesian language Tagalog (MacLachlan and Rambow 2002), the official language of the Philippines. Here there is a crossed dependency between the sentence-initial verbs and their argument nouns:

4. [Nagisip [na bumili si Pedro ng bulaklak]]
   AT.thought LK AT.buy NOM Pedro flower
   'Pedro thought to buy (of buying) a flower.' (LK = linker, AT = agent topic)

Generally speaking, center-embedded dependencies are more common than crossed dependencies in natural languages (Fodor 1978). In some cases, both crossed and center-embedded dependency readings of the same sentence can be possible, in which case the center-embedded reading is preferred (Dalrymple and King 2013). There seem to be no languages that allow only crossed dependencies in all types of phrases, or in which crossed dependencies are the default and center-embedded dependencies are exceptional (Steedman 1984: 36).

1.2. The cognitive bases of dependency structure: learnability and processing

Several cognitive factors related to learnability and processing may influence the distribution of dependency structures in natural languages, namely the prevalence of branching and comparative avoidance of center-embedding and crossed dependencies. Language typology seems in at least some cases to reflect the learnability and/or processability of different structures, as demonstrated through artificial language learning experiments.

Culbertson (2012) offers an overview of learnability experiments in relation to language universals, which generally show that more typologically common structures are also more learnable. For instance, in one such experiment, Culbertson et al. (2012) trained participants on simple artificial languages with inconsistent mixed rules for the ordering of nouns, adjectives, and numerals, and examined which orders participants produced when asked to label stimuli. The participants' learning biases reflected natural language word-order universals as described by Greenberg (1963), specifically a bias against the combination of adjective-noun and noun-numeral orders, which is rare in natural languages. In addition, while the participants were all English speakers, they did not preferentially replicate English word order (numerals and adjectives both prenominal). Culbertson et al (2020) also demonstrated that artificial languages with typologically more common features ("harmonic" word order patterns) are more learnable even for speakers of languages that do not follow the typologically typical order. This tendency for typology to correlate with learnability even holds for cases where the artificial language uses features not found in the subjects' native languages; for example, speakers of English (which lacks noun case marking but has number marking) and Japanese (which has case marking but no plural marking) both strongly preferred the typologically common ordering of number before case (Saldana et al 2021). While learnability is of course not the only factor in language typology-- historical contingency clearly plays a large role-- it is likely one major factor (Culbertson et al. 2012). In addition, processing difficulties may be relevant to typological frequency, either as a

factor in themselves or as a “filter” on learnability; i.e., if the input cannot be processed in the first place, it cannot be used as a basis for learning (Kirby 1999). However, as discussed further below, crossed and center-embedded dependencies may form a counterexample to the generalization that the most learnable features will also be typologically common.

1.2.1. Dependency length minimization

One specific cognitive factor potentially involved in dependency structure typology is minimization of dependency length. The processing difficulties presented by multiple center-embedding are well-attested (e.g. Blumenthal 1966, Fodor and Garret 1967, Blauberg et al.1974, Foss et al. 1970) and this difficulty likely arises from working memory load. The fact that written language features more prolific and deeper center-embedding than spoken language (Miller and Weinart 1998; Sakel and Stapert 2010; Karlsson 2007, 2009, 2010; Levison 2014) suggests that these structures pose a difficulty for working memory that is mitigated in writing and/or reading.

This working memory load may be attributed to the length of the dependencies involved: the distance between two syntactically related words (e.g. noun and verb, noun and adposition), and also the number of unresolved dependencies which must be maintained mid-sentence. During on-line sentence processing, multiple center-embedded phrases like 3C above present persistently unresolved dependencies, which challenge working memory more than a right- or left-branching structure (1) in which dependencies are resolved sooner (see Figure 1 for an illustration of different dependency lengths). Languages have a general tendency to minimize total dependency length and the number of dependencies open during sentence processing (e.g. Hawkins 1994, Futrell et al. 2015), which often rules out center-embedding, although long-distance dependencies are never eliminated entirely (Liu et al 2017). Dependency length minimization thus favors right- or left-branching phrases.

Languages have a general tendency to avoid center-embedding (thus lowering dependency length and cognitive load) by selecting combinations of word order parameters that

prevent it (Kuno 1974), or by avoiding such orders where they are grammatical. For example, Persian has obligatory extraposition of complement clauses which would otherwise be center-embedded (Dryer 1980), and Japanese has an optional scrambling rule (Lewis and Nakayama 2001). English grammatically permits both center-embedded sentences like 3c and branching dependencies like 1, but the paraphrase 1 is preferred over 3c because it is easier to process.

Kuno (1974) predicts that combinations of word orders forcing center-embedding, such as noun-initial noun phrases with postpositions, should not exist, as it would create a "hopeless situation of center-embedding and juxtaposition of prepositions". However, some natural language data suggest that these constraints are not absolute. Hagege (2010), cites counterexamples from the Nilo-Saharan language Moru, in which just such a "hopeless" pileup of adpositions occurs, and yet is fully acceptable:

5. [kokyɛ uni [toko dasi [odrupi ma ro] ro] ri] drate
   dog blackness woman main brother 1SG of of of die.PST
   'the black dog of the main wife of my brother is dead' (Tucker 1940: 165)

The corresponding right-branching, non-center-embedded version (with postpositions right after their head nouns) is not grammatical:

6. *kokyɛ uni toko dasi ri odrupi ro ma ro drate
   dog black woman main of brother of 1sg of die.PST

Other Nilo-Saharan languages show the same pattern (Hagege 2010: 32). The equivalent noun-final order with prepositions is also attested in Ancient Greek:

7.

to

the.ACC
tees
the.GEN
tou ksainontos
the.GEN wool.carder.GEN
tekhnees
art-GEN
ergon
work-ACC
'The work of the wool-carder's art' (Plato, the Statesman, 281a) (from Hudson 1996)

Hawkins (1994: 8) also notes that SOV word order, as in Japanese, Korean, and Turkish, forces center-embedded clauses; these are grammatical and acceptable, whereas similar sentences in English are generally rejected by native speakers.

Hawkins (1994) proposes a cross-linguistic hierarchy of permitted center-embedding: NP > VP > S (1994: 14), wherein categories that are typically longer and more complex are less embeddable (1994: 23). Natural language data like that cited by Hagege (2010) further suggest, in support of Hawkins' thesis, that whatever cognitive factors make center embedding difficult apply less strongly to NPs than to sentences.

There is also evidence linking learnability to dependency length minimization. Fedzechkina et al. (2018) found that, given an artificial language learning task with long dependencies between words, learners restructured the input to reduce dependency lengths by rearranging the order of short and long constituents, not necessarily recapitulating the syntax of their native language. This study did not explicitly address center-embedding of constituents within those of the same type, but as previously discussed, dependency length is likely a relevant factor in the processing of center-embedding versus branching. Our experiments set out to explicitly test learnability differences between dependency types.

However, it is unclear how crossed dependencies fit in with a dependency length minimization account. Like center-embedding, crossed dependencies create a longer total dependency length than right or left branching (Gómez-Rodríguez and Ferrer-i-Cancho 2017, Ferrer-i-Cancho 2006). The total dependency length is the same for comparable center-

embedding and crossed dependencies (and both are longer than branching), as can be seen by inspecting Figure 1, so on these grounds there is no reason either should be favored.[2] Another possible factor may be computational complexity. Crossed dependencies are more complex than center-embedded dependencies on the Chomsky hierarchy, because they require context-sensitive rewrite rules in addition to those that do not rely on context, while center-embedded dependencies can be produced with a more restrictive set of rules (Hunter 2021: 78). While center-embedded dependencies can be produced by a context-free grammar, crossed dependencies require a more complex (mildly) context-sensitive grammar (Öttl et al 2015, Partee et al. 1990: 505)[3]. Under the assumption that the dependency patterns that are more commonly found in human languages are the less computationally complex ones, the Chomsky hierarchy is consistent with natural language data in that center-embedded dependencies seem to be more common than crossed dependencies. As will be shown below, however, the Chomsky hierarchy does not appear to clearly map onto the real cognitive difficulty of the associated dependency structures (see also Chesi and Moro 2013), and the present experiments provide further evidence against a straightforward mapping.

1.2.2. Processing and learnability of dependency structures: Experimental evidence

1.2.2.1. Processing of natural language

Some experimental evidence (Bach et al. 1986) suggests that crossed dependencies are easier to process than center-embedded dependencies, which is surprising in light of both the marginal status of the former compared to the latter in natural languages.

---

[2] However, in a crossed-dependency pattern, the maximum length of individual dependencies is shorter than in center-embedding constructions. Therefore, if dependency length minimization is taken to set a limit on individual dependency lengths as well as total length, crossed dependencies would be favored over center-embedded. Thanks to an anonymous reviewer for this observation.

[3] But see Pullum and Gazdar (1982) and Gazdar and Pullum (1985) for discussion of whether the presence of crossed dependencies really demonstrates that a natural language is not context-free, with consideration of some specific examples and arguments.

In a comprehension experiment, Bach et al. (1986) found that Dutch speakers could comprehend crossed dependencies more easily than German speakers could comprehend equivalent center-embedded dependencies in their native languages. The difficulty of comprehension between two-verb and three-verb sentences increased sharply for both dependency types, but more so for center-embedding than for crossed dependencies. Right-branching paraphrases were found to be easier than either. These findings held for both subjective impressions of comprehensibility and performance on a comprehension test. However, the authors consider that some non-syntactic features of Dutch and German may have influenced sentence processing, namely case marking (which was factored out of the stimuli but could still influence the strategies adopted by speakers) and the stress patterns on the verb clusters (Bach et al. 1986: 261). Therefore it is not certain how their results generalize beyond these languages, or whether comparing across languages in this way is even a valid comparison.

There are several possible explanations for the greater ease of crossed dependencies. For instance, crossed dependencies may be more tractable in speech than center-embedded dependencies because they make it easier to integrate useful structures in real time -- i.e. in a crossed sentence of the form *N1 N2 N3 V1 V2 V3* one can construct the meaning “N1… V1” as soon as the verb cluster begins. The equivalent is not possible with a center-embedded structure (*N1 N2 N3 V3 V2 V1*), as it would be necessary to hear the whole before V1 is available to match up with N1. While “N3 V3” is available when the verb cluster begins, the listener still must wait to determine how this information is incorporated into the greater structure of the sentence, rather than having the more global information supplied by V1 (Bach et al. 1986: 260-1). While overall dependency length is comparable between equivalent crossed and center-embedded strings, the particular arrangement of dependencies may make processing easier for crossed -- because, as noted above in footnote 3, maximum dependency length is shorter than in a center-embedded phrase. The prevalence of center-embedding in natural languages, compared to crossed dependencies, is therefore hard to explain purely on the basis of differences in processing.

1.2.2.2. Sequence learning and artificial language learning

There is also evidence from nonlinguistic sequence learning and artificial grammar learning concerning the greater ease of crossed dependencies compared to center-embedded dependencies. However, how well these findings generalize to natural language is not clear.

Conway et al. (2003) confirmed that humans have more difficulty learning center-embedded patterns in meaningless visual or auditory sequences, compared to branching dependencies. In this experiment, the dependencies were matched pairs of letters. In another foundational experiment, Fitch and Hauser (2004) found that humans are easily capable of learning a center-embedded grammar of the form AnBn). However, while Fitch and Hauser (2004) analyzed the AnBn grammar as having center-embedded dependencies between A and B elements, this is not the only possible analysis: since any A element can form a dependency with any B, such strings can also be understood as crossed. In addition, the strings can be processed through count-and-match (N A's, N B's) with no need for dependencies between elements (Corballis 2007, Rogers & Pullum 2011). Perruchet and Rey (2005) attempted to replicate Fitch and Hauser's findings, but discovered that subjects did not succeed in learning the pattern when required to recognize center-embedding explicitly. Therefore, it is not entirely clear whether humans can easily learn and process center-embedded nonlinguistic strings.

Further experiments (De Vries et al. 2008, Uddén et al 2012) employed stimuli in which specific A and B elements formed dependency pairs. In the stimuli used by De Vries et al., dependencies consisted of specific pairs of nonsense syllables, such as "de… bo" or "gi... fo"; wherein A syllables had a front vowel and B syllables a back vowel. Their results showed that participants who learned the specific pairings could not easily distinguish hierarchical embedded sequences (of the form A1 A2 A3 B3 B2 B1) from superficially similar ones with the same elements ordered incorrectly (e.g. A1 A2 A3 B3 B1 B2). The results suggested that participants were attentive to surface features of the strings -- i.e. all As come before all Bs, and the number

of As matches the number of Bs -- but not dependency order. Uddén et al. (2012) performed a similar experiment with visually presented sequences of letters that formed arbitrary dependency pairs (e.g. F…L or D…P), and found that crossed dependencies were more easily learnable than center-embedded patterns. Fitch and Friederici (2012), reviewing this line of research, also point out that a string of the form AnBn can be analyzed several different ways (e.g. count-and-match, embedding, or crossed dependencies), and the ability to distinguish matched from mismatched sequences does not reveal which strategy is actually being used. Therefore, it is questionable whether recognizing or creating AnBn strings requires a grammar with multiple dependencies, center-embedded or crossed. To test learnability or comprehensibility of dependency structures, it is necessary for the stimulus strings to have some clear indication of how elements link to form dependencies.

Other sequence-learning studies have explored differences in learnability between center-embedded and crossed dependencies, using strings where dependencies are indicated by some similarity between elements. Vogel et al. (1996) predicted that crossed dependencies should be easier to learn and process, as they are amenable to a count-and-match strategy, while recognizing that center-embedded patterns require the additional operation of reversing half the string. However, experimental results do not clearly support this prediction. Öttl et al. (2015), for example, trained human participants on strings of nonsense syllables instantiating center-embedded or crossed dependencies. Syllables could fall into two categories depending on their vowel: A (vowel [e] or [i]) and B (vowel [o] or [u]). The dependencies were formed of syllables with the same initial and final consonant; e.g. “del” and “dol” or “tix” and “tux”. Strings could follow one of two grammars: center-embedded (A1 A2 A3 C B3 B2 B1) or crossed (A1 A2 A3 C B1 B2 B3). They found that there was no clear difference in learnability between grammar types. Again, this is unexpected given the typological distribution of these patterns, with crossed dependencies being much rarer than center-embedding. However, findings concerning nonlinguistic stimuli with dependencies indicated by phonological similarity may or may not generalize to artificial language

stimuli with semantics. The present experiment addresses this issue with semantically meaningful stimuli.

1.2.2.3. Our experiments

The stimuli in these previous studies of center-embedded and crossed dependencies were meaningless strings of sounds or syllables (i.e. not corresponding to any referent such as pictorial stimuli), in which cues to dependencies were absent or unclear. In the present study, we attempt to test for differences in learnability between dependency structures, and the role of learnability as a factor in language typology, using an artificial language task in which strings of novel words are associated consistently with visual stimuli which show the hierarchical relationship between elements described in the artificial language description. Previous studies suggest that participants may not readily perceive dependencies in nonlinguistic strings, but adding semantics in the form of associated visuals should make dependencies between elements clearer than they would be in either AnBn strings, or strings in which dependencies are indicated by a perceptual similarity, e.g. alliterative syllables.

In our experiments participants learn an artificial language consisting of nested adpositional phrases. We used NPs and adpositional phrases rather than clauses because, per Hawkins (1994), center-embedding is more tolerated in NPs than clauses, and also because it is easier to clearly represent spatial relations (e.g. *the cat on the mat next to the table*) than transitive verb actions (e.g. *the cat chased the rat that ate the malt*) in a static image.

First, we assessed the learnability of crossed, center-embedded, and branching dependencies; while previous studies have addressed the processing cost of different dependency types (Bach et al. 1986) or the learnability of long-distance dependencies (Fedzechkina et al. 2018), none has systematically compared the learnability of all three dependency structures in the same paradigm, or used artificial languages, which would eliminate the confounds introduced by natural languages like Dutch and German. The second experiment

used an iterated learning methodology to explore the development of consistent syntactic structures from initially random input; by iterating the language through multiple participants, individual learning biases are amplified and clarified. In both experiments, participant productions were analyzed to observe how well the input was replicated and how it was transformed through learning.

2. Experiment 1: Learnability

In this experiment, we set out to test the learnability of different dependency types in a small artificial language, in order to see which dependency orderings could be learned accurately and generalized to descriptions of items not seen in training. In addition, we compared participant-produced descriptions to different grammars to see how participants were interpreting or transforming the grammar – for example, by converting center-embedding into the apparently easier crossed grammar, or flipping head noun order.

2.1. Methodology

2.1.1. Participants

This study was conducted online; participants were Amazon Mechanical Turk workers who were compensated financially for their time ($5 per task). The task was advertised through mTurk as an “Alien Language Game.” Only participants who met the following criteria could work on the task: US residents, minimum approval rate of 97%, and at least 1000 HITs completed. All were native English speakers according to self-report, though some spoke other languages. Participants were assigned randomly to one of six syntactic order conditions, as discussed below. In total, 140 participants were tested, 20 of whom were excluded because they took written notes (see below).

2.1.2. Stimuli

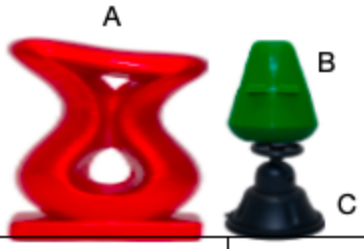
Stimuli were composed of images of four unfamiliar objects from the NOUN database (Horst 2015), arranged into scenes of 1-4 objects and labeled in an artificial language (Figure 3). This language consisted of two adpositions: *rae* “atop” and *moy* “to the (viewer’s) left of,” and four nouns, one for each object. Ten sets of four nouns were used, with each participant being allocated to a randomly-chosen set. Nouns were generated from a set of syllables of either CV or CVC structure, with four possible onset consonants (*t*, *v*, *s*, *k*), four vowels (*a*, *o*, *i*, *u*), and three possible endings (no consonant, *l*, *n*). Each noun had two syllables, such that both syllables had to share exactly two of the following: onset consonant, vowel, and final consonant (i.e. the syllables could not be fully identical). Within each set of four, each noun began with one of the onset consonants. This internal reduplication was used to make each noun more distinctive and easier to learn (cf. Ota and Skarabela 2016). In addition, each of the four nouns was required to have a mean Levenshtein distance of at least 5 from the others. Nouns matching English words (e.g. *salsa*, *tiki*) or which sounded like English words (e.g. *kuki* = *cookie*) were excluded. This produced sets of four distinct, recognizable novel nouns (examples in 7).

8. Example noun lexica:

   kilkul, vanva, tovo, sunsin

   kuka, vinkin, tultol, solsil

   kakan, volsol, tinkin, sulsal



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

Figure 3. Sample scene with description in each of the six syntactic orders. Objects are labeled with letters for clarity here and in other example scenes; no such labels were used on stimuli for the actual experiment.

The syntax of the language could follow one of six combinations of NP order and adpositional phrase order (cf. Kuno 1974: 127-8): two branching grammars (an English-like grammar with noun initial NPs and prepositions, and the reversed grammar with noun final NP with postpositions), two crossed-dependency grammars (with noun-initial NPs and postpositions, or with noun-final NP and prepositions)[4], and two center-embedding grammars (noun initial NPs with postpositions [N1 [N2 N3 Adp2] Adp1], noun final NPs with prepositions [Adp1 [Adp2 N3 N2] N1]). Examples

[4] There are two other logically possible crossed grammars: noun-initial with prepositions (Adp1 Adp2 N1 N2 N3) and noun-final with postpositions (N3 N2 N1 Adp2 Adp1). In these orders, the head noun would be at the middle of the phrase rather than the periphery; we were concerned that this could have introduced some additional confounds, e.g. by making the head noun less salient, so these orders were not included.

are shown in Figure 3. Head order was manipulated to investigate whether unfamiliar syntactic order in general, as opposed to the dependency order per se, negatively affected performance. A pilot experiment (detailed in Appendix 2, section 1) using only center-embedded and branching grammars, and in which results were free-typed, yielded similar results to those of the presently discussed experiment.

2.1.3. Procedure

Before the task, participants were asked not to take written notes ("Please do not take notes during the task! We are interested in what your brain can do, not what your brain plus a notepad can do."). At the end of the experiment, alongside other demographic questions, participants were asked whether they had taken written notes ("Did you write stuff down or take notes during the task? Please be honest - it won't affect your payment, we promise, and if you tell us now we can correct for this in our analysis without affecting the validity of the experiment."). Participants who indicated at this point that they had taken written notes were paid in full but their data were excluded from analysis. 20 participants were excluded in this way.

The experiment was organized into three main phases: an initial vocabulary training phase, during which participants were trained to criterion on the vocabulary items (nouns and adpositions), a multi-word expression training phase, during which participants were trained on multi-word labels describing scenes featuring multiple objects, and then a final test phase, in which participants were tested on their ability to produce descriptions of the scenes.

Phase 1: Vocabulary Training

In the initial vocabulary training phase, participants were trained on the meanings of the four nouns labelling the four novel objects, and then the adpositions. Participants were given training

trials in which an adposition or noun was presented passively, and comprehension trials in which participants were prompted with an adposition or noun and required to select the matching image. Example noun training and comprehension trials are shown in Figure 4; example adposition training and comprehension trials are shown in Figure 5.

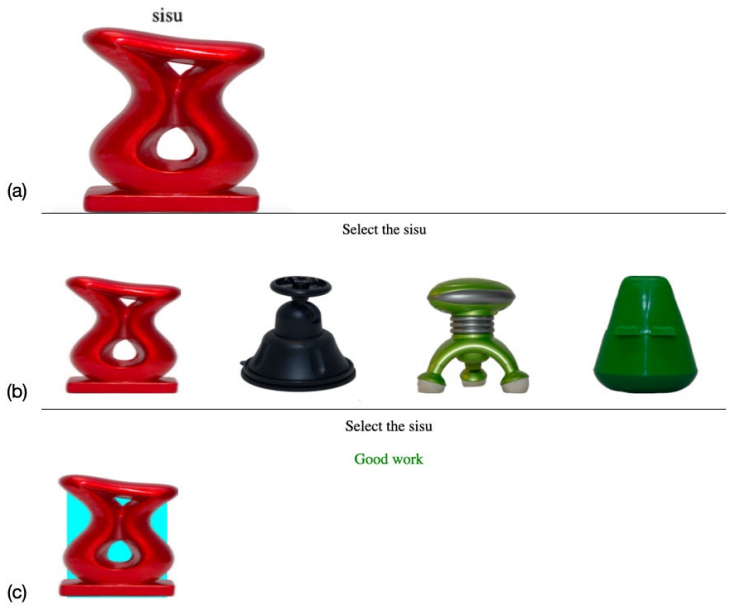


Figure 4: Noun training (a), comprehension (b), and feedback on comprehension (c)

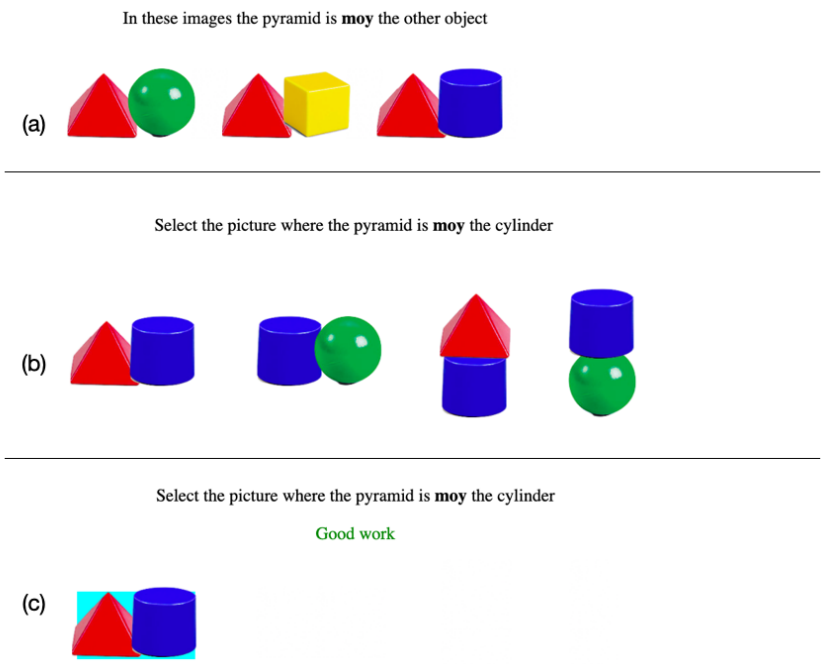


Figure 5: Adposition training (a), comprehension (b), and feedback on comprehension (c)

In order to ensure that the meaning of the artificial adpositions was clearly indicated, the adposition was used in an English descriptive sentence accompanying several example scenes; pilot experiments showed that adposition meaning was not always easily inferred from complex scenes and their associated artificial language descriptions (see Appendix 3). Training trials were self-paced; participants advanced to the next trial by button press.

Comprehension trials were structured as follows: the participant was shown the four novel objects (in the case of noun trials) or four spatial configurations of English-named objects (in the case of adposition trials) and asked to select the object or spatial configuration corresponding to a noun or adposition, respectively. After making a selection, participants were shown feedback indicating the correct selection. If they had selected correctly, the text “Good work” appeared; if they were

incorrect, the text “Sorry, incorrect” appeared. In either case, all scenes except the correct one (corresponding to the description) disappeared.

In the noun training phase, participants received four noun training trials (one per object), followed by four noun comprehension trials, and were trained to criterion (i.e. the training-comprehension sequence was repeated until all were answered correctly). Adposition training followed the same structure.

Phase 2: Multi-word description training

In the second phase of the experiment, participants were trained on multi-word descriptions for scenes involving either two or three objects. This phase progressed in difficulty: participants were given eight training and eight comprehension trials on scenes featuring two objects, then eight training and eight comprehension trials on scenes featuring three objects. As in the vocabulary training phase, participants received both passive training trials and comprehension trials (see Figure 6 for sample training and comprehension trials for a two-object scene); unlike in the vocabulary training phase, progression did not depend on performance on comprehension trials.

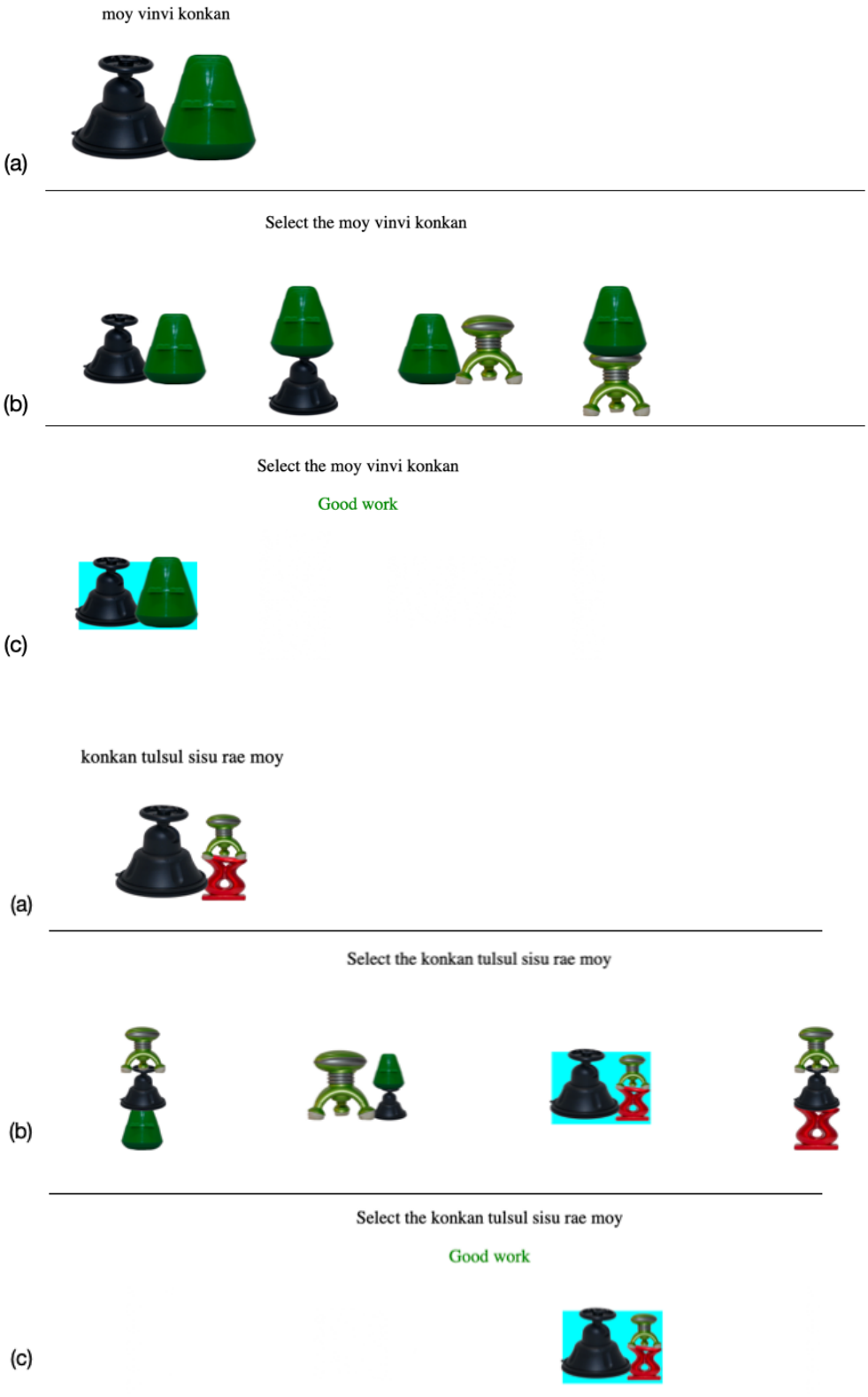


Figure 6: Phrase training (a), comprehension (b), and feedback on comprehension (c).

Training trials followed the same structure as in the one-word phase. After training trials were completed for one phase, participants would be prompted to move onto comprehension trials.

In comprehension trials for multi-word descriptions, participants were given a description in the target language and asked to select the appropriate object or configuration of objects from four possibilities with the same number of objects as the target. This set contained the target and three distractors:

1. The same objects as the target image, but in a different configuration.
2. Different objects from the target scene, but in the same configuration as the target. .
3. Different objects in a different configuration: the objects from distractor 2 in the configuration from distractor 1.

The four options (target plus three foils) appeared in random order (Figure 6b). Feedback on correct and incorrect responses was the same as for noun and adposition training.

Phase 3: Production test

Finally, participants moved on to a production test (Figure 7) in which they produced descriptions for scenes. Again, these trials progressed in difficulty, with four levels (one-object through four-object scenes). For levels 2 and 3, participants saw the eight scenes that they had been trained on, and four novel scenes; level 4 consisted of 12 novel four-item scenes. On each trial,

participants were presented with an image of a scene and provided with a set of six buttons[5] corresponding to their randomly selected lexicon: four two-syllable nouns plus *moy* and *rae*. Pressing each button entered its word into the text box, and participant entries were required to exactly match a required length: 2n - 1 for n objects, e.g. five words for an a scene of three objects. Once the required number of words was entered, the participant automatically progressed to the next trial.

Figure 7: Production trial at level 3: participants see a scene, and generate a description by clicking vocabulary buttons. A partially built description is shown.

Post-experiment demographics and debriefing

Following the production task, participants were given a brief survey in which they were asked to indicate their native language, list other languages they spoke, and provide any additional comments they had on the task. We also assessed whether they understood the meaning of the artificial adpositions. For each adposition, participants were given a multiple choice question from which they were required to select one of the following meanings: “To (your) left of”; “To (your)

[5] In the pilot experiment (see Appendix 1), responses were collected using a free text entry box; results were overall similar to the present experiment.

right of"; "Horizontal (either left or right)"; "Atop"; "Under", "Vertical (either atop or under)"; "Don't know"; and "Other" (with a free text entry box).

2.1.4. Analysis: Calculating accuracy for test trials and best-fit grammar

Comprehension performance (i.e. whether participants selected the correct referent at each comprehension trial during training) was analyzed with a logit mixed effects model, to give a measure of learning accuracy during training. Level was coded with successive differences contrast coding.

Accuracy for the final set of production test trials was measured by calculating the normalized optimal string alignment (optimal string alignment divided by length of the longer string, calculated with the R *stringdist* package, Van der Loo 2014) against the correct description in the language the participant was trained on, and subtracting the resulting normalized distance from 1 to get an accuracy score (an accuracy score of 0 indicates no correspondence between the participant-produced description and the trained label; an accuracy score of 1 indicates perfect correspondence). However, using raw string edit distance on the produced labels would result in a non-uniform penalty for errors: nouns are longer than adpositions, therefore noun errors would be penalized more heavily. Furthermore, nouns vary in length, and thus not all noun errors would be weighted equally. We therefore replaced each word in each string to be compared with its initial letter prior to calculating string edit distance (e.g. *rae vuntun vuntun* > *rvv*), ensuring that all errors were weighted equally. The resulting correctness score for each participant-produced description was used in a mixed effects model to analyze production test performance.

In order to assess how well grammatical features were preserved in different conditions (and generalized to new stimuli), and how grammars were modified through learning, we estimated which of many possible grammars was most likely to underlie each string in each participant's productions. We considered a set of 64 possible grammars. Grammars varied in their

dependency type and word order, and in the semantics associated with the adpositions. Grammars could have the following orders: branching, center-embedded, or crossed dependencies. Branching and center-embedded grammars could be noun-final or noun-initial, and crossed dependencies could have four total combinations of noun positions (initial or final), and pre- or postpositions. This gives eight possible word orders (including two crossed orders not represented in the stimulus conditions: noun-initial with prepositions and noun-final with postpositions; see footnote 2 in section 2.1.2). In addition, there were four possible combinations of correct and reversed interpretations for the adpositions *rae* and *moy*: 'atop'/ 'below' and 'left of'/'right of' respectively, and the possibility that the vertical and horizontal adpositions could be switched in meaning (e.g. *rae* = 'left of', *moy* = 'atop').[6] This produced eight possible adposition interpretations.

For each scene seen by a participant, the corresponding string in each of the 64 grammars was generated, using the initials of the object labels that participant had seen. These strings were then compared to the participant-produced description, using optimal string alignment and normalisation as described above, to allow us to measure how well the string from each of those 64 grammars matched the participant's production for that scene. Possible grammars were further narrowed down using the participant's stated interpretation of adpositions from the post-task survey. For example, if a participant indicated that *moy* meant "atop", all grammars featuring another meaning for that adposition were eliminated from consideration. Only directionally-specific glosses allowed candidate grammars to be removed; unspecified (e.g. the "either left or right" answer) or uninterpretable ones (e.g. a filled-in answer that did not correspond to any identifiable meaning) did not eliminate any grammars. In many cases, multiple grammars were

[6] Grammars with directionally unspecified adposition glosses ("next to" for horizontal or "vertically adjacent to" for vertical) were not included because they would be inherently indeterminate as to head order. If these grammars were allowed, all strings matching a grammar with a specific adposition interpretation would also match the nonspecific one, causing nonspecific grammars to win out over specific ones.

applicable to the same string. For example, a noun-initial grammar with the trained adposition meanings (*moy* = left of, *rae* = atop) would produce identical strings to a noun-final grammar with the opposite adposition meanings (*moy* = right, *rae* = under), as shown in Figure 8. It was not always possible to eliminate one of these competing grammars if participants had not specified their adposition interpretations sufficiently.

Another ambiguity involved strings of the form *N1 N2 N3 Adp1 Adp2*, which could be interpreted as either a crossed or a center-embedded grammar, if the string of adpositions was palindromic or all were identical. Indeterminacy between crossed and center-embedded could also occur in three-word descriptions of the form N1 N2 Adp, if the adposition directionality was not specified. These were labeled as their own category ("center-embedded/crossed"). (See Figures a1 and a2 in Appendix 1.)

2.2. Results

Full summary tables for all statistical analyses can be found in Appendix 2, Section 3. Analysis code can be found at https://github.com/EmilyDavis47/learnability-emergence

2.1.1. Comprehension

We analyzed comprehension testing accuracy for scenes of two and three objects (Figure 8). The single-object and adposition levels, which were trained to criterion, are excluded. Comprehension performance was analyzed with a logit mixed effects model with three fixed effects: dependency type (branching, crossed, center-embedding), head noun position (initial, final) and level (two objects or three). Dependency type was coded with Helmert contrasts, yielding a model with two dependency type contrasts, comparing crossed to center-embedding and then branching to the

mean of crossed and center-embedding. Head position and level were sum coded. The model included by-participant random intercepts and by-participant random slopes for level.

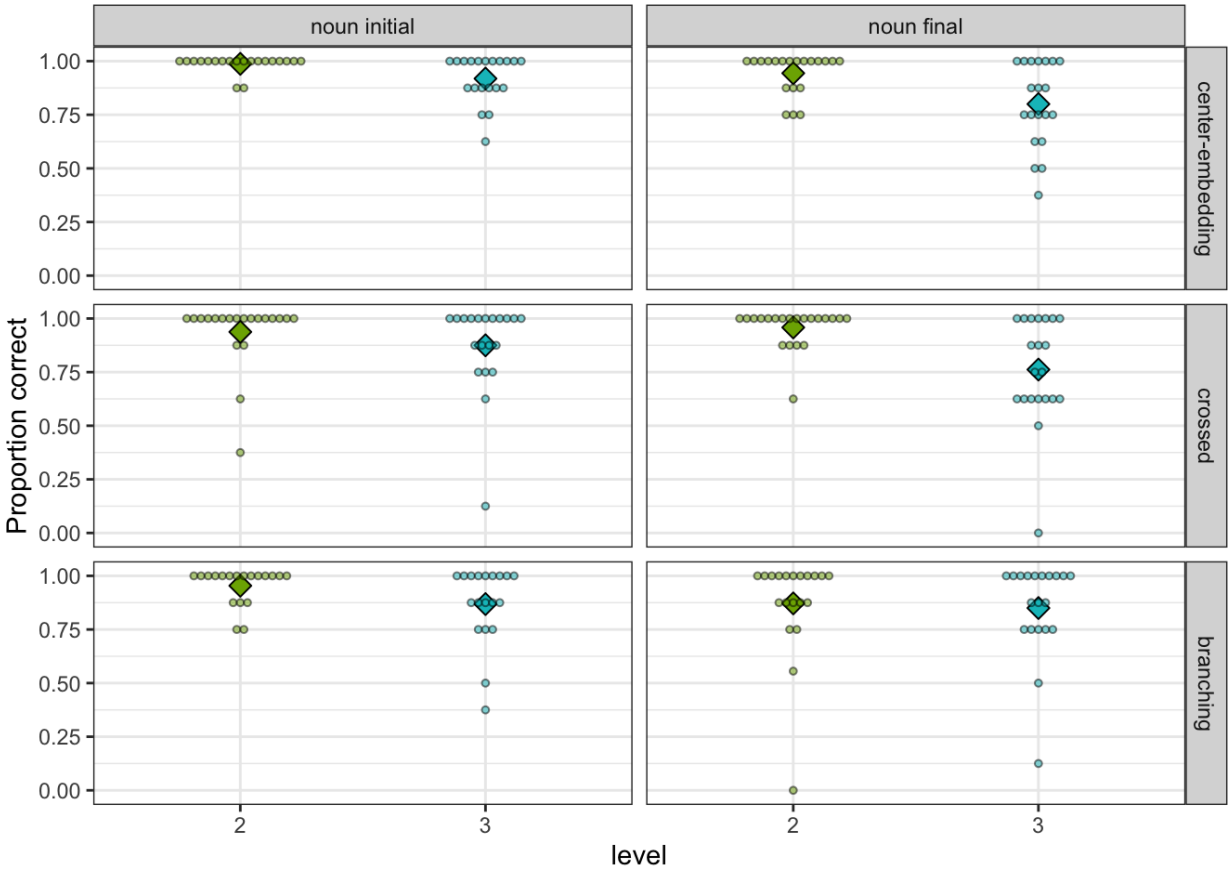


Figure 8: Performance on comprehension task, by condition and level (two and three objects). Each point represents one participant's performance at that level. Total n = 20 per cell. Large diamonds represent mean performance across all participants.

Performance in comprehension trials was generally very high; the mean comprehension for crossed and center-embedded conditions was marginally lower than for branching (B = -0.25, SE = 0.32, p = .14). Comprehension accuracy was marginally lower in the noun-final condition than the noun-initial condition (B = -0.43, SE = 0.25, p = .088), and was significantly lower at the three-item level than the two-item level (B = -1.7, SE = 0.38, p < .001). In addition, a significant

interaction occurred between dependency type and level, where performance declined more between levels in the crossed and center-embedded conditions than in the branching conditions (B = 0.33, SE = 0.17, p = .049). There were no significant interactions between level, dependency type and initial or final order.

2.2.2. Production accuracy

Figure 9 shows participant accuracy in the final-phase production task, across all levels.

In the first analysis, covering all data, dependency type (Helmert coded), headnoun order (sum coded), and level (successive differences) were fixed effects. As expected, branching grammars were the easiest to learn: the mean learning accuracy on center-embedded and crossed dependencies was significantly lower than for branching structures (B = -0.02, SE = 0.001, p = .04). Within the two harder dependency types, learning accuracy for crossed grammars was not significantly different from center-embedding grammars (B = -0.01, SE = 0.02, p = .52). There was also a significantly greater decline with level in noun-final than in noun-initial conditions (B = -0.03, SE = 0.01, p < .001). The effect of noun order was not significant and there were no other significant main effects or interactions. [7]

[7] For all scenes that were seen in training, there were no significant effects of dependency type or head order except that performance on head-final orders was significantly lower (B = -0.03, SE = .01, p = .01, all other p > .1), nor interactions between syntax type and other variables. For novel items, there was no significant main effect of dependency (p > .5 for both comparisons); however, performance was again significantly lower in head final (B = -0.07, SE = 0.02, p < .001) See appendix tables for full details of these analyses.

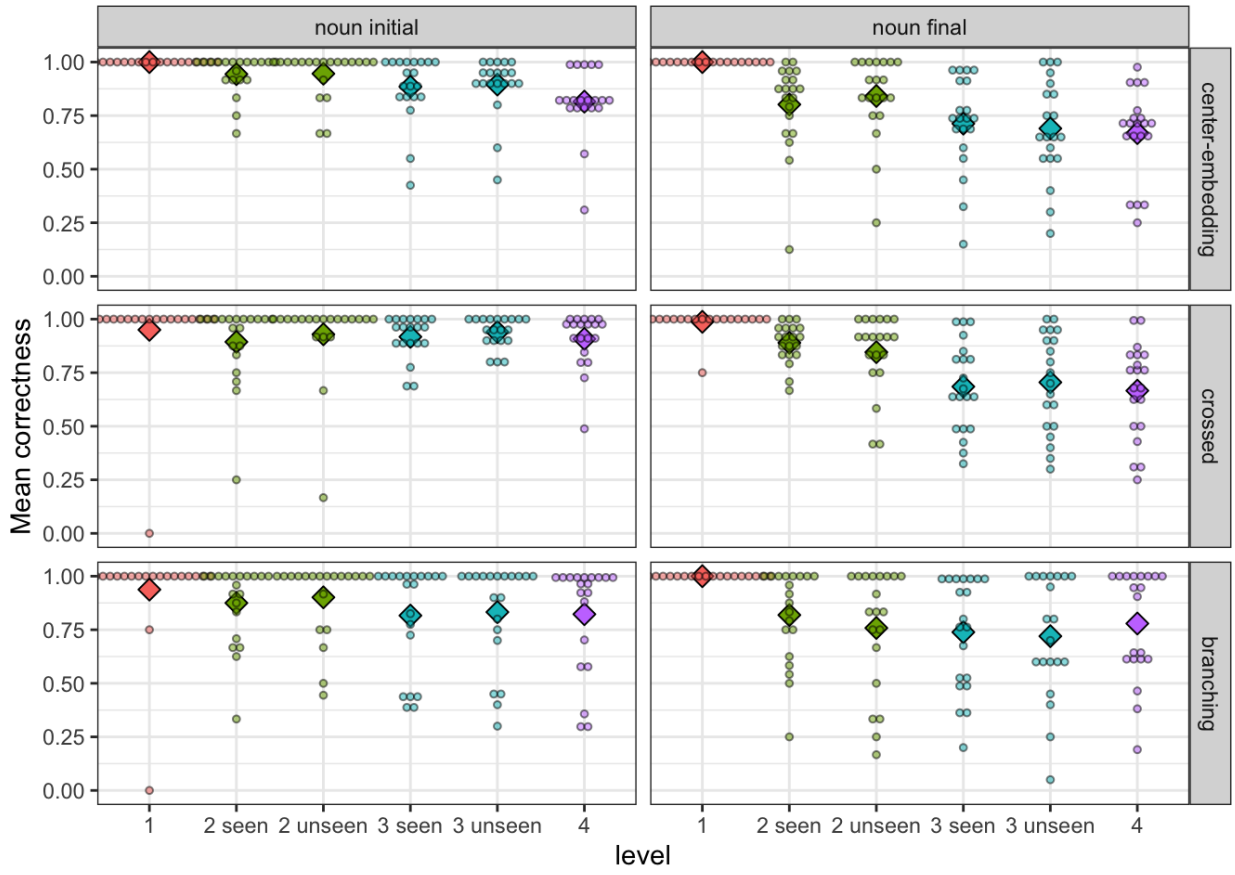


Figure 9: Production correctness (1 minus normalized edit distance from correct) by level and condition. Each point represents one participant's performance at that level; large diamonds represent mean performance across all participants.

In summary, production performance on branching grammars was significantly better overall than for the other two types, which did not differ significantly from each other. The interaction of noun order with level (number of objects) was also significant: i.e. performance declined with level more sharply in the noun-final condition, although noun order did not produce a main effect.

To assess how well the grammar was being generalized to new items, for levels in which both familiar and novel scenes appeared in the production task (two and three items), accuracy was compared across previously seen and unseen stimuli. This model had the same fixed effects as

the previous, plus seen vs. unseen. There was no significant difference in accuracy between the 8 scenes seen in training and the 4 novel ones at each level (B= 0, SE = 0.005, p = .976); there was, however, an interaction between seen/unseen and noun order. Specifically, the difference between seen and unseen was more pronounced in the noun-final condition, with lower performance on unseen test items (B = -0.01, SE = 0.01, p = .03). This shows that in the noun-final conditions, participants found it harder to generalize the learned language to items not seen in training, whereas dependency type did not appear to make a difference. There were no other significant main effects or interactions. We also analyzed performance on 4-object scenes alone to determine if there were any differences in generalization across conditions. There was no significant difference between branching and other dependency types (B = -0.01, SE = .013, p = .2), or between crossed and center-embedded (B = -.007, SE = .023, p = .75), but performance was significantly lower in the noun-final conditions (B = -.07, SE = .023, p < .0001). Overall, as described above, performance was best on the branching dependency type, but participants in this condition did not perform significantly better than those in the other conditions at generalizing to unseen stimuli, including scenes with four objects.

2.2.3. Grammars of participant productions

Figure 10 shows the best match grammar for each string produced by participants, organized by condition. Branching syntax was well-replicated, with most participants who were trained on a branching grammar producing descriptions which were consistent with a branching grammar (although not necessarily preserving order). For many participants, noun order was indeterminate because their adpositions were underspecified (see section 2.1.4), but noun-initial order was better preserved than noun-final. Participants trained on center-embedded syntax did not replicate

this pattern well, instead often producing strings best matched by a crossed grammar (dark green) or indeterminate grammar (light green). Conversely, crossed grammars were better preserved than center-embedding grammars, and were not reanalyzed as center-embedding (as can be seen in the rarity of strings matching this grammar, coded as pink cells, in the figure). Note also that generalization takes place: when a grammar is inferred for the “seen” items, it is in most cases carried over to the “new” items where an entirely new label must be composed.

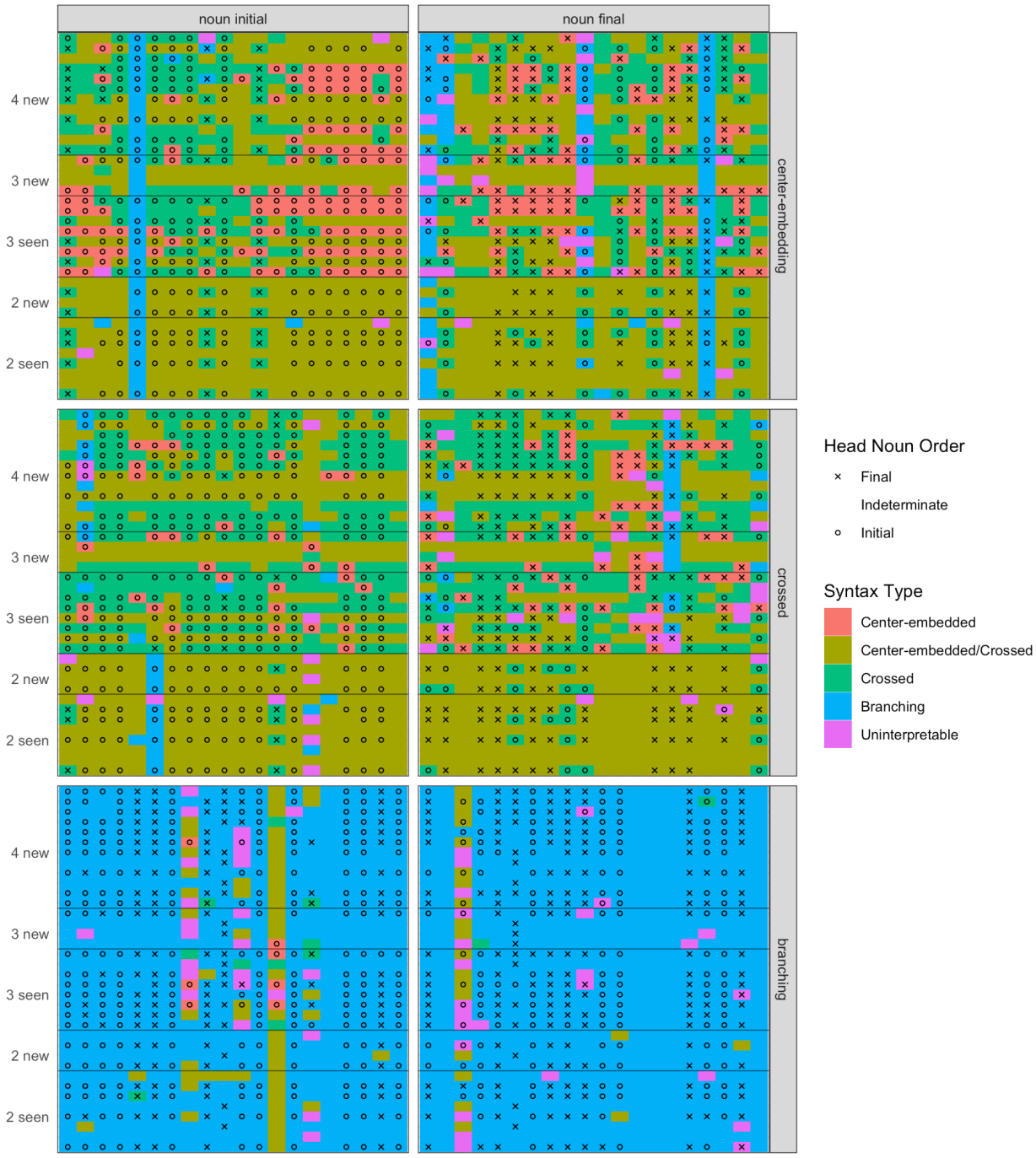


Figure 10: Participant production grammar by string and condition, Experiment 1. Each column within a facet (condition) corresponds to a participant's productions, with each cell being one string. Color represents dependency type, and X/O symbols represent head order, as listed in the legend. The sidebars indicate the number of objects in a scene and whether it was seen in training, or new to the production task.

2.3. Discussion

These results, particularly the lack of a significant difference between accuracy for center-embedded and crossed grammars, and the occasional appearance of a crossed grammar as a best fit in the center-embedded condition (while the converse never happened), are consistent with earlier findings (Bach et al. 1986) that crossed grammars are not more difficult to learn than center-embedding grammars. This is a surprising finding under the assumption that typology and learnability should be aligned; as discussed above (in section 1), crossed dependencies are quite rare in natural languages, whereas the harder center-embedded dependencies are more common. No participants reported speaking Dutch, Swiss German, or Tagalog (which allow crossed dependencies), so natural-language influence was unlikely to have been a factor. On the other hand, as evident in the production tests, the difficulties caused by center-embedding (relative to other grammars) seem less pronounced than those caused by an order where nouns follow adpositional phrases, which would be unfamiliar to and difficult for native English speakers.

It is possible that grammars resembling crossed or center-embedded syntax have arisen from the constraints of the task: participants may have been filling in the names of objects, then adding adpositions to fill out the length of the full required phrase (as was necessary to advance to the next item). This would have yielded either (what looks like) a noun-initial center-embedded grammar (N1 N2 N3 P2 P1), a noun-initial crossed grammar with postpositions (N1 N2 N3 P1 P2), or a noun-final crossed grammar with postpositions (N3 N2 N1 P2 P1), which did occur for some items in some participants' productions. This "reading off" strategy would favor crossed grammar more than center-embedding: the strategy here would be to list objects, then list spatial relations, in the same order (or spatial relations, then objects). To replicate a center-embedded syntax with this strategy, either the list of objects or the list of spatial relations would have to be reversed. The asymmetry between crossed and center-embedded grammars – center-embedded

becomes crossed but not vice-versa – would be consistent with participants using this listing-off strategy, as well as being consistent with a preference for crossed dependencies over center-embedding.

Some uncertainty in the results was introduced by unclear adposition reporting. While the horizontal adposition *moy* was specifically defined as “right of”, or “left of,” for these meanings, the English *next to* was readily available and unspecified results were common. There is no equivalent unspecified vertical term (a few participants did gloss the “atop” adposition *rae* as unspecified vertical). Therefore, it was not always possible to disambiguate between two grammars. In the following experiment, we attempted to clarify how participants understood the meaning of adpositions by refining the final survey, which was also necessary to produce iterable results for the next generation in the chain.

3. Experiment 2: Iterated learning and the emergence of consistent branching direction

This experiment set out to examine the emergence and stabilization of a consistent grammar in nested locative phrases, using the same type of stimuli as in the previous learnability experiment, and furthermore to observe whether the results reflected natural language typology. We hypothesized that branching syntax would be predominant in the results, due to its greater learnability, whereas center-embedding and crossed grammar would be rare, in keeping with previous experimental results.

3.1. Background: Iterated learning and artificial languages

Iterated learning is a process in which participants are asked to perform a cognitive task, such as viewing and then repeating a sequence, and the response of one participant becomes

the stimulus for the next. Each participant's response to the input that he or she is exposed to is referred to as a *generation*, and the "genealogy" of all participant responses in the same line of descent constitutes a *chain*. The first participant is trained on experimenter-designed stimuli, henceforth referred to as *generation 0*.

Through the process of transmission in iterated learning, the initial stimulus is modified due to learners' cognitive and memory biases (along with whatever constraints are imposed by the task itself), which are amplified with each subsequent generation. If the stimulus for the first participant (generation 0) is initially random and structureless, it tends to develop structure through repeated transmission, and irregularities are smoothed out. Artificial languages which initially lack consistent structure develop compositionality and regularity through transmission (Kirby et al. 2008, Beckner et al. 2017, Kirby et al. 2014, Saldana et al 2019). Artificial language learning experiments (e.g. Hudson Kam & Newport 2005, Fedzechina et al. 2018) confirm that learners regularize unpredictable input (but see Perfors 2016 for a counterexample). In addition, iterated learning can drive the regularization of features such as mapping of words to meanings (Reali et al 2009) and marking of plurals (Smith and Wonnacott 2010), increasing regularity by generation.

While we are not aware of any experimental studies looking at iterated learning and dependency ordering, there is at least one computational simulation of a similar process, using artificial neural networks to model language learning (Reali and Christensen, 2009). In their simulation, networks were first pretrained on sequential learning, and then on a simple language based on phrase structure grammars. At the initial "generation," phrase structure rules were flexible, such that phrasal heads and dependents could occur in either order (e.g. possessive before NP or vice-versa). Once trained, networks were tested on different grammars, and the best-learnt grammars were passed on to the next generation of networks, for hundreds of generations. Over time, the prevalence of consistent phrase structure rules with consistent head order across rules increased. As previously discussed, a combination of head-initial and head-

final orders (e.g. head-initial noun phrases plus postpositions) can produce syntactic center-embedding. Therefore, consistent head order across phrase types produces a consistent branching grammar and decreases or rules out center-embedding. We find comparable results appear with human participants in a few generations.

3.2. Methodology

3.2.1. Participants

As in Experiment 1, participants were recruited online through Mechanical Turk and compensated $5 for their time. In total, 192 participants were run, but all but 60 were excluded due to taking written notes, not providing a diverse enough set of labels, and/or not providing clear enough adposition responses (see Procedure below). Data from these 60 participants were analyzed, forming 12 chains of 5 generations each.[8]

3.2.2. Stimuli

Visual stimuli were the same as in Experiment 1. Each chain was assigned a randomly chosen set of 4 nouns from the 10 sets. At generation 0, the language used for scene labels was randomly generated; at each subsequent generation, the previous participant's labels from the production task were used for training, along with given adposition meanings.

Random stimuli for generation 0 were produced by listing the nouns and adpositions for the objects and spatial relations in each scene, and then scrambling the word order. Therefore, each description contained words representing the correct objects and spatial relations, but with

[8] A pilot experiment with a small number of 10-generation chains demonstrated that 5 generations were sufficient for the artificial grammar to stabilize on a consistent syntax.

no consistent syntactic order. For example, a scene with object B atop object A could bear the description (in English) “A B atop,” “atop B A,” “B atop A” and so on.

3.2.3. Procedure

The overall task followed the same structure as that in Experiment 1. At generation 0, four possible combinations of adposition meanings were used for training (left/atop, left/under, right/atop, right/under), with each being used in three chains (this was done to avoid possible confounds from learnability of different adposition meanings).

To clarify the adposition meaning that participants had arrived at, and to determine which meanings should be used to train the next generation, a visual component was added to the final survey, asking participants to identify images corresponding to the meaning of each adposition (see Figure 11). A participant’s adposition meaning was judged as consistent if the following criteria were met: a specific answer (e.g. *left* or *right* as opposed to *horizontal*) was given on the survey question on adposition meaning, and all and only the three images corresponding to that meaning were selected in the visual survey (a column in both of the scenes of images in figure 11, as highlighted). Since adposition meaning was iterated along with labels, it was necessary to have fully consistent adposition meanings to create training trials at the next generation, and prevent adpositions from becoming underspecified (as they frequently were in preliminary pilot experiments as well as experiment 1); a participant’s data was therefore only used for iteration if they produced consistent adposition meanings. Many participants did not fulfill this criterion. Participants were run until all chains were filled to the required five generations.

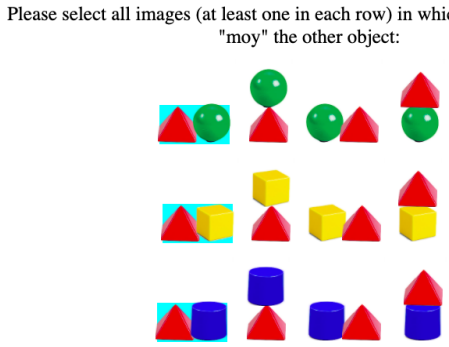


Figure 11: Visual adposition selection question (blue highlighting indicates selection)

At the production phase, as in Experiment 1, each participant saw and was asked to label forty total scenes: 20 that were seen in training with descriptions and 20 that were novel. In the training phase, Participant 1 learned a randomly generated Generation 0 language, which consisted of labels for 20 scenes. In the subsequent production phase, Participant 1 then produced the labels for these 20 familiar scenes, and was also asked to produce labels for 20 novel scenes. These 40 labels constituted the Generation 1 language. The labels for the 20 familiar scenes that Participant 1 produced in the production phase were used as in the input in Participant 2's training phase. In Participant 2's production phase, the task was the same as for Participant 1: provide labels for the 20 familiar scenes from the training phase, and also for the same 20 novel scenes that had been shown to Participant 1. The same process was repeated for subsequent participants. Figure 12 summarizes the process.

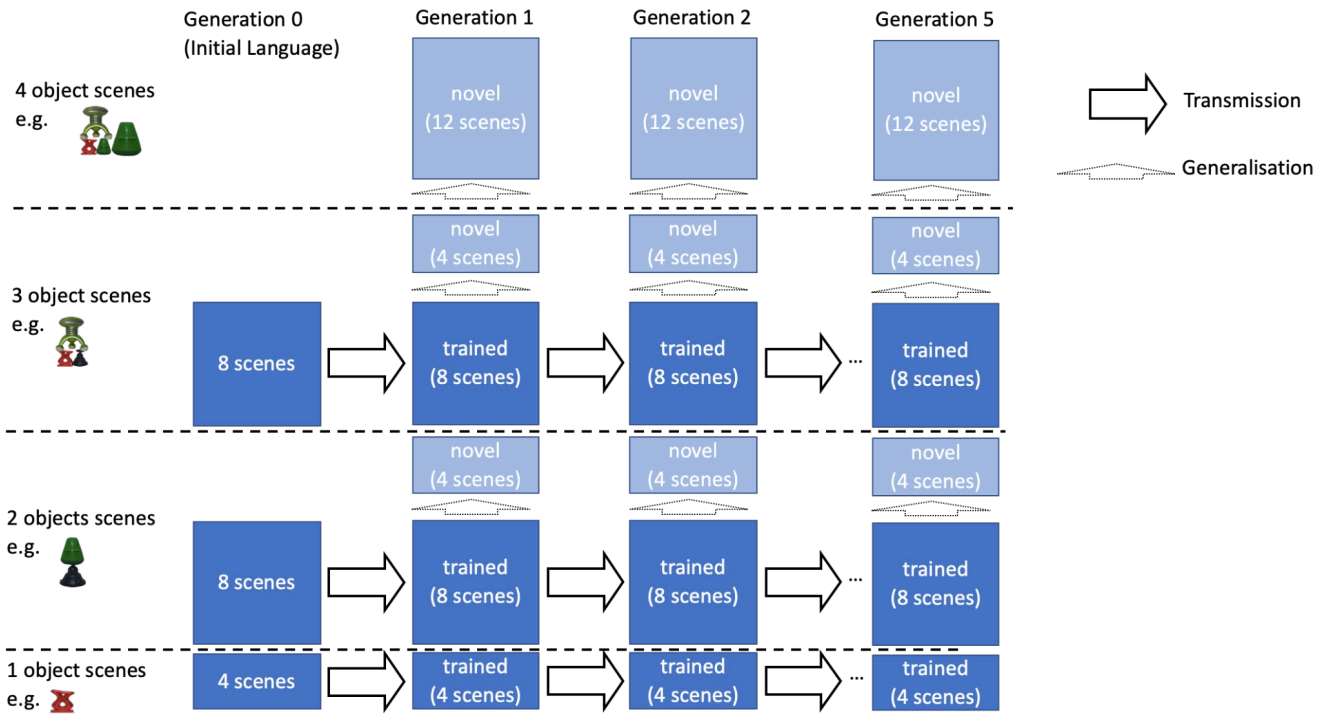


Figure 12: schematic of iteration in Experiment 2. Each participant saw and was asked to label forty total scenes. Participant 1 learned a randomly generated Generation 0 language, which consisted of labels for 20 scenes. In the subsequent production phase, Participant 1 was asked to label those scenes plus 20 novel ones. These 40 labels constituted the Generation 1 language. The labels for the 20 familiar scenes that Participant 1 produced in the production phase were used as in the input in Participant 2's training phase, and so on.

In order to be iterated to the next generation, participant output also had to adhere to criteria of minimum diversity (defined below). This was necessary to prevent the language from becoming degenerate and underspecified, i.e. using the same word to refer to multiple different objects, or only one adposition for both spatial relations (cf. Kirby et al., 2008; Beckner 2017). Criteria for diversity were as follows: Iterable productions required four different labels for the four objects, and at least eight different labels for the twelve scenes (multi-object) at each subsequent level (which included eight seen in training and four new for two-and-three-object scenes, and twelve new scenes with four objects). This would allow enough unique labels for each of the eight scenes per level seen in training at the next generation. Since all participants provided unique labels for at least the eight seen-in-training items during the production task, the items did not

have to be reordered to create a set of eight unique labels; consequently, all participants saw the same eight items in training at each level in each generation. A pilot experiment, detailed in Appendix 1, did not use adposition training but produced generally similar results to the present experiment with regard to dependency type.

3.2.4. Analysis

The same analyses were performed as for Experiment 1 on comprehension and production, but without condition, and including generation as a numeric.

Grammars at each generation were analyzed for best fit with possible grammars in the same way as Experiment 1. For each generation of each chain, the entropy of the grammar distribution – the list of all possible grammars for all strings in a generation – was also measured and analyzed with a linear regression model using generation as a fixed effect and chain as a random effect. This provided a quantitative analysis of how the grammar was regularizing and becoming consistent in terms of dependency type.

3.3. Results

Complete statistical analysis of all measures can be found in Appendix 2, section 4.

3.3.1. Comprehension

Comprehension performance was analyzed using a mixed effects binomial model with level (coded by successive differences contrast) and generation (coded numeric) as fixed effects, and generation/level slope by chain and participant as random effects. Comprehension scores were

well above chance (25%) from the start and did not change with generation (B= -.092, SE = .107, p =.392). Comprehension scores for 3-object scenes were significantly lower than for 2-object scenes (B = -0.747, SE = 0.359, p =.037). There was no significant interaction between level and generation (p = .96).

3.3.2. Production: accuracy and consistency

Unlike comprehension, the production task showed significant generational changes (Figure 13). Descriptions for seen and unseen scenes were analyzed separately here. Production accuracy (i.e. accurately reproducing the training label, which is the label produced at the previous generation) for descriptions seen in training improved with generation (B = 0.03, SE = 0.02, p = .045). As evident in Figure 16, performance for 1-object items was at or near ceiling for most generations and chains. Performance on 2-object items was significantly lower than on 1-object items (B = -0.32, SE = 0.06, p < .001), and in turn, performance on 3-object items lower than for 2-object items (B= -0.11, SE = 0.05, p = .03). No generation-level interaction was observed.

Production accuracy for unseen items also improved with generation (B = 0.12, SE =0.01, p < .001). There were no significant effects of level (number of objects per scene), and no interactions (all p > .29). Since participants had no access to any labels for unseen items created by previous-generation participants, the production of accurate labels is a clear sign of the grammar stabilizing on a generalizable structure, allowing participants at successive generations to produce similar labels even if the later participant did not actually see the label produced by the participant at the earlier generation. Even though comprehension scores did not increase significantly with generation, an improvement in production accuracy and consistency suggests that the languages were becoming more learnable, and more readily generalized to new scenes.

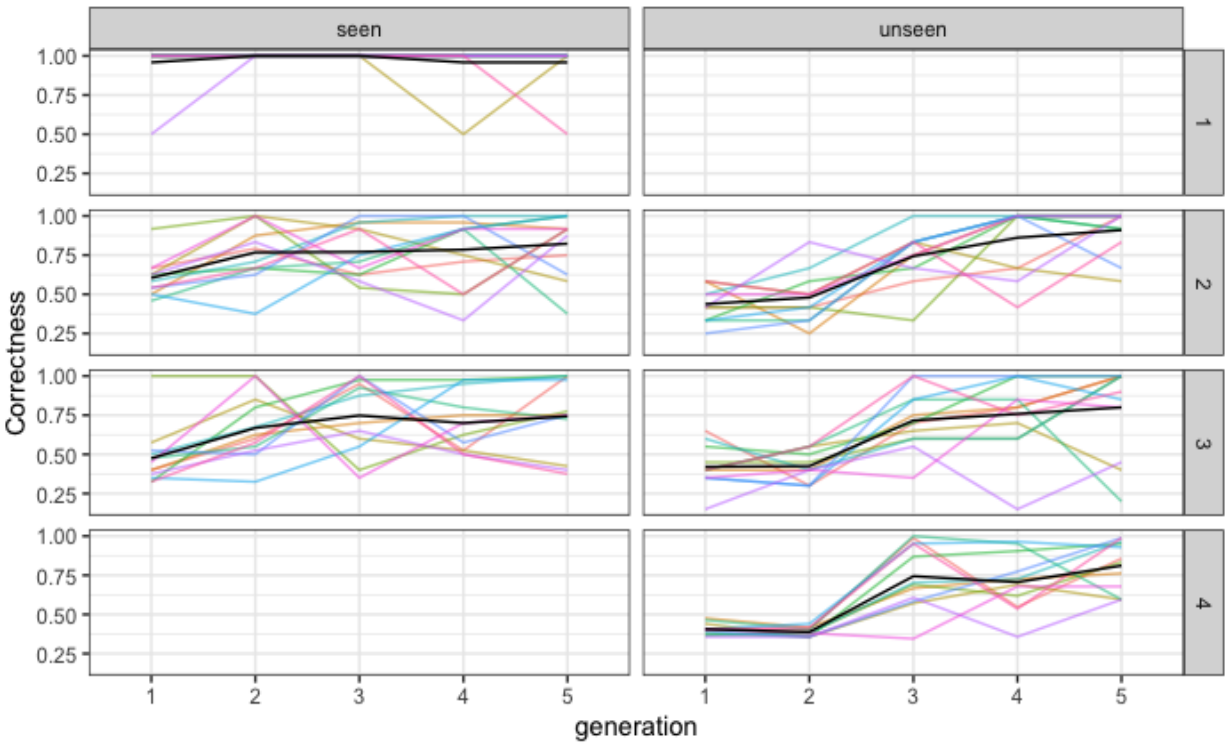


Figure 13: Production accuracy for seen descriptions, and consistency of descriptions for novel scenes. Facets 1-4 indicate the number of objects, colored lines are chains, the black lines represent the mean across all chains.

3.3.3. Grammar of participant productions

Analysis of the best-fit grammars for participant productions shows that a branching, English-like grammar in most chains as irregularities and variations are smoothed out.

Figure 14 shows the best fit grammar for each string, organized by chain, with each column showing a generation. There is a clear takeover of branching dependency structure, usually noun-initial, in most chains. Entropy of the grammar distribution decreased significantly over generations ($B = -0.21$, $SE = 0.03$, $p < .001$; Figure 15), consistent with an increase in uniformity.

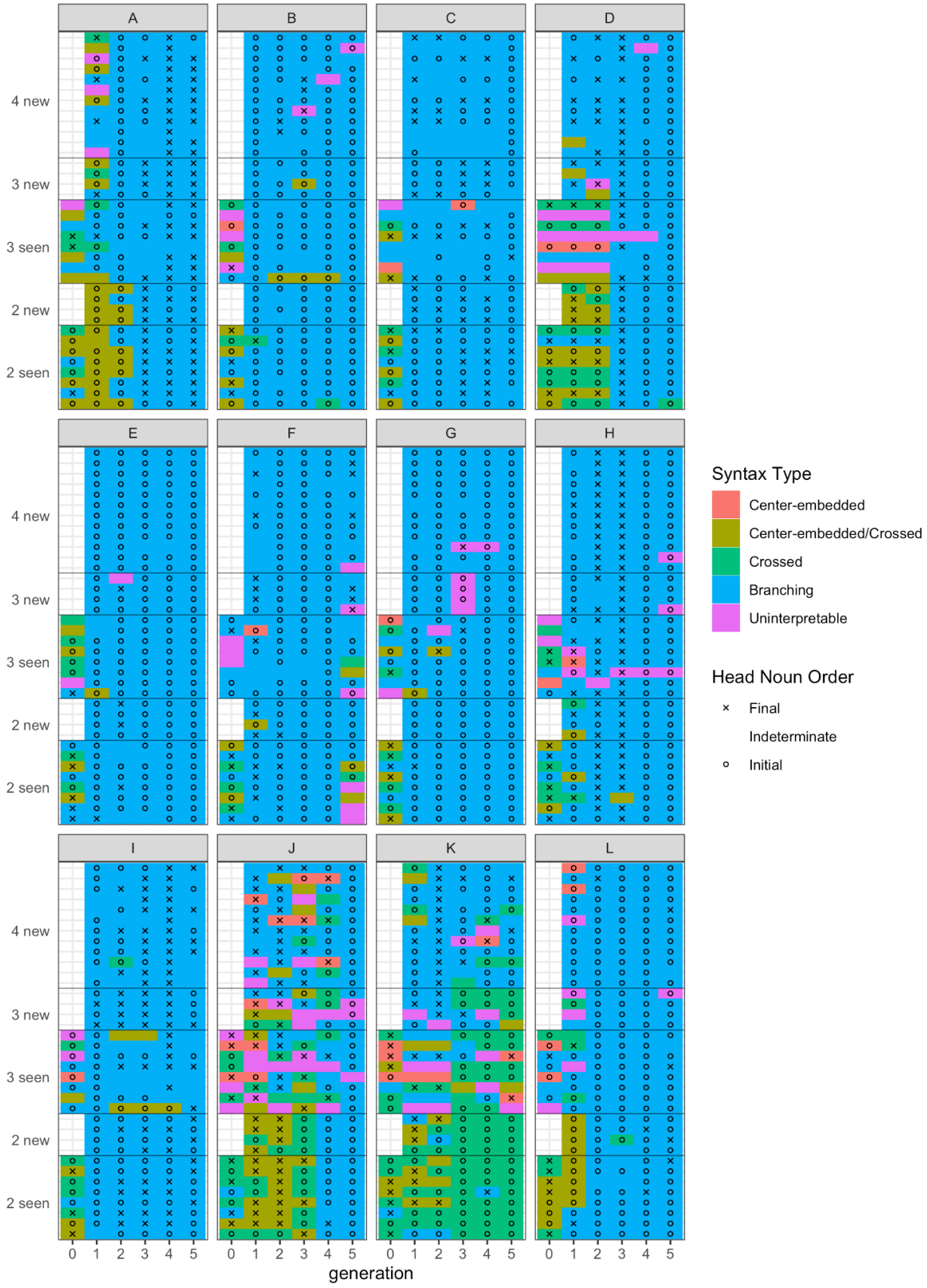

A
B
C
D
E
F
G
H
I
J
K
L
4 new
3 new
3 seen
2 new
2 seen
0
1
2
3
4
5
generation
Syntax Type
Center-embedded
Center-embedded/Crossed
Crossed
Branching
Uninterpretable
Head Noun Order
Final
Indeterminate
Initial

Figure 14: grammar by string, chain, and generation, Experiment 2. Each facet is one chain; generations (participants) are columns. Color represents dependency type and O/X symbols represent head order, as listed in the legend.

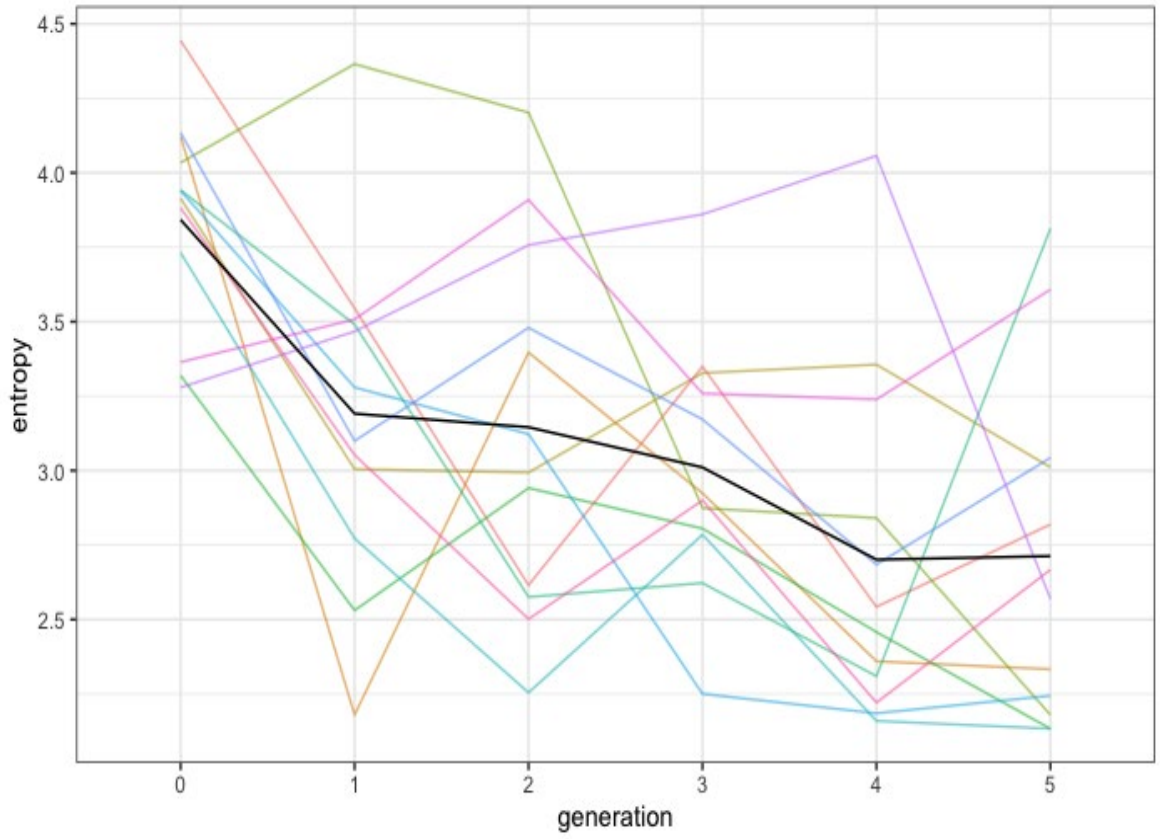


Figure 15: entropy of grammar distribution by generation and chain. Colored lines are chains; black line is the mean of all chains.

3.4. Discussion

The results of this experiment confirm the emergence of branching syntax as preferred over either center-embedded or crossed; branching emerged readily and usually stabilized once established.

Examining Generation 5 languages shows predominant English-like syntax (branching, noun initial). Chain E, some items from which are shown in Figure 16, was typical in this respect.

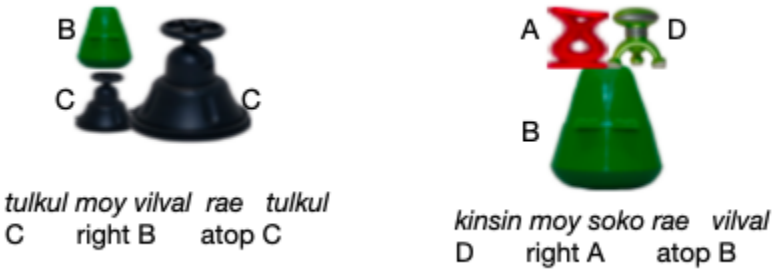


Figure 16: Items and labels from Generation 5, Chain E.

The chain designated A was typical in syntax type, but unusual in that a noun-final grammar predominated in the last three generations. This switchover coincides with, and may have been caused by, a change in the interpretation of the horizontal adposition: from left to right (see section 2.1.4 for how adposition meaning can determine the best fit grammar). One chain, K, was clearly anomalous, developing a crossed grammar for two-item pairs. Figure 20 shows some examples of pairs where the spatial adposition preceded the two object names. The given adposition glosses confirm a crossed grammar where the adposition comes before the head noun of the whole phrase.

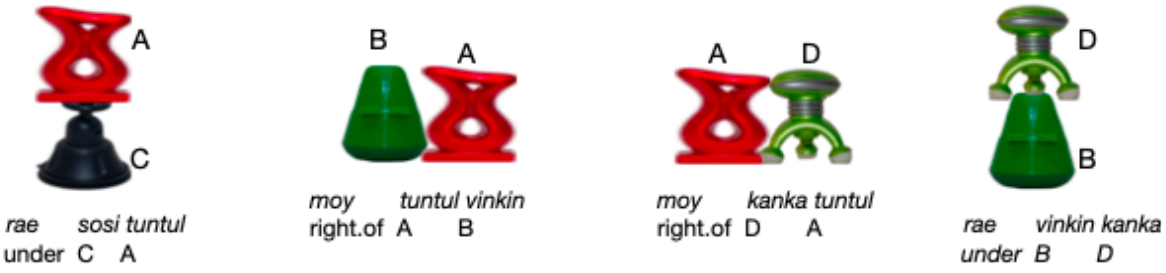


Figure 17: Some two-item scenes from Generation 5, Chain K, showing crossed grammar.

As shown in Figure 18, a unique rule may be operating in Chain K for adjacent pairs of identical objects: for these, the spatial adposition precedes both. Otherwise, the labels for larger scenes appear consistent with branching syntax.

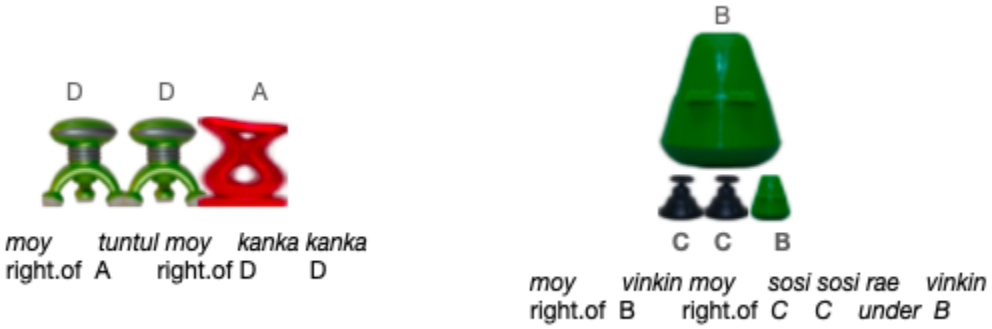


Figure 18: Some larger scenes in Generation 5, Chain K.

Other than this chain, which may not have fully stabilized, there were no clear occurrences of crossed grammar in the stable languages. Nor did center-embedding arise and remain stable. The prevalence of branching grammar may have been due to a combination of structural factors (shorter total dependency length) and participants' linguistic backgrounds (all were English speakers). English influence also likely accounts for the prevalence of noun-initial syntax.

One major difficulty which we faced in both this experiment and Experiment 1 was the difficulty of getting clear adposition meanings, which made it hard to narrow down grammars (see appendix figures). Even with training-to-criterion on adposition labels with geometric objects, many participants had to be discarded because adposition meanings given were nonspecific or inconsistent. Seeing a language with randomized word order, or one in which order had not fully stabilized, may have disrupted participants' learned meaning for the adpositions.

4. General discussion

Taken together, these experiments are consistent with much psycholinguistic research showing that center-embedding is difficult to learn. However, in the learnability experiment, center-embedding was not more difficult to learn than crossed grammar. The difference between previously seen descriptions, and descriptions generated for unseen items, serves as a metric for how well the grammar was generalized. Here, generalization seemed to be more difficult in the noun-final than noun -initial conditions, and dependency type did not make a difference.

The iterated learning experiment further showed that branching syntax tends to arise "naturally" out of initially unstructured input, and the development of a stable syntax was confirmed (in most chains) by the independent creation of identical descriptions for novel scenes by separate participants. The resulting branching grammars feature shorter dependency lengths, which may have been a factor in preference for branching grammars. Another likely factor was native language influence, as all participants spoke English. In all chains except three (A, K, and I), the word order was clearly English-like, and in one of the exceptions, head-final branching order (i.e. backwards English) prevailed. In one chain, K, a crossed dependency grammar may have emerged.

The use of artificial adpositions used in English phrases in the adposition training trials ("the sphere is *moy* the other object") may also have biased participants toward replicating English order. This training was intended to make sure adposition meaning was imparted clearly and consequently to help identify the grammar used by participants, and based on the results of preliminary experiments, it appeared to be necessary because adposition meaning was not

evidently inferred from viewing scenes with descriptions.

Despite these caveats, the results provide additional evidence for the somewhat unexpected observation that crossed dependency structure, rarely seen in natural languages, is not harder to learn than the widely-found center-embedded structure. The emergence of a crossed-dependency grammar in one iterated learning chain is particularly interesting, especially in light of the fact that center-embedding grammar was not observed in any. This not only conflicts with the observed patterns in natural languages, but with the ordering of context-free (center-embedding) and context-sensitive (crossed dependency) languages in the Chomsky hierarchy (Partee et al 1990, Öttl et al 2015). These results are also significant in light of the fact that this experiment featured meaningful semantic strings, demonstrating that the anomaly of crossed and center-embedded dependencies can arise in a meaningful language-like system as well as in nonsemantic symbol sequences (e.g. DeVries et al 2008, Uddén et al 2012). The role of dependency length minimization in learnability is also relevant. A preference for branching dependencies, which shorten dependency length and are widely preferred across natural languages (Hawkins 2004), was strongly confirmed through both experiments. However, an apparent preference for crossed over center-embedded dependencies demonstrates that dependency-length minimization cannot fully account for the patterns observed in our results; while crossed dependencies have shorter maximum length for each individual dependency, they still generate greater dependency length than branching, and a total dependency length comparable to center-embedding. Some other factor(s) must be at work in both natural language and artificial language and sequence learning. For example, crossed dependency structures may be easier than center-embedded structures because they allow a count-and-match strategy rather than last-in-first-out strategy necessary for center-embedding (Vogel et al 1996). This is a plausible strategy for nonlinguistic sequence learning, and our findings suggest it may be at work in artificial language learning as well.

Under the (plausible) assumption that the prevalence of branching dependencies, generally head-initial, was influenced by the shared English language background of our subjects, a useful approach for follow-up studies would be performing comparable studies with speakers of languages such as Japanese and Korean. These languages are head-final and also more tolerant of center-embedding; small-scale corpus studies (Davis forthcoming) suggest that multiple clausal center-embedding is far more common in these languages than in SVO languages such as English.

5. Conclusion

We have shown that center-embedded dependency structures are difficult to learn in artificial languages, and do not readily arise through iterated learning (at least with English-speaking participants). On the other hand, a crossed-dependency grammar was no harder to learn than a center-embedding grammar, and there were suggestions in our data that crossed grammars were more likely to be faithfully reproduced. This is surprising given evidence in other artificial language learning tasks that learnability and typological frequency are often aligned, and considering that crossed dependencies are almost nonexistent in natural languages, whereas center-embedded dependencies do appear widely despite the processing difficulties they may present. Branching syntax was found to be easier than either center-embedded or crossed dependencies, which is consistent with language typology and prior experiments, and branching also predominated in the iterated learning results. Crossed grammar also appeared in the iterated learning experiment, albeit only sporadically and inconsistently in one chain. Why learnability experiments sometimes produce results at odds with language typology remains to be clarified, and our experiment shows

this seeming anomaly applies to artificial languages with semantic meaning as well as to nonlinguistic sequence learning.

Acknowledgements

Many thanks to Robert Kluender for feedback on earlier versions of this article. This research received funding from the European Research Council under the European Union's Horizon 2020 research and innovation program (Grant 681942, held by K. Smith).

Bibliography

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

Bever, T.G (1970). The cognitive basis for linguistic structures. In: J.R. Hayes, Editor, Cognition and the development of language, Wiley, New York (1970), pp. 279–362.

Beckner, C., Pierrehumbert, J. B., & Hay, J. (2017). The emergence of linguistic structure in an online iterated learning task. *Journal of Language Evolution*, *2*(2), 160-176.

Blaubergs, M. S., & Braine, M. D. S. (1974). Short-term memory limitations on decoding self-embedded sentences. Journal of Experimental Psychology, 102, 745–748.

Blumenthal, A. L. 1966 Observations with self-embedded sentences. Psychon. Sci. 6, 453 – 454.

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

Chesi, C., and Moro, A.. (2014). Computational complexity in the brain. *Measuring grammatical complexity*, 264-280.

Chomsky, Noam (1963). "Chapter 12: Formal Properties of Grammars". In Luce, R. Duncan; Bush, Robert R.; Galanter, Eugene (eds.). Handbook of Mathematical Psychology. II. John Wiley and Sons, Inc. pp. 323–418.

Conway, C. M., Ellefson, M. R., & Christiansen, M. H. (2003). When less is less and when less is more: Starting small with staged input. In Proceedings of the Annual Meeting of the Cognitive Science Society (Vol. 25, No. 25).

Corballis, M. C. (2007). Recursion, language, and starlings. Cognitive Science, 31(4), 697-704.

Culbertson, J. (2012). Typological universals as reflections of biased learning: Evidence from artificial language learning. Language and Linguistics Compass, 6(5), 310-329.

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

Culbertson, J., Schouwstra, M., & Kirby, S. (2020). From the world to word order: deriving biases in noun phrase order from statistical properties of the world. *Language*, *96*(3), 696-717.

Dalrymple, M., & King, T. H. (2013). Nested and crossed dependencies and the existence of traces. From quirky case to representing space: papers in honor of Annie Zaenen, 139-152.

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

Dryer, M. S. (1980). The positional tendencies of sentential noun phrases in universal grammar. Canadian Journal of Linguistics/Revue canadienne de linguistique, 25(2), 123-196.

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

Ferrer i Cancho R (2006) Why do syntactic links not cross? Europhys Lett 76(6):1228.

Fitch, W. T., & Friederici, A. D. (2012). Artificial grammar learning meets formal language theory: an overview. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1598), 1933-1955.

Fitch, W. T., & Hauser, M. D. (2004). Computational constraints on syntactic processing in a nonhuman primate. *Science*, *303*(5656), 377-380.

Fodor, J. A. & Garret, M. 1967 Some syntactic determinants of sentential complexity. Percept. Psychophys. 2, 289 – 296. (doi:10.3758/BF03211044)

Foss, D. J., & Cairns, H. S. (1970). Some effects of memory limitations upon sentence comprehension and recall. Journal of Verbal Learning and Verbal Behavior, 9, 541–547.

Futrell, R., Mahowald, K., & Gibson, E. (2015). Large-scale evidence of dependency length minimization in 37 languages. Proceedings of the National Academy of Sciences, 112(33), 10336-10341.

Gazdar, G. (1988). Applicability of indexed grammars to natural languages. In *Natural language parsing and linguistic theories* (pp. 69-94). Springer, Dordrecht.

Gentner, T. Q., Fenn, K. M., Margoliash, D., & Nusbaum, H. C. (2006). Recursive syntactic pattern learning by songbirds. Nature, 440(7088), 1204-1207.

Gold, E. M. (1967). Language identification in the limit. Information and control, 10(5), 447-474.

Gomez-Rodríguez, C., & Ferrer-i-Cancho, R. (2017). Scarcity of crossing dependencies: a direct outcome of a specific constraint? Physical Review E, 96, 062304.

Greenberg, J. (1963). Some universals of grammar with particular reference to the order of meaningful elements. In J. Greenberg (Ed.), Universals of language (pp. 73–113). Cambridge, MA: MIT Press.

Hagège, Claude (1976). "Relative clause, center-embedding and comprehensibility", *Linguistic Inquiry* 7/1: 198–201

Hagège, C. (2010-04-29). Adpositions. : Oxford University Press. Retrieved 17 Jan. 2020, from https://www-oxfordscholarship-com.ezproxy.is.ed.ac.uk/view/10.1093/acprof:oso/9780199575008.001.0001/acprof-9780199575008.

Hawkins, J. A. (1994). *A performance theory of order and constituency* (Vol. 73). Cambridge University Press.

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

Horst, J. S., & Hout, M. C. (2015). The Novel Object and Unusual Name (NOUN) Database: A collection of novel images for use in experimental research. *Behavior Research Methods, 48,* 1393-1409. doi: 10.3758/s13428-015-0647-3.

Hudson, R. (1996). The difficulty of (so-called) self-embedded structures. Work. Pap. Linguist, 8, 283-314.

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

Hunter, T. (2021). The Chomsky Hierarchy. In Allot, N., Lohndal, T., and Rey, G. (eds.) *A Companion to Chomsky*, John Wiley and Sons, 74-95.

Huybregts R. The weak inadequacy of context-free phrase structure grammars. In: de Haan GJ, Trom- melen M, Zonneveld W, editors. Van periferie naar kern. Dordrecht: Foris Publications; 1984. pp. 81–99.

Karlsson, F. (2007). Constraints on multiple center-embedding of clauses. Journal of Linguistics, 43 , 365–392.

Karlsson, F. 2009. Origin and maintenance of clausal embedding complexity. In Sampson, G., Gil, D., & Trudgill, P. (Eds.) Language complexity as an evolving variable. Oxford University Press.

Karlsson, F. (2010). Working memory constraints on multiple center-embedding. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 32, No. 32).

Kluender, R. (1998). On the distinction between strong and weak islands: A processing perspective. Syntax and semantics, 241-280.

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

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

Kuno, Susumu (1974). "The position of relative clauses and conjunctions", Linguistic Inquiry 5/1: 117–36.

Levison, S. (2014). Pragmatics as the Origin of Recursion. In F. Lowenthal and L. Lefebvre (eds.), *Language and Recursion.* New York: Springer Science+Business Media.

Lewis, R.L., & Nakayama, M. (2001). Syntactic and positional similarity effects in the processing of Japanese embeddings. In M. Nakayama (Ed.), Sentence Processing in East Asian Languages (pp. 85–113). Stanford, CA.

Liu, H., Xu, C., & Liang, J. (2017). Dependency distance: A new perspective on syntactic patterns in natural languages. Physics of Life Reviews, 21, 171- 193.

Maclachlan, A., & Rambow, O. (2002, May). Cross-serial dependencies in Tagalog. In *Proceedings of the Sixth International Workshop on Tree Adjoining Grammar and Related Frameworks (TAG+ 6)* (pp. 252-258).

Mazuka, R. and K. Itoh. 1995. Can Japanese Speakers Be Led Down the Garden Path? Japanese Sentence Processing, R. Mazuka and N. Nagai, eds. Hillsdale, N.J.: Lawrence Erlbaum, 295-329.

McCoy, R. T., Culbertson, J., Smolensky, P., & Legendre, G. (2021). Infinite use of finite means? Evaluating the generalization of center embedding learned from an artificial grammar. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 43, No. 43).

Miller, J. E., Miller, J., & Weinert, R. (1998). *Spontaneous spoken language: Syntax and discourse*. Oxford University Press on Demand.

Ota, M., & Skarabela, B. (2016). Reduplicated words are easier to learn. *Language Learning and Development*, *12*(4), 380-397.

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

Perfors, A. (2016). Adult regularization of inconsistent input depends on pragmatic factors. Language Learning and Development, 12(2), 138-155.

Perruchet, P., & Rey, A. (2005). Does the mastery of center-embedded linguistic structures distinguish humans from nonhuman primates?. Psychonomic Bulletin & Review, 12(2), 307-313.

Pullum, G. K., & Gazdar, G. (1982). Natural languages and context-free languages. *Linguistics and Philosophy*, *4*(4), 471-504.

Reali, F., and Christiansen, M. H. (2009). Sequential learning and the interaction between biological and linguistic adaptation in language evolution. Interact. Stud. 10, 5–30. doi: 10.1075/is.10.1.02

Reali, F., & Griffiths, T. L. (2009). The evolution of frequency distributions: Relating regularization to inductive biases through iterated learning. Cognition, 111(3), 317-328.

Rogers, J., & Pullum, G. K. (2011). Aural pattern recognition experiments and the subregular hierarchy. Journal of Logic, Language and Information, 20(3), 329-342.

Sakel, J. and Stapert, E. (2010). Pirahã: In need of recursive syntax? In van der Hulst, H. (ed.), *Recursion in human language*, 3-16. Berlin: Mouton de Gruyter.

Saldana, C., Kirby, S., Truswell, R., & Smith, K. (2019). Compositional hierarchical structure evolves through cultural transmission: an experimental study. *Journal of Language Evolution*, *4*(2), 83-107.

Saldana, C., Oseki, Y., & Culbertson, J. (2021). Cross-linguistic patterns of morpheme order reflect cognitive biases: An experimental study of case and number morphology. *Journal of Memory and Language*, *118*, 104204.

Shieber, S. M. (1985). Evidence against the context-freeness of natural language. In Philosophy, language, and artificial intelligence (pp. 79-89). Springer, Dordrecht.

Smith, K., & Wonnacott, E. (2010). Eliminating unpredictable variation through iterated learning. Cognition, 116(3), 444-449.

Steedman, M. (1984) "On the generality of the Nested Dependency Constraint and the reason for an exception in Dutch." In: B. Butterworth, B. Comrie, and 6. Dahl, eds., *Explanations for Language Universals.* Mouton, New York.

Suh, S. (2000). Multiple subject NPs and processing overload. Language research, 36(2), 279-307.

Suh, S.(2005). The minimal chain principle and parsing Korean.Language Research 41(2), 363-378.

Tucker, Archibald, N. (1940). *The Eastern Sudanic Languages*. Oxford: International African Institute.

Uddén J, Ingvar M, Hagoort P, Petersson KM. Implicit acquisition of grammars with crossed and nested non-adjacent dependencies: Investigating the push-down stack model. Cogn Sci. 2012; 36:1078–101. doi: 10.1111/j.1551-6709.2012.01235.x PMID: 22452530

van der Loo, M. P. J. (2014). The stringdist package for approximate string matching. R Journal 6(1) pp 111-122

Vogel, C., Hahn, U., & Branigan, H. (1996, August). Cross-serial dependencies are not hard to process. In *Proceedings of the 16th conference on Computational linguistics-Volume 1*(pp. 157-162). Association for Computational Linguistics.

Vosse, Theo & Kempen, Gerard (1991). A hybrid model of human sentence processing: parsing right branching, center-embedded and cross-serial dependencies. In: *Proceedings of the Second International Workshop on Parsing Technologies* (Cancun, Mexico, February 1991).

Yngve, V. H. (1960) "A Model and a Hypothesis for Language Structure," Proceedings of American Philosophical Society I04, 444-466.