The Semantics - Syntax Interface: Learning Grammatical Categories and Hierarchical Syntactic Structure through Semantics

Fenna H. Poletiek[1,2] , Padraic Monaghan[3,4] Maartje van de Velde[1], and Bruno R. Bocanegra[5]

[1] Institute of Psychology, Leiden University, Netherlands

[2] Max Planck Institute of Psycholinguistics, Nijmegen, Netherlands

[3] Lancaster University, United Kingdom

[4] University of Amsterdam, Netherlands

[5] Erasmus University Rotterdam, Netherlands

Author Note

Correspondence concerning the article should be addressed to: Fenna Poletiek, Department of Cognitive Psychology, University of Leiden, PO Box 9555, 2300 RB Leiden, The Netherlands. Electronic mail may be sent to poletiek@fsw.leidenuniv.nl

Abstract

Language is infinitely productive because syntax defines dependencies between grammatical categories of words and constituents, so there is interchangeability of these words and constituents within syntactic structures. Previous laboratory-based studies of language learning have shown that complex language structures like hierarchical center embeddings (HCE) are very hard to learn, but these studies tend to simplify the language learning task, omitting semantics and focusing either on learning dependencies between individual words or on acquiring the category membership of those words. We tested whether categories of words and dependencies between these categories and between constituents, could be learned simultaneously in an artificial language with HCE's, when accompanied by scenes illustrating the sentence's intended meaning. Across four experiments, we showed that participants were able to learn the HCE language varying words across categories and category-dependencies, and constituents across constituents-dependencies. They also were able to generalize the learned structure to novel sentences and novel scenes that they had not previously experienced. This simultaneous learning resulting in a productive complex language system, may be a consequence of grounding complex syntax acquisition in semantics.

Keywords: Language learning; Artificial grammar learning; Center embedded hierarchical grammar; Semantics; Syntactic category learning.

The semantics - syntax interface: Learning grammatical categories and hierarchical syntactic structure through semantics.

One of the defining features of human language is its productivity (Pinker & Jackendoff, 2005): From a finite set of words, an infinite set of sentences can be composed. Realizing this productivity requires operations that enable simple grammatical sentences (constituents) to be inserted in another grammatical constituent to form a new hierarchically constructed grammatical sentence and recursivity has been considered to be the property of language providing this expressivity (Chomsky, 1957; Fitch, Hauser & Chomsky, 2005; Hauser, Chomsky, & Fitch, 2002). Whether or not recursivity is observed in all languages is a point of conjecture (Everett, 2005), but most linguists agree that it occurs in nearly all languages, and has been proposed to be a defining feature of human communication and a distinction from other animal communication systems (Corballis, 2007; Fitch & Hauser, 2004, though see Cholewiak, Sousa-Lima, and Cerchio (2013) for discussion of this issue with regard to humpback whale song).

A long-standing linguistic assumption has been that the grammaticality of a sentence is independent of its meaning (Chomsky, 1957). Thus, the hierarchical center-embedded (HCE) structure with two constituents (constructed by inserting a simple sentence inside a sentence) *The dog [the cat chases] runs* and *The cat [the dog chases]runs,* are grammatically identical but have different meanings. Considering the first sentence, its meaning is derived from the meaning of the words (e.g. *cat, dog, chases*), the dependencies between the grammatical categories (e.g. noun *dog* being subject of the verb *runs)*, and the dependencies between the constituents (e.g. noun verb constituent *dog runs* being object of noun verb constituent *cat chases*). Thus, there are two levels of grammatical dependencies in these sentences: the dependency 1) between *runs* (verb) and *dog* (noun), and 2) between *dog runs*

(main constituent of noun-verb pair) that has an object dependency relation to *cat chases* (the subordinate noun-verb pair).

Alternatively, syntax and lexical semantics have recently been proposed to be more integrative, such that particular words define nuanced constraints on permitted combinations on the basis of distributional and semantic information, rather than constraints being determined by rules between linguistic units (Goldberg & Suttle, 2010; Jackendoff, 2010; Reeder, Newport & Aslin, 2013; MacDonald, 2016; Perek & Goldberg, 2015; 2017; Poletiek & Lai, 2012). Yet, there are still undisputedly many broad, abstract constructions in natural languages, such as the HCE examples above, which permit replacing almost any noun and verb in the sequence without affecting the grammaticality of the sentence. However, a key issue in the study of these structures remains; that acceptability is affected by the intended meaning of the constructions.

In order to interpret recursive structures the learner must determine firstly the meaning of individual words, secondly their relation (where appropriate) to referents in the world around them (e.g., Smith & Yu, 2008), and thirdly the mutual relations between higher order units. Consequently, to paraphrase Pinker (1994), a learner can comprehend why the phrase *the dog the man bites* makes the news, whereas *the man the dog bites* does not, if she knows the meaning of the specific nouns and verbs and the object dependency relation between constituents, i.e. of *man* to *dog bites*. The final skill to acquire, crucial for expressive communication, is the productive use of the language. The learner must acquire an understanding that there are categories of words within sentences, which permit replacement of words of the same category, and constituents (word category sequences) that can be replaced with similarly formed constituents. This would enable the learner who already knows that *mouse* and *owl* belong to the same category as *cat* and *dog*, that *observes* and *squeaks* belong to the same category as *runs* and *chases*, and that *the dog runs* stands in an object-

relation to *the cat chases*, and that these object-relations can occur in combination, to interpret *the mouse the owl observes squeaks* even without prior exposure to these combinations of particular words. Hence, recognizing that 1) words in a HCE sentence belong to syntactic categories and 2) that groups of words belong to constituents that depend on their mutual positions in a HCE sentence, is necessary in order to *use* them to productively express or comprehend meaning.

Recursive HCEs are cognitively challenging. Even in adults, accuracy of interpretation of these structures in natural language is effortful and not entirely accurate (Bach, Brown, & Marslen-Wilson, 1986; Blaubergs & Braine, 1974; Foss & Cairns, 1970). Given their substantial difficulty, how are such structures acquired, and what contributes to their learning? There has been substantial work exploring these questions using artificial languages in order to isolate particular aspects of learning. Establishing an artificial language learning paradigm involving these complex structures enables the processes associated with their acquisition for usage, to then be appraised (de Vries, Monaghan, Knecht, & Zwitserlood, 2008).

However, previous artificial language learning studies of HCEs have not yet adequately addressed the productive use of recursive structures expressing dependencies between categories of words and constituents. Hence, thus far, artificial language studies cannot yet inform us about the natural acquisition of these complex sequential structures typical for natural language (Levelt, 2019). For instance, previous studies have isolated only aspects of HCE structures (that append one constituent to the end of another, or insert one constituent within another) using finite state grammars (Fitch et al., 2005). In these studies, sequences either corresponded to a A^nB^n or a $(AB)^n$ structure, applying over two categories of words: A and B, with constituents being grammatical AB-pairs. For A^nB^n sequences, the grammar produces a sequence of As succeeded by a sequence of a matching number of Bs. Such sequences can, but need not (Perruchet & Rey, 2005), be constructed by a HCE

grammar, where one pair of words, e.g., A_2B_2 , comprising a constituent (i.e. sentence in the language) can be inserted into another sentence, e.g., A_1B_1, to make a longer sentence $A_1A_2B_2B_1$. In the case of an $(AB)^n$ structure, a constituent sequence A_2B_2 can be added to the end of another sequence A_1B_1 to make a longer sequence $A_1B_1A_2B_2$.

In experimental studies, such a distinction between an $(AB)^n$ and an A^nB^n sequence is evident to human participants but not to macaque monkeys (Fitch & Hauser, 2004). However, without a salient cue to differentiate the category of A and B words, the distinction was not evident even to humans (Perruchet & Rey, 2005), suggesting that even humans cannot learn HCE without additional cues. Whether or not species other than humans can learn A^nB^n sequences remains a matter of debate (Corballis, 2007; Gentner, Fenn, Margoliash, & Nusbaum, 2006; Petkov & Wilson, 2012; Wilson, Spierings, Ravignani, Mueller, Mintz, Wijnen, Van der Kant, Smith & Rey, 2018). Crucially, such studies highlight *one* aspect of HCEs involving a set of As followed by a set of Bs; they are not able to directly test the *dependencies* between particular As and Bs. They are also unable to inform about how these structures are used for communicating meaning. As such, the typical task in an AGL experiment (evaluating whether a structure is 'grammatical'), can be solved by alternative shallow strategies, such as counting categories of words (the As and then the Bs), rather than learning the dependency structure of the sequence (de Vries et al., 2008).

Understanding the dependencies between particular As and Bs, and the dependencies between AB pairs, is needed to use HCE-grammars to construct the intended meaning of the sentence. An attempt to address the first requirement of HCEs – that there are dependencies between particular As and Bs – has been tested in several artificial language studies (Bahlmann, Shuboltz & Friederici, 2008; Friederici, Bahlmann, Heim, Schubotz & Anwander, 2006). In these studies, sequences were again either of the form A^nB^n or $(AB)^n$, but particular pairs of A and B words always co-occurred together in the sequences. For

instance, whenever the A word *de* occurred, the B word *fo* always appeared in the position corresponding to the dependency between that A and B. Participants were able to learn these sequences, but de Vries et al. (2008) noted that in these previous studies the A and the B category words shared phonological properties, which permitted a simple counting strategy during testing. Without the possibility of applying a counting strategy, de Vries et al. (2008) showed that participants failed to learn the HCE with these materials.

More recent studies have found that learning HCEs *can* occur when additional cues are provided to the learner. Lai and Poletiek (2011) and Poletiek et al. (2019) found that HCEs could be learned if particular AB pairs were first acquired in a starting small training regime, and Mueller, Bahlmann, and Friederici (2010) demonstrated that prosodic cues may help in the acquisition of non-adjacent dependencies in the HCEs. However, even though dependencies were included in these artificial languages, they did not instantiate the HCE dependencies between syntactic *categories* of As and Bs, and AB pairs; rather they implemented dependencies between particular words, which limits the productivity of the learned language.

Learning that dependencies apply between categories of words (and hence apply to any word belonging to that category), and that these dependencies determine the interpretation of the sentence, requires that artificial languages relate to meanings. Otherwise, it would not be possible to determine whether the learner had acquired the dependencies expressed in the grammar, or merely a surface-level heuristic. Consider, for instance, an A^nB^n language where any word from category A and any word from category B can occur in dependency pairings. Again, participants could then be tested on their knowledge that there are an equal number of A and B words (e.g., AAABBB) but it would not be possible to determine if participants had acquired the particular AB dependencies (e.g., $A_1A_2A_3B_3B_2B_1$). By providing referents

alongside the sentences one can distinguish between the effect of pairing A_1 with B_1 from the effect of pairing A_2 with B_1, because A_1B_1 would *mean* something else compared to A_1B_2.

A few previous artificial language learning studies have added a semantic domain to an artificial language (Amato & MacDonald, 2010; Moeser & Bregman, 1972; Moeser & Ohlson, 1974; Morgan & Newport, 1981; Oetll, Dudschig, & Kaup, 2017) in order to test learning of various grammatical structures. In these studies, knowledge of the artificial language is typically tested with a grammaticality judgment task. Adding meaning to an artificial sequential system facilitated the learning of the system: For example, an early artificial grammar study featuring 4 word categories in fixed positions referring to visually presented objects (whose colors and orientation were determined by the words in the sequence) was shown to be learned better when the visual displays closely mirrored the words in the string (Moeser & Bregman, 1972). Fedor, Varga, and Szathmáry (2012) used a complex HCE grammar (A^nB^n) with words taken from the participants' natural language, and particular associations occurring between pairs of specific words, as in Bahlmann et al.'s (2008) study. When the dependencies were supported by words with associated meanings (e.g., the category A word *me* always appeared with the word *you* in the corresponding category B position), participants were able to learn the HCE structure, but when words had unrelated meanings (e.g., A word *me* and B word *lake*) the dependencies were not learned.

As in the early studies, the dependencies specified in Fedor et al. (2012) were between particular words, not categories of words. Moreover, the learning of higher order dependencies between HCE constituents was not investigated. That is, relative positions of the constituents – word pairs – in the sentences did not affect the *meaning* of the sentence. In these respects, the grammars used in AGL studies enriched with semantic features were still importantly distinct from complex natural language structures.

Artificial grammar studies have shown statistical learning of simple linear grammars without semantics (Gomez, 2002; Saffran et al., 1996; Reeder et al., 2013), and also that multiple cues (phonological and prosodic) are useful for learning linguistic regularities (Cassidy & Kelly, 2001; Kelly, 1992; Monaghan, Christiansen, & Chater, 2005; Morgan & Newport, 1981; Naigles, 1990; Lai & Poletiek, 2011). However, it remains unclear on the basis of AGL studies, how natural, complex HCEs are acquired. How are the categories of A and B words in the sequences derived, the dependencies between those categories and between higher order constituents learned, and these sentences understood?

The purpose of our study is to explore, by experimentation, the contribution of meaning in this dual learning process. How do learners acquire a fully productive recursive structure, in which words in categories are interchangeable, affecting the meaning but not the grammaticality of the sentence? It may be that the various possible meanings of a HCE sentence make the structure hard to detect. Alternatively, it may be that grounding formal HCE sequences with multiple meanings (various words appearing within categories) facilitates this learning.

In two sets of experiments, we explore the extent to which flexible language comprehension can be acquired from a A^nB^n HCE artificial language. We test the learning of two relations needed to derive meaning from HCE sentences: (a) the relations between A's and B's, and (b), the relations between constituents AB pairs. In all experiments participants were first exposed to sentences of the artificial language, together with the picture representing its meaning. Next, they were tested on their knowledge of the grammar with a comprehension test. In effect, we are simulating how natural language learners exposed to sentences with multiple clauses like *The dog (A_1) the cat (A_2) chases (B_2) runs (B_1)* extract the subject-verb relations (*dog* to *runs* and *cat* to *chases*) and the hierarchical object relation between *cat chases* and *dog runs*, from exposure to the simultaneous presentation of the

sentence and a visual scene where As are observably related to Bs, and AB pairs to each other.

In Experiments 1a and 1b, we tested whether participants are able to correctly interpret sentences that they have not previously seen, but that contain pairs of particular A and B words that have been experienced during training. Analogous to natural language, learners would be familiarized during training with the noun-verb pairs: *the boy laughs* (A_1B_1), *the girl kisses* (A_2B_2), *the dog likes* (A_3B_3), and *the man eats* (A_4B_4) presented in HCE sentences such as *the boy* (A_1) *the girl* (A_2) *kisses* (B_2) *laughs* (B_1) and *the man* (A_4) *the dog* (A_3) *likes* (B_3), *eats* (B_4). During testing, they would be exposed to sentences containing these familiar A_iB_i events, but in new grammatical combinations: such as *the man* (A_4) *the girl* (A_2) *kisses* (B_2) *eats* (B_4). In our experiments, A-words referred to shapes, B-words to colors, and AB-pairs to objects (colored shapes; e.g. a red square).

In Experiment 1a, the relation between AB-constituents (objects being colored shapes) was specified in the visual scene by their ordering in space: $A_1A_2B_2B_1$ referring to an A_1B_1-object being positioned left of the A_2B_2-object; In Experiment 1b, however, the relations between the constituents had no reference in the visual display of the objects; the objects were randomly positioned. Only the relations between As and Bs were expressed in the visual merge of particular shapes (A-words) and colors (B-words). Hence, only the individual AB pairings could be 'checked' in the visual display of a sentence, not the relations between AB pairs. Experiment 1b was the only experiment in which the higher order semantic reference about the relation between constituents (AB pairs) was absent. Our novel implementation of both types of dependencies between words and constituents, allows us to test the essential role of semantics in learning complex structures akin to natural hierarchical language.

In Experiments 2a and 2b, we investigated whether learning of the grammar extended further than known AB word pairs (objects), testing comprehension with sentences containing

novel AB pairs (and hence novel visual objects) that had not occurred during training. Hence, learners were tested with sentences containing novel AB pairs, though they had been exposed to each of the individual A and B words in other AB pairings. The same analogy to natural language can be made as for the explanation of Experiment 1a and 1b. Participants would be tested on sentences like *the girl* (A) *the boy* (A) *kisses* (B) *eats* (B), but now with *the girl eats* and *the boy kisses* representing new AB-events.

In this manner, we made two changes to the standard artificial language learning procedure that has been used in the literature (e.g., de Vries et al., 2008; Friederici et al., 2006; Lai & Poletiek, 2011; Mueller et al., 2010) to test the learning of HCEs: First, the training sentences were presented along with a picture representing their meaning, and, second, the test task was a comprehension task, rather than a grammaticality judgment task. The comprehension task could not be successfully completed without knowledge of the structure, because the structure determined the unique semantic representation of the word sequence. In contrast to grammaticality judgment tasks, the comprehension task thus reveals how the positional rules in the language are *used* by participants to represent a particular meaning. This usage is the very goal of the natural language learning process (Christiansen & MacDonald, 2009).

In all experiments, comprehension was measured with a picture matching task: participants choose one of two pictures they believe to represent the meaning of the test sentence (see Amato and McDonald, 2010, for a similar approach). Accurate picture matching was taken to indicate that learners had acquired the HCE structure for semantic sentence processing. The semantic referent domain comprised objects (colored shapes) aligned in a row. In the lexicon, each A word represented one of four shapes and each B word represented one of four colors. Grammatical AB pairs (constituents) then determined the color (B) and the

shape (A) of an object in the display (see Figure 1). Sentences in the artificial language could describe 1, 2, or 3 colored shapes (objects) in the reference domain, and in Experiments 1a, 2a, and 2b, the HCE grammatical structure determined the position of object. For example, in the sentence $A_1A_2B_2B_1$, the first object is described by the first A-word (A_1) in the sentence and the final B-word (B_1), the second object by the second A-word A_2 and the first B-word (B_2) (for an example of a longer sequence see Figure 1).

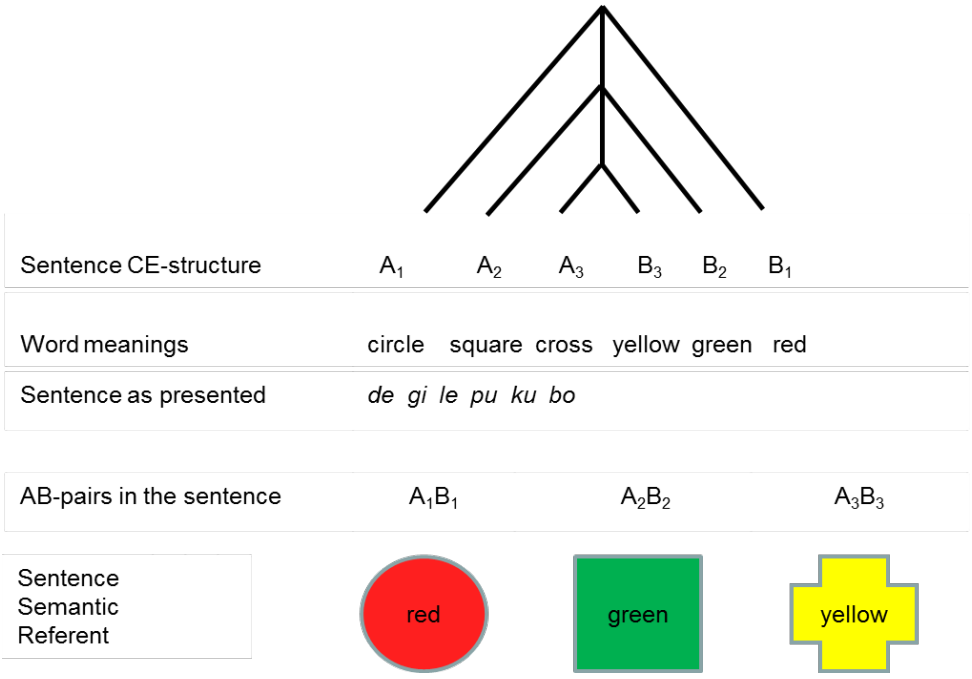


Figure 1: Example of a sentence and semantic referent of the artificial A^nB^n language used in Experiments 1a, 2a and 2b. The sentence *de gi le pu ku bo*, for example, of the form $A_1A_2A_3B_3B_2B_1$, described a row of three objects, positioned from left to right being a red circle, left to a green square, left to a yellow cross. A-words were shape words (e.g. *de* (A_1) is circle), and B-words were color words (e.g., *bo* (B_1) is red). In Experiment 1a, 2a and 2b, the position of an object in the row determined the level of embedding of the corresponding AB pair in the sentence (so A_1B_1, is left to A_2B_2, is left to A_3B_3). In Experiment 1b, the position of an object was unrelated to the position of the AB pair in the sentence. Hence, in Experiment 1b, the example sentence in Figure 1, would represent the three objects in whatever locations in the display.

Experiment 1a

In Experiment 1a we tested whether learners could learn an artificial language with a semantic reference domain made of sequences of colored shapes-objects, referred to by a HCE grammar(A^nB^n). For example, a series of one yellow square positioned to the left of a blue circle, would be described in the artificial language with the sentence: A(*square*) A(*circle*) B(*blue*) B(*yellow*). Thus object *AiBi was positioned left to the object* $A_{(i+1)}B_{(i+1)}$. This semantic representation resulted in the shapes of the referent objects being positioned in the same positions as the A-words in the sentence.

Since the aim of our experiments was to establish learning the HCE *dependencies* (i.e. correct A to B word pairing and AB relations with regard to each other), test sentences would also always correctly describe the positions of the shapes (A-words) while the violations in the test items would always be an incorrect ordering of the colors (B-words), violating HCE dependencies. Keeping the shape ordering constant allowed to avoid confounds in the interpretation of test errors. Indeed, incorrect comprehension of sentences with both A and B words wrongly positioned, might then be caused by either a simple lexical error or a dependencies error. Therefore, both the correct and the incorrect picture would display correctly the positions of the shapes, the color (B) words being only ordered in accordance with the properties of the objects in the *correct* picture. We measured participants' performance with novel sentences made of objects (AB pairs) seen during training. We tested learning using a picture matching task, measuring learners' capability of *comprehending* the test sentence.

Method

Participants. 19 Dutch speaking participants (9 women), between the ages of 17-27, students from Leiden University, participated in this experiment. We based sample size on a previous study of learning of hierarchical center embeddings (Experiment 1 of Lai and Poletiek (2011), effect size was d = 1.125, with observed power = .973 from 14 participants).

In the four studies in this paper, we hypothesized that the effect size would be similar. However, since knowledge of the basic AB structures is a prerequisite for learning of the more complex HCE structure (Lai & Poletiek, 2011; 2013), we analyzed the data with and without excluding participants that failed to learn the basic AB pairs (as indexed by an accuracy at or below chance level on these test items). Taking this criterion into account, we aimed for at least 10 participants in the first Experiment (1a) who passed this criterion of effective learning of the basic AB pairs, resulting in predicted power of .88. Participants were tested in small groups of two to three participants, and data collection was stopped once 10 or more participants, showing successful learning of the basic AB pairs, had been tested. On the basis of participants' average performance on the basic structure (items without embeddings) observed in the first experiment, the number of participants tested in the follow up studies was set at 20 participants.

Materials. The vocabulary for the artificial language comprised four words in each of two grammatical categories. The A category words (referring to shapes) were *de* (circle)*, gi* (square)*, le* (cross*),* and *ri* (triangle), and the B category words (referring to colors) were *bo* (red)*, fo* (blue)*, ku* (green), and *pu* (yellow). The words were derived from Friederici et al. (2006). The words 'de' and 'le' in this set are articles in Dutch and French respectively. The use of these words in our artificial language was unrelated to both their meaning and their syntactic category in these natural languages. Sentences in the language were made of pairs of words taken from the A- and B- categories respectively. The language could produce 16 unique AB pairs, referring to 16 objects (colored shapes). Complex sentences were constructed according to the hierarchical structure A^nB^n, such that AB pairs could intervene between other pairs. Sentences had either 0, 1, or 2 levels of embedding (LoE). Examples of sentences generated by the grammar are *de fo* (0-LoE), *gi [de fo] pu* (1-LoE), *ri [gi [le fo] ku] bo* (2-LoE). Though each color and each shape would be presented during training,

importantly, four arbitrarily chosen objects (shape-color combinations) were not presented during training. They were *de bo, gi fo, le ku, ri pu.* Therefore, the four semantic referents (red circle, blue square, green cross, and yellow triangle) of these AB pairs were not displayed at any point at training, either, in any of the experiments. There were 30 distinct sentences used for training: the 12 unique sentences with 0-LoE (i.e., objects described by AB pairs) left over after omitting the 4 unpresented items, nine unique sequences with 1-LoE (AABB sequences, representing 2 objects), and nine unique sequences with 2-LoE (AAABBB sequences representing three objects).

Sentences were accompanied by pictures of the objects. The dependencies between AB word pairs in the sentences were illustrated by the color(s) and shapes of the objects visually presented. The order of the sequence of shapes corresponded to the order of A words in the sentence. Thus, the first shape was described by the first A word, the second shape by the second A word, and so on. Analogously to natural language, then, sentences such as *the boy the girl kisses laughs* and *the girl the boy kisses laughs* could both be represented as grammatical in the language but with different dependencies between A and B category words, altering the meaning of the sentence. Note that processing the dependencies between A and B category words, and detecting the role of the relative positions of the dependencies, are necessary in order to correctly match sentences and pictures, in our stimuli.

Another 30 sentences were used for testing, eight each with 0-LoE, eleven with 1-LoE, and eleven with 2-LoE. 1 and 2-LoE test sentences were all different from the training sentences and each was unique. Training and test sentences were balanced for the frequency and position of each particular AB pairing. Each test sentence was accompanied by two pictures – one was the target which illustrated the colored shape(s) associated with the sentence, and one was a foil, which did not respect the dependencies between the A and B category words in the sentence.

The 0-LoE test sentences were a subset of items presented during training. Hence, participants had seen each 0-LoE test item during training. Moreover, the test task for 0-LoE's was slightly different from the 1-LoE and 2-LoE items. The task for 0-LoEs was necessarily a lexical test. For test sentences with 0-LoE (representing one object: a colored shape), the foil picture featured the correct shape, i.e., the shape corresponding to the A-word in the sentence, and a color whose name was *not* in the sentence. For example the sentence *de fo* meaning blue circle, would be presented with one picture of a blue circle, and one picture of a yellow circle. Hence, the 0-LoE test items contained a lexical error rather than a syntactic error: the color represented by the B-word in the test sentence was absent in the incorrect picture. For the 0-LoE test items, both the correct picture and the foil picture, had figured in the training trials. As a result, the participant had to select from two familiar objects.

For test sentences with 1-LoE, the sentence comprised a novel sequence of words, but contained only AB pairs that individually had been experienced during training. The target picture corresponded to the sentence, and the foil picture displayed two correct shapes in the correct positions, but with reversed colors. Thus, the correct picture could only be selected based on linking shapes and colors as described by the grammatical dependencies between A and B words. For example, for the sentence *gi de fo pu*, meaning yellow (*pu*) square (*gi*) and blue (*fo*) circle (*de*), the incorrect picture would display a blue square and a yellow circle. Again, the choice was always between two rows of familiar colored shapes.

For test sentences with 2-LoE, the sentence was again novel, and composed of objects (AB pairs) that individually had occurred during training. The target picture corresponded to the sentence. The foil picture presented the correct shapes in the correct positions, but the colors of two of the shapes were swapped. For example, the test sentence *de gi le bo pu fo* was presented with its correct meaning being a row of three colored shapes: a blue-circle, a yellow-square and a red-cross, and with a foil having the colors of two of the shapes swapped

around, for example a blue-circle, a red-square and a yellow-cross. The position of the swap could be the first and second, the first and third, or the second and third. This was to ensure that alternative solution strategies that did not involve computing the HCE dependencies, were not sufficient to solve the task. For instance, if the color of the first shape was always different in the foil picture, then participants could solve the task merely by choosing the picture where the first shape had the color described by the last word in the sequence, i.e. checking two words only. As for the 1-LoE items, the foil pictures were constructed such that they comprised only shape-color combinations that had been experienced during training. See Figure 2 for examples of 2-LoE test items in each experiment.

The shapes (A) and colors (B) were balanced. The shape-color combinations defining the objects were balanced both in terms of their frequency of occurrence in the training set, and of their positions across training and test sets.

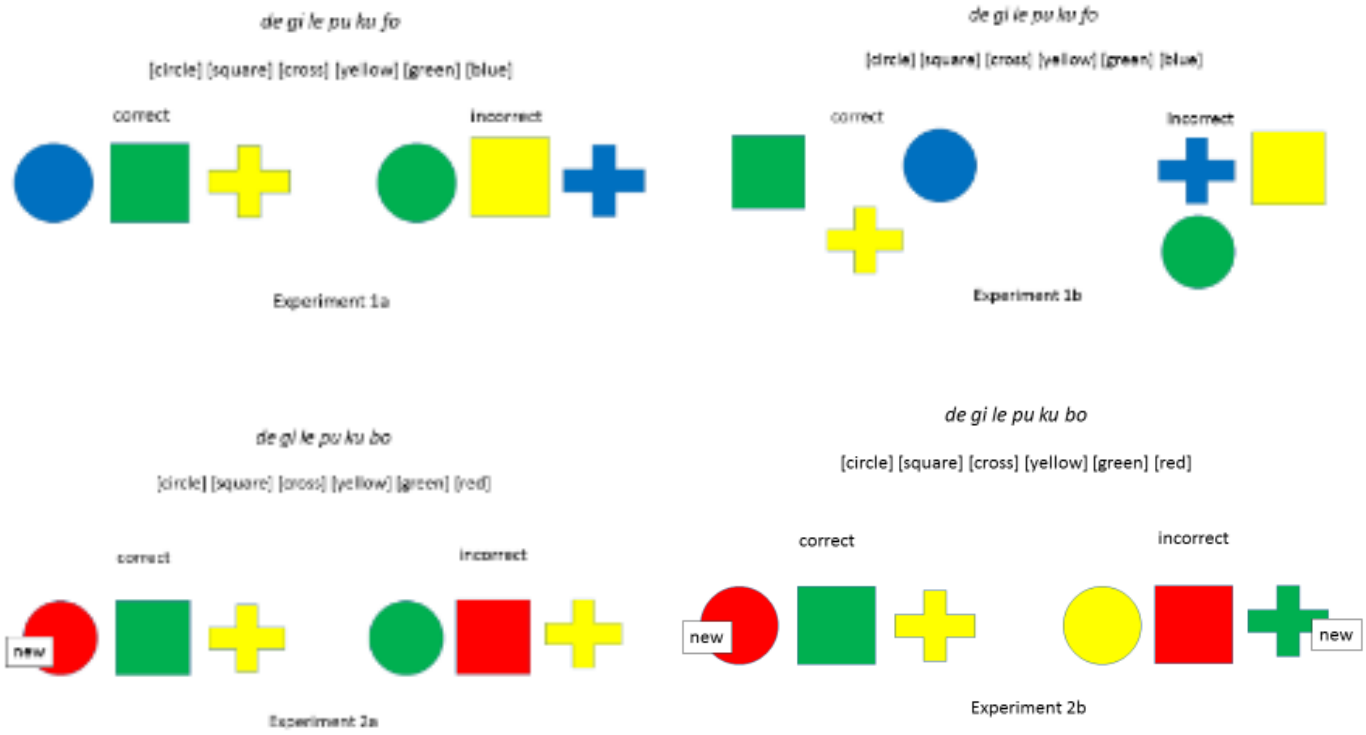


Figure 2: Examples of test items with 2 levels of embedding used in each of the four experiments. In Experiment 1a, 2a and 2b, the language determined both the correct A to B pairings and the relative positions of the objects (AB pairs) in the semantic reference domain. Both the correct and the foil picture proposed in the picture matching task had identical sequences of shapes (A words); the foil had the colors mentioned in the sentence, but incorrectly distributed across the shapes according to the CE-grammar. In Experiment 1b, the positions of the AB pairs in the sentence had no reference to the position of the objects in the reference domain. In Experiment 2a and 2b, the test sentence featured one new AB pair (object) never seen before. In Experiment 2a, the foils featured familiar objects only. In Experiment 2b, both the target and the foils could feature a never seen object.

Procedure. In Experiment 1a, the stimuli were presented on a screen in a PowerPoint presentation to groups of 2 to 4 participants positioned at maximum distance from each other. In the training phase, participants experienced the sentences appearing one at a time in written form on the screen accompanied by their picture referents. Participants were instructed to memorize the items. No reference was made to rules in the instructions. Sentences and pictures appeared on the screen for 2000 ms (0-LoE), 3000 ms (1-LoE) and 4000 ms (2-LoE). After presentation of a sentence, a blank screen appeared during which participants were instructed to rehearse silently what they had seen on the screen. Then, the same sentence and picture referent would appear briefly again, during respectively 1000 ms (0-LoE), 2000 ms (1-LoE) and 3000 ms (2-LoE). This procedure was used to enhance active processing of the training stimuli. The training items were presented in a staged fashion. The 30 training sentences were presented two times each: the 12 0-LoE sentences were presented in a random order first, followed by the same twelve items again randomized, then the nine 1-LoE sentences, and then the nine sentences with 2-LoE would be presented according to the same procedure.

During the testing phase, participants were presented with each test sentence together with its target and foil pictures. Unlike the training procedure, the test items were presented in fully random order. The pictures were presented next to each other, and the sentence was presented immediately above the pictures. Position of the target picture (left or right) were

randomized. Participants were instructed to indicate which of the two pictures matched the sentence. Participants recorded their answers on a sheet of paper. Each test item was presented once, and participants were not able to see one another's response sheets during testing by sitting participants distantly from one another in the room.

There were three different versions of the training and testing power-point slides with different random orderings of the items. No feedback on responses was given.

Results and Discussion

In order to (a) control for potential dependencies in our repeated measures for participants at the level of individual test items and (b) to exclude the possibility of observing spurious effects due to potential nonlinearities in our dependent performance measure (Jaeger, 2008), we fit a logit mixed model (Generalized Linear Mixed Model for binomial outcomes) with Laplace approximation using the *glmer()* function in the R package *lme4* (Bates, Maechler, Bolker & Walker, 2015), with picture matching accuracy for each test item (0 = incorrect, 1 = correct) as the outcome variable for all nineteen participants (very similar results were obtained using conventional ANOVAs; see Supplementary Analyses). We included random intercepts for participants and items, and random slopes for LoE by participants. We included a mean-centered fixed effect for LoE in order to be able to interpret the model intercept. The logit model ($N = 570$; log-likelihood = -332.8) showed that the intercept was significantly larger than 0 ($\beta_0 = 0.80$, $SE = 0.31$, $Z = 2.59$, $p < .01$), indicating that, on average, participants performed better than chance. Note that the model is a regression on log-odds (logits) where $\text{logit}(p) = \log(p / p - 1)$. Here, chance performance has a log-odds = 0. Negative log-odds values indicate below chance performance, whereas positive values indicate above chance performance. The exponentiated and transformed log-odds intercept indicated an estimated mean proportion of correct selections of .69, which was

significantly larger than .50. LoE was not significant (β_{LoE} = -0.05, SE = 0.14, Z = -0.31, p = .75), indicating that the mean proportion of correct selections did not differ for different number of embeddings (see Figure 3). When we excluded the nine participants that failed to perform above chance level on the 0-LoE items (i.e., with accuracy ≤ .50), we observed the same pattern of results (see Supplementary Analyses).

The results suggest that when the training input of a HCE-language is accompanied by a visual scene, learners can acquire the HCE-structure, as indicated in a comprehension task showing participants' ability to use their knowledge of word meanings and the grammatical dependencies between word categories (positions of shape A- and color B-words that specify an object) and clauses (relative positions of AB-pairs that specify the relative positions of the objects) induced during training. This knowledge was used to comprehend the meaning of sentences to which they had not previously been exposed, indicating that the learning was not merely at the lexical word level, but required understanding the relations between word *categories*. Participants did not learn, for example, any relation between particular words and their absolute positions of the sequence (like a square can only occur in first position; or only after a circle). The knowledge acquired was generalized concerning the *relative* positions of shape-*category* words with respect to its dependency to color-*category* words, and the relative positions of the shape-color (AB) pairs. This crucial result contrasts with previous studies showing poor or no learning of a very similar artificial HCE structure after exposure to many more stimuli without meaning (de Vries, et al., 2008), where participants were tested on grammaticality judgments without semantic referents to the sentences available.

Though the task could not be performed by merely matching the shape words to the positional order of the pictures, because both the correct and the incorrect test pictures contained correctly positioned shapes, the hierarchical relation between constituents (AB pairs) simply mirrored in the spatial alignment of the objects, might have simplified the task

overall. Notice however, that analogically, semantic referencing in natural language with a visual scene displaying *who* is doing *what* (A to B pairings) and *to whom* (as in Object Relative clauses determining the relation between AB-units) can be an extremely effective though simple semantic cue for parsing a complex sentence. The artificial language studies by Reeder et al. (2013) and Amato & MacDonald (2013) also suggested a general usefulness of visual cues for grammatical parsing.

To control for the possibility that the straightforward semantic reference of the hierarchical rule might have driven performance during the test, we conducted a control Experiment 1b that removed any cue about the mutual spatial relation between the objects (AB constituents) in the visual display, while keeping the CE binding rule (A- to B- pairings) constant. In other words, the relations between the constituents was semantically unconstrained in Experiment 1b, implying that any A_iB_i object could be at any position in the display of objects mentioned in the sentence. All other conditions were kept identical to Experiment 1a. In Experiment 1b, for example, a set of a yellow square and a blue circle could be described grammatically with either the sentence A(*square*) A(*circle*) B(*blue*) B(*yellow*) or with A(*circle*) A(*square*) B(*yellow*) B(*blue*), both sentences conforming to the HCE structure that now determines only the color of each shape in the sentence.

If the HCE with semantics can still be learned without the visual cue for the constituents dependencies, we expect above chance performance on the comprehension task. However, the random positions of the objects in the reference domain, might make semantic parsing more difficult overall, especially for longer sentences describing multiple objects. Indeed, for these sentences the location of a shape in the sentence cannot be predicted, but has to be searched for in the set of shapes. In sum, in Experiment 1b, participants could not use the positions of the shapes in the pictures anymore to find the correct match to the sentence,

but they could observe the actual colors of the shapes in the scene (A to B pairings) to determine whether they are described by the CE sentence.

Experiment 1b

Method

Participants. 20 students (13 women) from Lancaster University, between the ages of 18-33, participated in this experiment. Participants were tested individually, and paid £3.50 or given course credit for taking part. Participants were native or proficient in English.

Materials. The same artificial grammar was used as in Experiment 1a, with four A category and four B category words, each referring to –respectively– a shape and a color. Training sentences (with 0 to 2-LoE) were exactly the same as in Experiment 1a. As in Experiment 1a, during training, sentences were shown together with their referent pictures comprising colored shapes. However, whereas in Experiment 1a the order of the shapes on screen corresponded to the order of A-words, in Experiment 1b each colored shape was placed randomly in one of five positions on the screen (top center, center, bottom center, center left and center right; forming a cross). Hence, for the sentence *gi de fo pu*, meaning yellow (*gi*) square (*pu*) and blue (*de*) circle (*fo*), the yellow square and blue circle would not be depicted as a sequence, but the yellow square and the blue circle could each appear in any of the 5 positions. This way, shape or color sequencing could not form the basis of matching to the sentence. Conversely, sentences describing the same colored shape combination, but with a different nesting structure (e.g., *gi pu de fo* and *pu gi fo de*), could be depicted in the same manner.

In the test phase, the same 30 sentences were used as in Experiment 1a, including eight sentences with 0-LoE, eleven with 1-LoE, and eleven with 2-LoE. Again, each test sentence was accompanied by a target and a foil picture, in which shape-color combinations

were identical to those used in Experiment 1a. Hence, in 0-LoE sentences, foils featured a lexical error, while in 1 and 2-LoE sentences foils contained a dependency error (i.e., color words were reversed, such that AB pairs were swapped).

However, in both target and foil pictures, colored shapes were placed randomly on one of five positions, either on the right or the left side of the screen. Hence, participants had to choose between two picture configurations (on the left or the right side of the screen), each consisting of one (for 0-LoE sentences) or more colored shapes. As in Experiment 1a, both target and foil pictures comprised only shape-color combinations that had been experienced during training.

Procedure. The task and the trial-structure were the same as in Experiment 1a, except that now the experiment was run on a computer with the experimentation software E-prime (Psychology Software Tools, 2012), with different randomizations of training and test items for each participant. Participants were tested in individual booths.

Results and Discussion

A logit mixed model on all participants ($n = 20$) with random intercepts for participants and items, random slopes for LoE by participants, and a mean-centered fixed effect for LoE ($N = 600$; log-likelihood = -355.4) showed that the intercept was significantly larger than 0 ($\beta_0 = 0.79$, $SE = 0.21$, $Z = 3.78$, $p < .001$), indicating that, on average, participants performed better than chance with an estimated mean proportion of correct selections of .69. LoE was significant ($\beta_{\text{LoE}} = -0.69$, $SE = 0.19$, $Z = -3.62$, $p < .001$), indicating that the mean proportion of correct selections decreased by approximately .14 for each additional level of embedding (see Figure 3). When we excluded the five participants that failed to perform above chance level on the 0-LoE items (i.e., with accuracy ≤ .50), we observed the same pattern of results (see Supplementary Analyses).

Aggregating over Exp 1a and 1b (including all participants $n = 39$), a logit mixed model with random intercepts for participants and items, random slopes for LoE by participants and Experiment by items, and mean-centered fixed effects for LoE and Experiment ($N = 1170$; log-likelihood = -692.1) showed the same pattern of results as the previous analyses: the intercept was significantly larger than 0 ($\beta_0 = 0.79$, $SE = 0.18$, $Z = 4.27$, $p < .001$), indicating that, on average, participants performed better than chance. Both the main effect of LoE ($\beta_{LoE} = -0.37$, $SE = 0.11$, $Z = -3.26$, $p < .01$), and the LoE*Experiment interaction effect were significant ($\beta_{LoE*Exp} = -0.61$, $SE = 0.21$, $Z = -2.88$, $p < .01$). However, the main effect of Experiment was not significant ($\beta_{Exp} = 0.03$, $SE = 0.36$, $Z = 0.07$, $p = .94$). A between-subjects Bayesian t-test for the factor Experiment showed a JSZ Bayes factor BF = 3.00 (Rouder et al., 2009), which indicates anecdotal evidence in favor of the null-hypothesis. This shows that overall performance was comparable in Exp 1a and 1b, but the decrease in performance over LoE was larger in Exp 1b compared to Exp 1a (see Figure 3).

In Experiment 1a, an artificial HCE structure with semantics could clearly be learned in the presence of two semantic referencing rules expressing how 1) the A's are related to the B's and 2) the AB's to each other. The knowledge participants acquired concerned the relation between A and B word categories, as well as between AB pairs. No overall difference in learning between Experiment 1a and 1b was shown. However, as predicted, the absence of a semantic representation of the hierarchical dependencies rule between AB constituents in Experiment 1b made it more difficult to understand sentences as their complexity (the number AB pairs) increased.

Even though the test sentences included new hierarchical orderings of AB pairs, all AB pairs (objects) displayed in the test items in Experiment 1a and 1b, were already familiar to the learner This raises the following question: Can and do learners trained on a subset of all possible instantiations of category dependencies, parse new dependencies between words,

that refer to meanings not previously encountered during training? A crucial question is whether and how learners can acquire a productive language system that both generalizes to new organizations of familiar meaning (i.e., known AB-objects), *and* creates new meaning (i.e., represent novel AB-objects). Experiments 2a and 2b investigate how language learners learn to apply grammatical dependencies to word categories, and then describe new semantic content that has never been experienced or talked about before.

In contrast to Experiments 1a and 1b, Experiments 2a and 2b used test sentences that referred to color – shape combinations that had not been seen before. If the grammar is acquired as a generalizable, productive system then sentence comprehension for sentences with new colored shapes (new AB pairings) should be similar to performance in Experiment 1a and 1b, where all AB pairings occurring in the test, had been seen during training. Indeed, this would suggest grammar learning at the word category level, and the constituents level, independent of the meaning of the words (Onnis, Monaghan, Christiansen & Chater, 2004).

Since comprehending a new meaning (as we test in Experiment 2a and 2b) cannot rely merely on memory, it requires a parse of the sentence structure, to build its meaning. If, however, learners have acquired only a system of grammar that retrieves items from a finite memory, then we should see poorer comprehension than we observed in Experiments 1a and 1b. In Experiment 2a, participants had to choose between the correct referent picture that contained a new object, and an incorrect picture that contained only familiar objects. Any preference for the incorrect picture might indicate that learning had been experience- rather than structure-based. Any preference for the correct picture might indicate that learning had abstracted away from the specific semantic content of the objects in the sentence. Note, however, that the novelty of the object in the correct picture, might also unintendedly bias participants responses. To control for that possibility, we carried out Experiment 2b.

Experiment 2a

Crucially, Experiment 2a aimed to establish whether learners infer knowledge about grammatical dependencies between *word categories* or between *words*. If participants learn grammatical dependencies between categories of words, they should select the picture containing a new object, that was correctly described by the target sentence, rather than a picture containing familiar objects that was not correctly described by the sentence (see Figure 2).

Method

Participants. 20 students (aged 17-27, 16 female) of Leiden University, participated in this experiment. They earned 3 euro or course credit. The participants had not taken part in Experiments 1a or 1b. Sample size and stopping rule were determined as for Experiments 1a and 1b.

Materials. The same training sentences were used as in Experiment 1a, with the same four AB pairs reserved from the training sentences. Also, 30 test sentences were used. The test sentences, however, differed from those used in Experiments 1a and 1b. For the 0-LoE test sentences, the four AB pairs reserved from training were used. The target picture accompanying test sentences was therefore an object (colored shape) that had not been seen during training. The foil picture had the same shape but a different color to the target. As in the previous experiments, the test task for the 0-LoE items was necessarily a lexical selection task. HCE was tested with the embedded test sentences. 14 1-LoE sentences were used. For the 1-LoE test sentences, either the first or the second AB pair was one of the pairs reserved from training. As in Experiment 1a and 1b, the foil picture presented the same shapes as the target picture, but with the colors swapped between the shapes. Hence, the target picture presented a novel colored shape, but the foil picture only featured previously seen colored shapes. So, the foil picture contained familiar components only, but it did not represent the

meaning of the sentence. For the 12 2-LoE test sentences, again one of the three AB pairs in each sentence was one of the pairs reserved from the training sentences, either in the first, second, or third position. The foil picture presented the same shapes as the target picture in the same positions, but with two of the colors of the pictures swapped, either between the first and second, the first and third, or the second and third shapes.

Procedure. The task and the trial-structure were the same as in Experiment 1a, except that now the experiment was run on a computer with the experimentation software E-prime, with different randomizations of training and test items for each participant. Participants were tested in individual booths. In contrast to Experiments 1a and 1b, the test stimuli comprised novel AB pairs, referring to novel objects.

Results and Discussion

A logit mixed model on all participants (n = 20) with random intercepts for participants and items, random slopes for LoE by participants, and a mean-centered fixed effect for LoE (N = 600; log-likelihood = -267.7) showed a significant intercept (β_0 = 1.90, SE = 0.43, Z = 4.43, $p < .001$), indicating that, on average, participants performed better than chance with an estimated mean proportion of correct selections of .87. However, the effect of LoE was not significant (β_{LoE} = -0.27, SE = 0.30, Z = -0.90, p = .37), indicating that the mean proportion of correct selections did not differ for different number of embeddings (see Figure 3). When excluding one participant that failed to perform above chance level (i.e., with accuracy ≤ .50), on the 0-LoE items, we observed the same pattern of results (see Supplementary Analyses).

The test sentences could thus be parsed effectively when their precise meanings had never been seen before. In particular, 0-LoE test sentences describing one new object (colored shape combination) were comprehended almost perfectly, indicating that participants learned the HCE structure and applied it to categories of words in order to productively interpret

sentences with novel objects. However, given that correct parsing always required them to select the sentences containing a new object, participants might have learned over time this contingency between grammaticality and novelty. This was a consequence of the purpose of Experiment 2a to separate memory based comprehension from building a parse, and hence to establish participants' understanding that the system is productive. To control whether novelty per se has biased the respondents choice independently of parsing, we carried out Experiment 2b.

In Experiment 2b we tested whether HCE parsing could also occur independently of semantic novelty. During test, both the correct and the incorrect picture could contain a previously unseen object. If participants are still able to parse the HCE structure correctly, performance should be above chance. On the other hand, if performance is driven by the mere presence of novel semantic content, then performance in Experiment 2b should drop as compared to Experiment 2a.

Experiment 2b

Method

Participants. 20 new students (13 women, aged 18-23) from Lancaster University, participated in this experiment for £3.50 or course credit.

Materials. Materials were as in Experiment 2a, except that now both the target and the foil pictures for 1- and 2-LoE test items contained novel objects. In order that both pictures contained the same color and shape terms, this required the 1-LoE test items to contain a repetition of either shape or color. For example, for the sentence *de de pu bo* meaning red circle yellow circle, the foil picture depicted yellow circle red circle. In both cases, *de bo*, a red circle, was a novel object to the participants. Target 2-LoE test pictures were identical to the 2-LoE test items in Experiment 2a. Regarding the foils: for half the trials we could create

pictures containing a novel object by swapping color-shape pairs from the correct items. In the other half of the trials, foils did not contain a novel object. Overall, then, participants in Experiment 2b could not rely solely on identifying which of the pictures featured a new object (color-shape) combination.

Procedure. The procedure was identical to that of Experiment 2a but used the set of test items described in the Materials that contained novel AB pairs in targets and foils.

Results and Discussion

A logit mixed model on all participants (n = 20) with random intercepts for participants and items, random slopes for LoE by participants, and a mean-centered fixed effect for LoE (N = 600; log-likelihood = -293.7) showed a significant intercept (β_0 = 1.51, SE = 0.46, Z = 3.30, p < .001), indicating that, on average, participants performed better than chance with an estimated mean proportion of correct selections of .82. However, the effect of LoE was not significant (β_{LoE} = -0.24, SE = 0.24, Z = -0.98, p = .33), indicating that the mean proportion of correct selections did not differ for different number of embeddings (see Figure 3). When we excluded the five participants that failed to perform above chance level on the 0-LoE items (i.e., with accuracy ≤ .50), we observed the same pattern of results (see Supplementary Analyses).

Aggregating over Experiments 2a and 2b (including all participants n = 40), a logit mixed model with random intercepts for participants and items, random slopes for LoE by participants and Experiment by items, and mean-centered fixed effects for LoE and Experiment (N = 1200; log-likelihood = -560.9) showed the same pattern of results as the previous analyses: the intercept was significantly larger than 0 (β_0 = 1.70, SE = 0.32, Z = 5.39, p < .001), indicating that, on average, participants performed better than chance. Neither the main effect of LoE (β_{LoE} = -0.26, SE = 0.20, Z = -1.27, p = .20), nor the LoE * Experiment interaction effect ($\beta_{\text{LoE*Exp}}$ = 0.17, SE = 0.27, Z = 0.62, p = .54), nor the main effect of

Experiment was significant (β_{Exp} = -0.46, *SE* = 0.61, *Z* = -0.76, *p* = .45). A between-subjects Bayesian t-test for the factor Experiment showed a JSZ Bayes factor BF = 1.83, which indicates anecdotal evidence in favor of the null-hypothesis (see Figure 3).

Experiment 2b suggests that it was not the contingency between grammaticality and semantic novelty that drove the learning effect in Experiment 2a: when both target and foil pictures at test could contain a novel object, participants were still able to select the correct meaning of the HCE structure at above chance-levels. Participants were thus able to generalize the grammar that they learned at the training phase to new semantic content never actually been seen in the world.

Figure 3 displays the mean accuracy scores in all four experiments (1a, 1b, 2a and 2b) each for the full set of participants and for the selection of participants meeting the 0-LoE learning criterion. Additionally, Supplementary Figure A shows the accuracy scores on the comprehension task, in each experiment for the individual participants. We will further compare experiments in the next section. Regarding the effects of the number of levels of embedding analyzed in all four experiments, it should be noted that the number of levels of embedding in a sentence correlates with sentence length. Hence, sentence length in itself might play a role in the acquisition of complex structures. If so, we might expect this effect to be seen in all conditions, however, and we did not. Also, recent AGL study without semantics that disentangled the effects of sentence length and sentence complexity (number of LoE's) suggests an influence of complexity, not sentence length per se in learning complex structures (Poletiek et al., 2018).

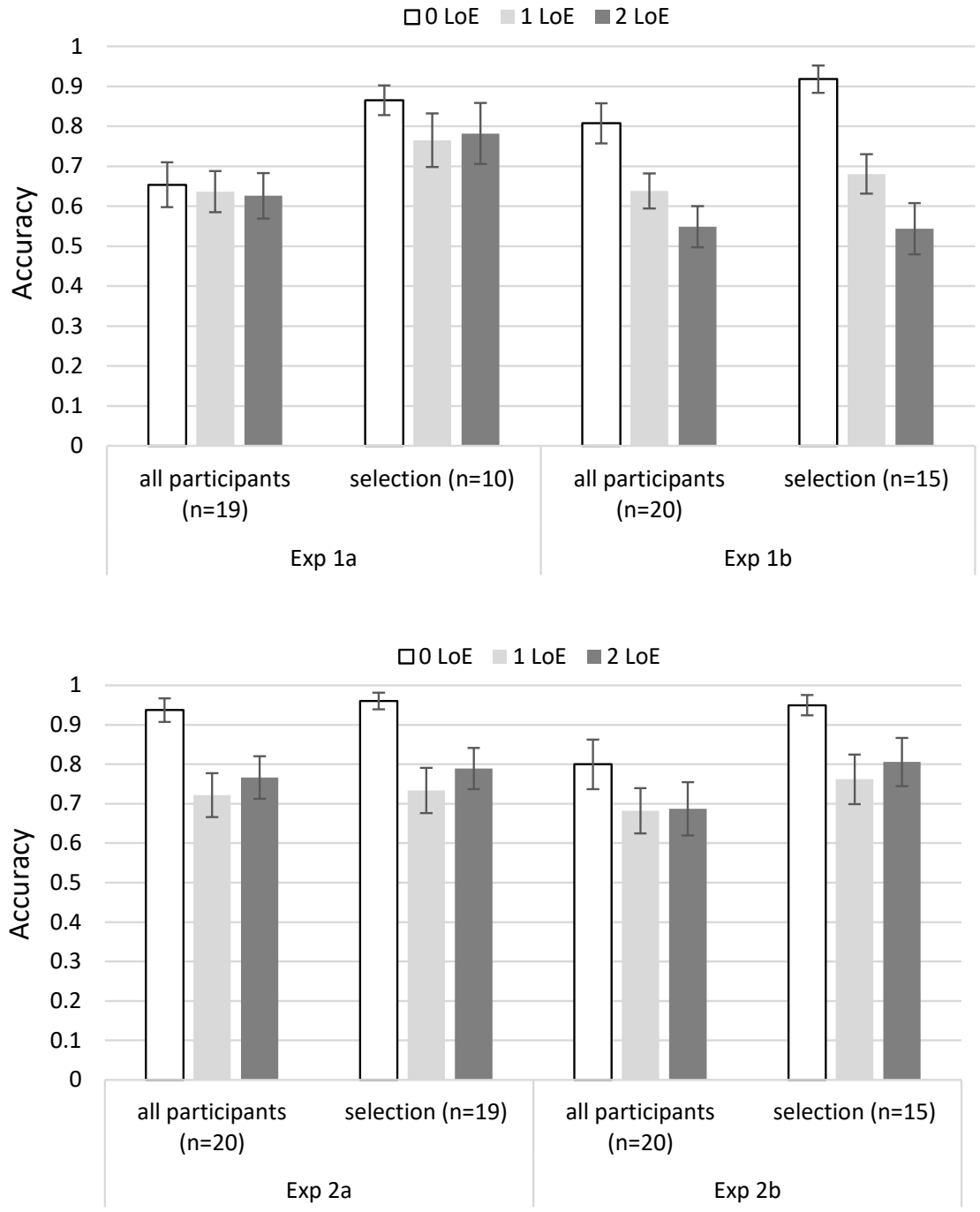


Figure 3: Accuracy scores on comprehension task, in each experiment for the full set of participants and for the subset showing learning of the basic AB structures only (selection). In Experiment 1a and 1b, the test sentences comprised familiar objects only, in Experiment 2a and 2b, the test sentences were about novel objects. In Experiment 2b, the semantic relation between the objects described by the AB clauses, was unspecified (In Experiment 1a it was specified). In Experiment 2b, both correct and foil picture of the test sentence comprised novel objects (In Experiment 2a only the correct picture featured a novel object). Error bars represent SEM.

Generalizing grammar knowledge to novel semantic content:

Comparison of Experiments 1a and 2b

Experiment 2a and 2b suggests that participants could use the knowledge inferred during training about grammatical dependencies between word categories and between constituents, to correctly extend the interpretation of test sentences to information that was novel to them. In order to test whether generalization of the HCE structure to novel AB pairs was different than performance only to novel sentences containing familiar and AB pairs, we performed an exploratory analysis that compared Experiment 1a and 2b. These two experiments are similar in terms of design except for the variable of interest; i.e., the novelty of semantic content in the test task. In both experiments the semantic cue for the relative positions of AB pairs in a sentence was the same (objects described by outer pairs were positioned left to those described by inner pairs) and test items could be comprehended on the basis of grammar knowledge only. A significant effect of the factor Experiment would indicate that novelty affects performance on the task. If learners perform better on sentences featuring old AB pairs compared to new AB pairs, then experience will have driven learning. If however, sentences with new AB pairs (and hence new objects) show equal or better performance compared to old AB pairs, this would be consistent with category learning and generalization across categories.

Aggregating over Exp 1a and 2b (including all participants $n = 39$), a logit mixed model with random intercepts for participants and items, random slopes for LoE by participants and Experiment by items, and mean-centered fixed effects for LoE and Experiment ($N = 1170$; log-likelihood = -627.3) showed that the intercept was significantly larger than 0 ($\beta_0 = 1.14$, $SE = 0.27$, $Z = 4.23$, $p < .001$), indicating that, on average, participants performed better than chance. Neither the main effect of LoE ($\beta_{LoE} = -0.13$, $SE =$

0.13, Z = -0.95, p = .35), nor the LoE*Experiment interaction effect ($\beta_{LoE*Exp}$ = -0.15, SE = 0.21, Z = -0.75, p = .46), nor the main effect of Experiment was significant (β_{Exp} = 0.58, SE = 0.53, Z = 1.10, p = .27), showing that, overall performance was comparable in Exp 1a and 2b. A between-subjects Bayesian t-test for the factor Experiment showed a JSZ Bayes factor BF = 2.02, which indicates anecdotal evidence in favor of the null-hypothesis. This suggests that overall performance was comparable in Exp 1a and 2b (see Figure 3).

Overall, we found no difference in performance between sentences with a HCE structure containing new meaning not experienced before and similar sentences with familiar meaning. In fact, 'comprehending' the linguistic description (AB) of single objects was not more difficult for new objects than for familiar objects. Our finding is consistent with the idea that learners acquire the productive feature of the language system to describe *any* information be it previously experienced or never experienced before.

General Discussion

Learning complex recursive structures from artificial languages in the laboratory has proven a challenge for previous studies. Up to now, learning effects have only been shown when additional phonological cues, memory cues, or cues stemming from the organization of the learning sample indicate the language structure (e.g., Fedor et al., 2012; Lai & Poletiek, 2011, 2013; Mueller et al., 2010; Poletiek & van Schijndel, 2009; Poletiek & Lai, 2012; MacDonald, 2016; Poletiek et al., 2018). Furthermore, these previous studies have often oversimplified the complexity of dependency rules in natural language, by investigating dependencies only between particular words or non-word tokens, rather than categories of words, and neglecting higher-order dependencies between parts of sentences. Finally, previous work in the artificial language paradigm has often tested participants only on a grammaticality judgment task. In contrast, the present study used a comprehension task as an

indicator of learning, assuming that production and comprehension are the essential goals of language learning. In four experiments, we have shown that learning recursive grammatical constructions for dependencies between categories of words and for dependencies between constituents, can be readily accomplished from exposure to a language that is accompanied by visual referents expressing the semantics of the language. Participants were able to correctly interpret novel sentences under these conditions.

Learning dependencies between word categories and constituents

Previous studies of hierarchical structures using artificial languages have mostly focused on dependencies between items rather than categories, but studies of other linguistic structures have been tested in terms of relations between categories. Endress and Bonatti (2007), for instance, trained participants on a language with three categories of words (A, X, and B) that were defined by their position in a sentence (AXB). Pairs of words in the A and B categories always co-occurred during training – so if word A_1 occurred in the first position, word B_1 occurred in the third position. After training, participants were tested on whether they had learned dependencies between particular A_i-B_i pairs, by testing on preference for A_iXB_i sequences, or whether they generalized to accept sequences involving words of the same category which did not respect the precise dependencies but conformed to the positional constraints, i.e., A_iXB_j.

The results of Endress and Bonatti (2007) demonstrated that participants were able to learn a grammar defined in terms of categories of words appearing at different positions in the sentence. Yet, the study did not distinguish between learning *dependencies* between those categories of A and B words, and learning the *relative positions* of words in these sequences. Our present findings further demonstrate that such category dependencies between *word categories* and *constituents* in a complex hierarchical structure can be acquired by participants learning a novel language.

Studies of language learning, such as the Endress and Bonatti (2007) experiments, have been interpreted in terms of triggering symbolic manipulations that apply to linguistic stimuli rather than indicating statistical learning sufficient for deriving syntactic structure (though see also Frost & Monaghan, 2016; Marcus, Vijayan, Rao, & Vishton, 1999; Peña, Bonatti, Nespor, & Mehler, 2002; Wonnacott, Newport & Tanenhaus, 2008). The results of our studies do not necessitate assuming that the learning is rule-based or algebraic, rather than statistical. Participants learn that the language contains categories, and that the syntax indicates relations between those categories. Acquisition of such dependencies is difficult for simple statistical learning mechanisms, such as simple recurrent networks (Endress & Bonatti, 2007, though see Onnis, Christiansen, Chater, & Gomez, 2003, for an indication that such learning is possible). But simple recurrent networks tend to instantiate only very simple, local statistical associations in predicting the upcoming word. Word *category* dependencies learning is likely to require clustering of words into categories and then determining the (statistical) dependencies between those groups of words. Such an approach is entirely consistent with statistical learning that can efficiently compute the structure of a set of stimuli (see, e.g., Gerken, 2010; French, Addyman, and Mareschal, 2011).

Semantics driven HCE learning

It is interesting to consider the role of the semantic referents in learning HCEs, in the current study. First, acquiring the dependencies between categories involves being able to illustrate how the dependencies modify the meanings of sentences as words appear in different positions. If the stimuli had no meaning, for instance, it would not be possible to distinguish whether participants had learned the semantic *effect* of the relation between the A and B words in an $A_1A_2B_2B_1$ sequence, from learning the relation between A and B words in an $A_1A_2B_1B_2$ sequence. Second, learners need semantic reference to learn how the positions of *constituents* (AB pairs) determine their dependencies and affect sentence meaning of a

HCE structure with multiple clauses. This was evidenced in Experiment 1b, where this type of reference was absent. Without semantics, the language would reduce to a sequence of As followed by a sequence of Bs, as in the Fitch and Hauser (2004) studies, which are not sufficient to test whether dependencies between As and Bs and between constituents AB, have been acquired (de Vries et al., 2008). The role of semantics in illustrating the language structure may have been fundamental in directing participants to the dependencies, which are otherwise difficult to track because of their distant separation within HCE sentences.

The two types of referential hints (about the relation between A's and B's and between AB pairs) embodied in the spatial configuration of events in the world, are often likely to be present in the semantic events speakers are talking about in natural language use. For example, in the sentence: *the girl the boy kisses laughs*, the binding pattern can be derived quite easily from observing a boy kissing (who is doing what) and a girl laughing, and from which action is done to whom (*boy* to *girl*) (see Poletiek & Lai, 2012). Although the relation between AB-pairs we implemented in our artificial language study is not as rich as role assignment rules for constituents in natural languages, spatial cues can be very strong for parsing natural sentences as well (Chang, 2002). The explanation of the learnability of these notoriously difficult structures (Gomez, 2002; Newport & Aslin, 2004), is then grounded in experience of world knowledge.

Another potentially important function of semantics is in facilitating generalization of the language system to new content, by transferring characteristics of the domain of the referents to the characteristics of the language (e.g., Chang, 2002; Poletiek & Lai, 2012). For example, if learners see many combinations of colors and shapes in the world, and they know the words for shapes and colors, it is a short step to infer that new colored shapes might be described in the same way as has been previously experienced. This semantic bootstrapping process might have been induced in our artificial language study where the characteristics of

experienced objects (AB pairs) are easily generalized the four objects that are omitted from training (see, e.g., Gerken, 2010).

Implicit versus explicit learning

A question often raised in studies of language learning using artificial materials is whether the learning is explicit (akin to a reasoning process) or implicit (without awareness of the knowledge acquired) and what this tells about the nature of complex language learning. (Van den Bos & Poletiek, 2010; Rohrmeier, Fu & Dienes, 2012). The standard assumption is that adults in the artificial language tasks, and children with natural language, learn the rules implicitly, as they are unable to verbalize their knowledge about complex dependencies. The nature of the learning process was not the focus of the present study and does not affect role of world knowledge semantics on learning, suggested by our results. However, our paradigm and results suggest the possibility that natural language learning recruits some 'reasoning', 'problem solving', and 'cross items learning' mechanisms (e.g., learning the positional rules of word categories by comparing red ball, green ball and red house) typically referred to as 'explicit' learning. Our study cannot inform conclusively about which of the two processes underlie semantics based HCE learning; rather it questions the distinction itself. As the present results suggest, explicit reasoning about the outer world can be a strong and helpful cue for learning implicitly the complex sequential rules of language.

Testing grammar knowledge in artificial language studies

Our study also demonstrates the importance of the type of test of grammar knowledge used in the artificial grammar learning paradigm, for the generalizability of the results outside the lab. Grammaticality judgments to test HCE's where dependencies are defined over particular words rather than word categories, can be highly accurate in the context of an artificial grammar learning experiment, and a poor indication of learning a natural HCE. Moreover, grammaticality judgments and comprehension seem to reveal different

occasionally inconsistent aspects of language knowledge, as suggested by research with a non hierarchical artificial language that tested learners on both tasks (Wonnacott, Newport & Tanenhaus, 2008). In grammaticality judgment tasks for artificial languages, learning is "successful" only if participants rate new AB pairs of words as grammatically *unacceptable*. Interestingly, this response is essentially contradictory to the generalization requirement for grammar learning of *natural* language. Generalizing across words A's and B's and constituents AB's as our data suggest, is in fact what our participants were inclined to do, and should do to become proficient language users.

In conclusion, our data allow us to specify how language usage for semantic purposes interacts with complex syntax learning, involving long distance binding. Additionally, our design clarifies the difficulty of finding successful learning of HCEs with classical artificial language learning procedures (de Vries et al., 2008), where the influence of semantics is disregarded. Our experimental results support the view that binding in vision guides binding in the syntax (Chang, 2002). Our studies also offer a new perspective on the question about whether complex syntactic structures are processed hierarchically at all, or whether they are processed as linear sequences (Frank & Bod, 2011; Frank, Bod, & Christiansen, 2012): Even if sentence structure is processed linearly, sentence *meaning* might be the space within which hierarchical constructions are built.

Acknowledgements

Correspondence concerning this article may be sent to Fenna Poletiek poletiek@fsw.leidenuniv.nl or poletiek@mpi.nl. We thank Elise Hopman for very valuable comments on earlier versions of this article. We also thank Roy de Kleijn, Rebecca Frost and Christine Schoetensack for help with data collection, and Ben Wilson and Chris Petkov for their feedback on the present work. Padraic Monaghan was supported by the Economic and Social Research Council (UK), grant number ES/L008955/1.

References

Amato, M., & MacDonald, M. (2010). Sentence processing in an artificial language: Learning and using combinatorial constraints. *Cognition*, *116*(1), 143--148.

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

Bahlmann, J., Schubotz, R. I., & Friederici, A. D. (2008). Hierarchical artificial grammar processing engages Broca's area. *Neuroimage, 42,* 525-534.

Bates, D. M., Maechler, M., Bolker, B. & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software, 67*, 1–48.

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

Cassidy, K.W. & Kelly, M.H. (2001). Children's use of phonology to infer grammatical class in vocabulary learning. *Psychonomic Bulletin and Review*, *8*, 519-523.

Chang, F. (2002). Symbolically speaking: A connectionist model of sentence production. *Cognitive Science*, *26*, 609-651.

Cholewiak, D. M., Sousa-Lima, R. S., & Cerchio, S. (2013). Humpback whale song hierarchical structure: Historical context and discussion of current classification issues. *Marine Mammal Science*, *29*, E312-E332.

Chomsky, N.(1957). *Syntactic structures*. The Hague: Mouton.

Christiansen, M. H., & MacDonald, M. C. (2009). A usage-based approach to recursion in sentence processing. *Language Learning, 59*, 126-161.

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

de Vries, M.H., Monaghan, P., Knecht, S. & Zwitserlood, P. (2008). Syntactic structure and artificial grammar learning: The learnability of embedded hierarchical structures. *Cognition 107* (2), 763-774.

Endress, A. D., & Bonatti, L. L. (2007). Rapid learning of syllable classes from a perceptually continuous speech stream. *Cognition*, *105*, 247-299.

Everett, D. L. (2005). Cultural constraints on grammar and cognition in Piraha: Another look at the design features of human language. *Current Anthropology, 46,* 621-646.

Fedor, A., Varga, M., & Szathmáry, E. (2012). Semantics boosts syntax in artificial grammar learning tasks with recursion. *Journal of Experimental Psychology: Learning, Memory and Cognition, 38(3),* 776-782.

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

Fitch, W.T., Hauser, M.D. & Chomsky, N. (2005). The evolution of the language faculty: Clarifications and implications. *Cognition, 97,* 179-210.

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.

Frank, S.L., & Bod, R. (2011). Insensitivity of the human sentence processing system to hierarchical structure. *Psychological Science 22 (6),* 829-834.

Frank, S. L., Bod, R., & Christiansen, M. H. (2012). How hierarchical is language use? *Proceedings of the Royal Society B: Biological Sciences*, *279*, 4522-4531.

French, R. M., Addyman, C., & Mareschal, D. (2011). TRACX: A recognition-based connectionist framework for sequence segmentation and chunk extraction. *Psychological Review*, *118*, 614-636.

Friederici, A.D., Bahlmann, J., Heim, S., Schubotz, R.I. & Anwander, A. (2006). The brain differentiates human and non-human grammars: Functional localization and structural connectivity. *Proceedings of the National Academy of Sciences, 103,* 2458-2463.

Frost, R. L. A., & Monaghan, P. (2016). Simultaneous segmentation and generalisation of non-adjacent dependencies from continuous speech. *Cognition, 147*, 70-74.

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

Gerken, L. A. (2010). Infants use rational decision criteria for choosing among models of their input. *Cognition, 115*, 362-366.

Goldberg, A., & Suttle, L. (2010). Construction grammar. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*, 468-477.

Gómez, R. L. (2002). Variability and detection of invariant structure, *Psychological Science, 13*, 431-436.

Hauser, M. D., Chomsky, N., & Fitch, W. T. (2002). The faculty of language: What is it, who has it, and how did it evolve? *Science*, *298*, 1569-1579.

Jackendoff, R. (2010). *Meaning and the lexicon: The parallel architecture 1975–2010*. Oxford: Oxford University Press.

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language, 59*, 434–446.

Kelly, M.H. (1992). Using sound to solve syntactic problems: The role of phonology in grammatical category assignments. *Psychological Review*, *99*, 349-364.

Lai, J., & Poletiek, F.H. (2013). How “small” is “starting small” for learning hierarchical centre-embedded structures? *Journal of Cognitive Psychology, 25(4*), 423-435

Lai, J. & Poletiek, F. H. (2011). The impact of adjacent-dependencies and staged-input on the learnability of center-embedded hierarchical structures. *Cognition, 118(2)*, 265-273

Levelt, W. J. (2019). On Empirical Methodology, Constraints, and Hierarchy in Artificial Grammar Learning. *Topics in Cognitive Science*, 1-15.

MacDonald, M. C. (2016). Speak, act, remember: The language-production basis of serial order and maintenance in verbal memory. *Current Directions in Psychological Science*, 25(1), 47-53.

Marcus, G. F., Vijayan, S., Rao, S. B., & Vishton, P. M. (1999). Rule learning by seven-month-old infants. *Science, 283,* 77-80.

Mueller, J., Bahlmann, J., & Friederici, A. (2010). Learnability of embedded syntactic structures depends on prosodic cues. *Cognitive Science*, *34*(2), 338--349.

Moeser, S. D., & Bregman, A. S. (1972). The role of reference in the acquisition of a miniature artificial language. *Journal of Verbal Learning and Verbal Behavior*, *11*, 759-769.

Moeser, S. D., & Olson, J. (1974). The role of reference in children's acquisition of a miniature artificial language. *Journal of Experimental Child Psychology, 17*, 204-218

Monaghan, P., Chater, N., & Christiansen, M.H. (2005). The differential contribution of phonological and distributional cues in grammatical categorization. *Cognition*, *96,* 143-182.

Morgan, J. L., & Newport, E. L. (1981). The role of constituent structure in the induction of an artificial language. *Journal of verbal learning and verbal behavior, 20(1)*,67-85.

Naigles, L. (1990). Children use syntax to learn verb meanings. *Journal of child language*, *17*(2), 357-374.

Newport, E.L. & Aslin, R.N. (2004). Learning at a distance 1. Statistical learning of non-adjacent dependencies. *Cognitive Psychology, 48*, 127-162.

Onnis, L. Christiansen, M., Chater, N. & Gomez, R. (2003). Reduction of uncertainty in human sequential learning: Evidence from artificial language learning. *Proceedings of*

The 25th Annual Conference of the Cognitive Science Society. (pp.886-891). Mahwah, NJ: Lawrence Erlbaum.

Onnis, L., Monaghan, P., Christiansen, M. H., & Chater, N. (2004). Variability is the spice of learning, and a crucial ingredient for detecting and generalizing in nonadjacent dependencies. In *Proceedings of the 26th Annual Meeting of the Cognitive Science Society* (pp.1047-1052)

Öttl, B., Dudschig, C., & Kaup, B. (2017). Forming associations between language and sensorimotor traces during novel word learning. *Language and Cognition, 9(1),* 156-171.

Peña, M., Bonatti, L. L., Nespor, M., & Mehler, J. (2002). Signal-driven computations in speech processing. *Science, 298,* 604–607.

Perek, F., & Goldberg, A. E. (2015). Generalizing beyond the input: The functions of the constructions matter. *Journal of Memory and Language*, *84*, 108-127.

Perek, F., & Goldberg, A. E. (2017). Linguistic generalization on the basis of function and constraints on the basis of statistical preemption. *Cognition*, *168*, 276-293.

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

Petkov, C. I., & Wilson, B. (2012). On the pursuit of the brain network for proto-syntactic learning in non-human primates: conceptual issues and neurobiological hypotheses. *Philosophical Transactions of the Royal Society B: Biological Sciences, 367*(1598), 2077-2088.

Pinker, S., & Jackendoff, R. (2005). The faculty of language: What's special about it? *Cognition*, *95*(2), 201-236.

Poletiek, F. H. (2011). What in the world makes recursion so easy to learn? A statistical account of the staged input effect on learning a centre embedded hierarchical structure in AGL. *Biolinguistics, 5, 36-42.*

Poletiek, F.H., & Lai, J. (2012). How semantic biases in simple adjacencies affect learning a complex structure with non-adjacencies in AGL: A statistical account. *Philosophical Transactions of the Royal Society B,* 367, 2046–2054.

Poletiek, F.H., & Van Schijndel, T.J.P. (2009). Stimulus set size and grammar coverage in artificial grammar learning. *Psychonomic Bulletin and Review, 16(6),* 1058-1064.

Poletiek, F. H., Conway, C. M., Ellefson, M. R., Lai, J., Bocanegra, B. R., & Christiansen, M. H. (2018). Under what conditions can recursion be learned? Effects of starting small in artificial grammar learning of center-embedded structure. *Cognitive Science, 42(8)*, 2855-2889.

Reeder, P. A., Newport, E. L., & Aslin, R. N. (2013). From shared contexts to syntactic categories: The role of distributional information in learning linguistic form-classes. *Cognitive Psychology*, *66*(1), 30-54.

Rohrmeier, M., Fu, Q., & Dienes, Z. (2012). Implicit learning of recursive context-free grammars. *PloS one*, *7*(10), e45885.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*(2), 225-237.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*, 1926-1928.

Smith, L.B., & Yu, C. (2008). Infants rapidly learn word–referent mappings via cross-situational statistics. *Cognition, 106*, 1558–1568.

Van den Bos, E., & Poletiek, F. H. (2010). Structural selection in implicit learning of artificial grammars. *Psychological Research*, *74*(2), 138-151.

Wilson, B., Spierings, M., Ravignani, A., Mueller, J. L., Mintz, T. H., Wijnen, F., Van der Kant, A, Smith, K., & Rey, A. (2018). Non-adjacent Dependency Learning in Humans and Other Animals. *Topics in Cognitive Science. 1-16.*

Wonnacott, E., Newport, E. L., & Tanenhaus, M. K. (2008). Acquiring and processing verb argument structure: Distributional learning in a miniature language. *Cognitive Psychology*, *56*(3), 165-209.

Supplementary Analyses

Experiment 1a

ANOVAs. The mean proportion of accurate picture selection was above chance level (M = .64; SD =.23), $t(18) = 2.62$, $p < .05$. The mean (SD in parentheses) proportion of correct selections for 0-LoE sentences was .65 (.25), .64 (.23) for 1-LoE, and .63 (.25) for 2-LoE sentences. An ANOVA with LoE as within subjects variable showed no significant effect of LoE on performance in the picture matching task, $F(2,36) = .24$, p =.79, $\eta_p^2 = .01$.

Nine participants failed to perform above chance level at the 0-LoE items, suggesting that they had not learned the vocabulary of the study, or did not otherwise pay attention to the task. The mean proportion of accurate picture selection of the remaining ten participants (i.e., with accuracy ≤ .50) was higher and clearly above chance level (M = .80; SD =.19), $t(9) = 4.87$, $p < .01$. The mean (SD in parentheses) proportion of correct selections for 0-LoE sentences was .86 (.12), .76 (.22) for 1-LoE, and .78 (.25) for 2-LoE sentences. There was no significant effect of LoE on performance in the picture matching task, for these participants $F(2,18) = 1.95$, p =.17, $\eta_p^2 = .18$.

Logit mixed model. Nine participants failed to perform above chance level on the 0-LoE items, suggesting that they had not learned the vocabulary of the study, or did not otherwise pay attention to the task. We also analyzed the data excluding these participants (i.e., with accuracy ≤ .50), focusing only on performance of learners who could be assumed to have solid knowledge about the vocabulary, the basic AB pairs and their semantic referent objects. Indeed, as previous research suggests, early robust learning of the simple basic structures of a HCE is a precondition for learning the full complex structures subsequently (Lai & Poletiek, 2011). Participants who poorly recognized the individual objects on the basis of their shape and color names, might underperform on the higher levels of embedding items, just because of insufficient knowledge of the basic structures; i.e. AB pairs in our experiment. When we exclude the nine participants that failed to perform above chance level on the 0-LoE

items (N = 300; log-likelihood = -128.2), we observed the same pattern of results: the intercept was significantly larger than 0 (β_0 = 2.04, SE = 0.54, Z = 3.80, p < .001), indicating that, on average, participants performed better than chance with an estimated mean proportion of correct selections of .89. LoE was not significant (β_{LoE} = -0.02, SE = 0.38, Z = -0.04, p = .97), indicating that the mean proportion of correct selections did not differ for different number of embeddings.

Experiment 1b

ANOVAs. The mean proportion of accurate picture selection was above chance level (M = .65; SD =.17), $t(19) = 4.06$, $p < .01$. An ANOVA with LoE as within subjects variable, showed a significant effect of the level of complexity on performance in the picture matching task, $F(2,38) = 10.08$, $p < .001$, $\square_p\square = .35$. T-tests showed that 0-LoE items (M = .81; SD = .23) were better comprehended than both 1-LoE (M =.64; SD = .20), $t(19) = 3.03$, p <.01, and 2-LoE (M =.55; SD = .24) items, $t(19) = 3.59$, $p < .01$. There was only a marginally significant difference between comprehension of 1-LoE test items versus 2-LoE items, $t(19)$ = 2.01, p = .06.

As for Experiment 1a, we also ran the analyses excluding five participants with accuracy ≤ .50 on the 0-LoE items. Overall mean proportion of accurate picture selection was above chance level (M = .69; SD =.16), $t(14) = 4.65$, $p < .001$. An ANOVA with LoE as within subjects variable, showed a significant effect of the level of complexity on performance in the picture matching task, $F(2,28) = 19.82$, $p < .001$, $\eta_p^2 = .59$. T-tests showed that 0-LoE items (M = .92; SD = .14) were better comprehended than both 1-LoE (M =.68; SD = .20), $t(14) = 3.90$, p <.01, and 2-LoE (M =.54; SD = .26) items, $t(14) = 5.22$, $p < .001$. 1-LoE test items were better comprehended than 2-LoE items, $t(14) = 3.05$, $p < .01$.

Logit mixed model. When we exclude the five participants that failed to perform above chance level (i.e., with accuracy ≤ .50), on the 0-LoE items (N = 570; log-likelihood = -

332.4), we observed the same pattern of results: the intercept was significantly larger than 0 (β_0 = 0.85, *SE* = 0.21, *Z* = 4.00, $p < .001$), indicating that, on average, participants performed better than chance with an estimated mean proportion of correct selections of .70. LoE was significant (β_{LoE} = -0.72, *SE* = 0.20, *Z* = -3.61, $p < .001$), indicating that the mean proportion of correct selections decreased by approximately .15 for each additional level of embedding.

Comparison of Experiments 1a and 1b

ANOVAs. When we aggregate the data of Experiment 1a and 1b, we observed no difference between the experiments in overall learning, as indicated by performance on the picture selection task, was found (F (1,37) = .18, p = .68, η_p^2 = .01, in an ANOVA with experiment as between subjects and LoE as within subjects factor). Also, performance on HCE sentences without the hierarchical relation between the word pairs, became significantly worse, as the number of LoE increased: $F(2,46 = 5.86, p < .01$ for the interaction between LoE (1 vs 2) and Experiment (1a vs 1b). T-tests revealed a significant difference between Experiment 1a and Experiment 1b on the 2-LoE items only, $t(23) = 2.28$, $p < .05$. When the hierarchical reference rule was absent (in Experiment 1b), performance on complex sentences with 2-LoE items did not exceed chance level, $t(14) = .66$, $p = .52$.

Experiment 2a

ANOVAs. Mean proportion of accurate picture selection was above chance level (M = .77; SD =.21), $t(19) = 5.59$, $p <.001$. Picture matching performance differed significantly for items with different levels of embedding, $F(2,38) = 12.21$; $p < .001$, η_p^2= .39, with better performance for 0-LoE items (M = .94; SD = .14) than on both 1-LoE (M =.72; SD = .25), $t(19) = 4.15$, $p <.01$, and 2-LoE (M =.77; SD = .25) items, $t(19) = 3.66$, $p <.01$.

Next, one participant's data was removed from the analysis because accuracy was ≤ .50 for 0-LoE items. Mean proportion of accurate picture selection was above chance level (M

= .79; *SD* =.20), $t(18) = 6.08$, $p < .001$. Picture matching performance differed significantly for items with different levels of embedding, $F(2,36) = 12.36$; $p < .001$, $\eta_p^2 = .40$, with better performance for 0-LoE items (*M* = .96; *SD* = .09) than on both 1-LoE (*M* =.73; *SD* = .23), $t(18) = 4.24$, $p < .001$, and 2-LoE (*M* =.79; *SD* = .23) items, $t(18) = 3.48$, $p < .01$. 1- and 2-LoE items were not significantly different, $t(18) = 1.44$, $p = .16$.

Logit mixed model. When excluding one participant that failed to perform above chance level (i.e., with accuracy ≤ .50), on the 0-LoE items (*N* = 570; log-likelihood = -245.0), we observed the same pattern of results: the intercept was significant ($\beta_0 = 2.04$, *SE* = 0.44, $Z = 4.62$, $p < .001$), indicating that, on average, participants performed better than chance with an estimated mean proportion of correct selections of .88, and LoE was not significant ($\beta_{LoE} = -0.26$, *SE* = 0.32, $Z = -0.82$, $p = .41$), indicating that the mean proportion of correct selections did not differ for different number of embeddings.

Experiment 2b

ANOVAs. Overall, participants' accuracy at test was above chance level (*M* = .70; *SD* =.26), $t(19) = 3.50$, $p < .01$. An ANOVA with LoE as within subjects variable, revealed that picture matching performance differed significantly for test-strings with different levels of embedding, $F(2,38) = 3.28$, $p < .05$, $\eta_p^2 = .15$: 0-LoE items (*M* = .80; *SD* = .29) were better comprehended than both 1-LoE (*M* =.68; *SD* = .26), $t(19) = 2.16$, $p < .01$, and 2-LoE (*M* =.69; *SD* = .31) items, $t(19) = 2.32$, $p < .05$. Participants reached a similar level of accuracy on 1- and 2-LoE items, $t(19) = -.10$, $p = .93$.

Next, data from five participants were excluded from the analysis, since these participants reached an accuracy of ≤ .50 on the 0-LoE items. Overall, participants' accuracy at test was above chance level (*M* = .81; *SD* =.20), $t(14) = 5.85$, $p < .001$. An ANOVA with LoE as within subjects variable, revealed that picture matching performance differed significantly for test-strings with different levels of embedding, $F(2,28) = 6.63$, $p < .01$, $\eta_p^2 =$

.32: Again 0-LoE items (M = .95; SD = .10) were better comprehended than both 1-LoE (M =.76; SD = .25), $t(14) = 3.38$, p <.01, and 2-LoE (M =.81; SD = .25) items, $t(14) = 3.54$, p < .05. Participants reached a similar level of accuracy on 1 and 2-LoE items, $t(14) = -.77$, p = .46.

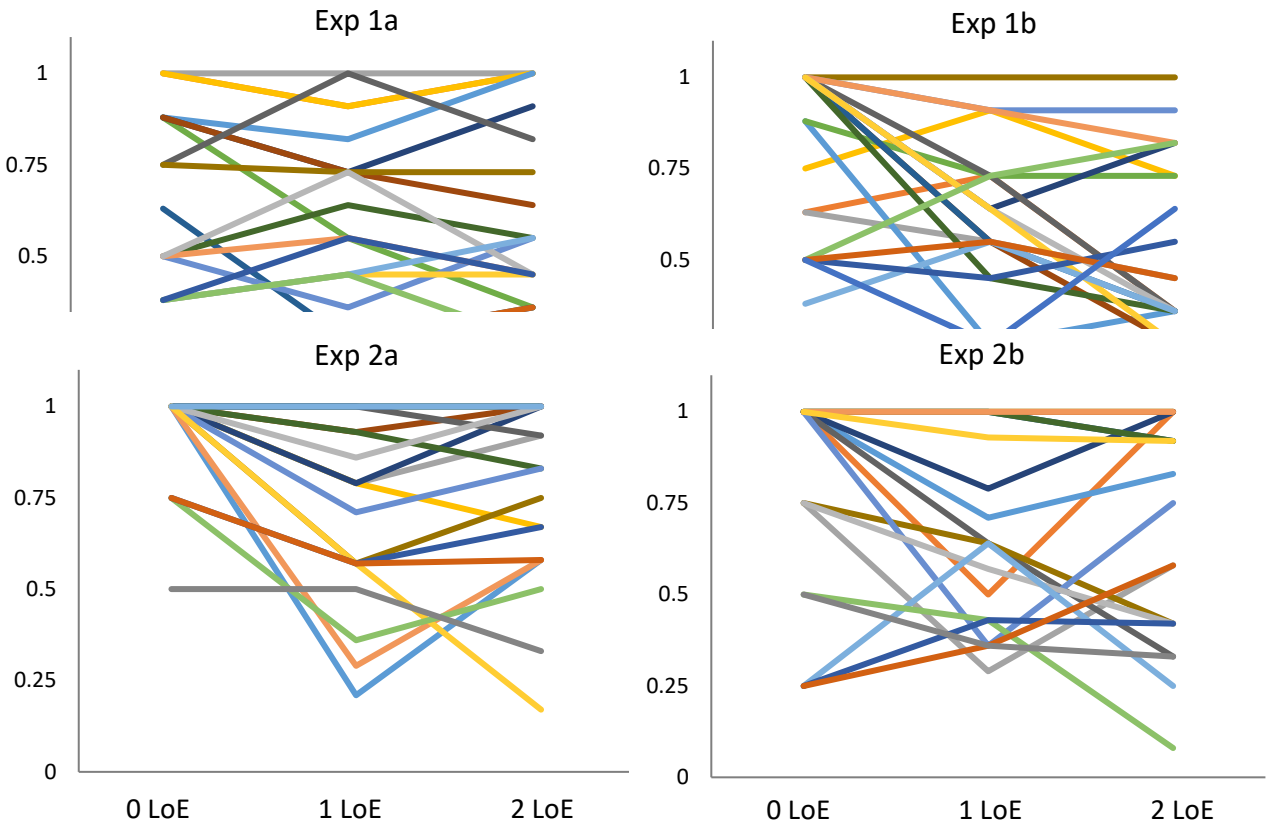
Overall performance in Experiment 2b (M = .81, SD = .20) was similar to that in Experiment 2a (M = .79, SD = .20), $F < 1$, and remained similar for items with different levels of embedding, $F < 1$ (for the interaction between LoE and Experiment: 2a vs 2b).

Logit mixed model. When we exclude the five participants that failed to perform above chance level (i.e., with accuracy ≤ .50), on the 0-LoE items (N = 450; log-likelihood = -181.5), we observed the same pattern of results: the intercept was significantly larger than 0 (β_0 = 2.34, SE = 0.55, Z = 4.23, p < .001), indicating that, on average, participants performed better than chance with an estimated mean proportion of correct selections of .91. LoE was not significant (β_{LoE} = -0.33, SE = 0.39, Z = -0.84, p = .40), suggesting that the mean proportion of correct selections did not differ for different number of embeddings.

Comparison of Experiments 1a and 2b

ANOVAs. We performed an ANOVA with Experiment as between subjects factor, and Level of Embedding as within subjects factor, on the combined data of Experiment 1a and 2b. Results indicate no significant difference between experiments: performance with novel AB pairs was as accurate as performance with old AB pairs, $F(1,37) = 0.26$, p = .28, η_p^2 = .03, see Figure 3. The main effect of LoE, $F(2,74) = 2.99$, p = .06, η_p^2 = .08, and the interaction between experiment and LoE, $F(2,74) = 1.52$, p = .23, η_p^2 = .04, were not significant.

Supplementary Figure A



Supplementary Figure A. Accuracy scores on comprehension task, in each experiment for the individual participants. In Experiment 1a and 1b, the test sentences comprised familiar objects only, in Experiment 2a and 2b, the test sentences were about novel objects. In Experiment 2b, the semantic relation between the objects described by the AB clauses, was unspecified (In Experiment 1a it was specified). In Experiment 2b, both correct and foil picture of the test sentence comprised novel objects (In Experiment 2a only the correct picture featured a novel object).