Journal of Experimental Psychology:
Learning, Memory, and Cognition

© 2012 American Psychological Association
0278-7393/12/$12.00 DOI: 10.1037/a0026986

RESEARCH REPORT

Semantics Boosts Syntax in Artificial Grammar Learning Tasks With Recursion

Anna Fedor
AQ: 1 Eötvös Loránd University of Sciences and Collegium Budapest

Máté Varga
TiVo, Inc., San Jose, California

Eörs Szathmáry
Eötvös Loránd University of Sciences, Collegium Budapest, and Parmenides Center for the Study of Thinking

Center-embedded recursion (CER) in natural language is exemplified by sentences such as “The malt that the rat ate lay in the house.” Parsing center-embedded structures is in the focus of attention because this could be one of the cognitive capacities that make humans distinct from all other animals. The ability to parse CER is usually tested by means of artificial grammar learning (AGL) tasks, during which participants have to infer the rule from a set of artificial sentences. One of the surprising results of previous AGL experiments is that learning CER is not as easy as had been thought. We hypothesized that because artificial sentences lack semantic content, semantics could help humans learn the syntax of center-embedded sentences. To test this, we composed sentences from 4 vocabularies of different degrees of semantic content due to 3 factors (familiarity, meaning of words, and semantic relationship between words). According to our results, these factors have no effect one by one but they make learning significantly faster when combined. This leads to the assumption that there were different mechanisms at work when CER was parsed in natural and in artificial languages. This finding questions the suitability of AGL tasks with artificial vocabularies for studying the learning and processing of linguistic CER.

Keywords: center-embedded recursion, artificial grammar learning, semantics, familiarity

Artificial grammar learning (AGL) tasks are widely used to test the abilities of different species in learning grammatical rules. During these tasks participants are usually trained and tested on a set of artificial sentences to assess whether they could master the grammatical rule underlying these sentences. Sentences are composed of a set of nonsense words (the vocabulary), which could be anything from actual letters to geometrical shapes but are usually consonant–vowel (CV) syllables. The theory behind this paradigm is that removing semantics from a language makes it possible to research its pure syntax.

After the influential paper of Hauser, Chomsky, and Fitch (2002), active research started using AGL tasks investigating a particular grammar, called center-embedded recursion (CER). In natural language, CER is exemplified by sentences such as “The malt that the rat ate lay in the house.” There are three main characteristics of this sentence: (a) A phrase (*that the rat ate*) is embedded within another (*the malt lay in the house*); (b) there are within-phrase dependencies between different classes of words (here, nouns and verbs; the *malt–lay* and the *rat–ate*); and (c) there is also a dependency between phrases: *the rat ate* qualifies *the malt* (only the malt that the rat ate would do, not just any malt).

The first generation of AGL experiments on CER (e.g., Gentner, Fenn, Margoliash, & Nusbaum, 2006) conformed only to the first characteristics of CER. In these experiments, a four-word-long sentence could be described by the formula of AABB (A^nB^n in general), where As and Bs are arbitrary words from two distinct classes of artificial words. This means that AB phrases are embedded within each other but the dependencies between and within phrases are not modeled. Due to these simplifications, it was possible to solve the tasks (discrimination between grammatical and ungrammatical sentences) without recognizing the recursive structure of sentences, simply by matching the number of As and Bs (Corballis, 2007a, 2007b; Perruchet & Rey, 2005).

A second generation of experiments tried to get around this problem by establishing A–B word pairs. These sentences could be described by the formula $A_1A_2B_2B_1$, where indices denote dependencies between As and Bs; the between-phrase dependencies are

Anna Fedor, Department of Plant Systematics, Ecology and Theoretical Biology, Eötvös Loránd University of Sciences, Budapest, Hungary; and Collegium Budapest, Budapest, Hungary; Máté Varga, TiVo, Inc., San Jose, California; Eörs Szathmáry, Department of Plant Systematics, Ecology and Theoretical Biology, Eötvös Loránd University of Sciences; Collegium Budapest; and Parmenides Center for the Study of Thinking, Munich/Pullach, Germany.

Thanks to Vera Vasas for suggestions on statistics and István Zachar for blind scoring. Work by Eörs Szathmáry has been supported by the National Office for Research and Technology (NAP 2005/KCKHA005) and Hungarian Scientific Research Fund (OTKA; National Scientific Research Fund NK73047).

Correspondence concerning this article should be addressed to Anna Fedor, Eötvös Loránd University of Sciences, Room 7.506, 1/C Pázmány Péter Sétány, Budapest 1117, Hungary. E-mail: fedoranna@gmail.com

still missing, but the within-phrase dependencies are present. These experiments yielded various results. In the experiment of Perruchet and Rey (2005), human participants were not able to learn the grammar after 3 min of habituation. Similarly, in de Vries, Monaghan, Knecht, and Zwitserlood (2008), 50 min of alternating familiarization and test blocks with feedback (230 sentences in sum) was not enough for participants to recognize the structure of sentences. However, there were two studies in which participants managed to learn CER: those of Bahlmann, Schubotz, and Friederici (2008) and Lai and Poletiek (2011). These studies share a number of methodological points (one or more of which were missing from the previous studies): Word pairs and word groups (A and B) were distinguished by phonological cues; training was staged with sentences of increasing length (starting small paradigm); there were alternating habituation and testing blocks; and feedback was provided during testing.

These experiments require different computations from learners at different training stages. The starting small paradigm with staged input means that learners are exposed first to shorter and then to increasingly long sentences. In the case of CER, the first stage of learning involves two-word-long sentences (i.e., word pairs); the second stage involves four-word-long sentences; and the third stage involves six-word-long sentences. During the first stage, where two-word-long sentences are presented, associative learning is required for memorization of word pairs (which is supposedly helped by phonological cues). On the next stage, the task is to recognize the center-embedded structure of sentences. Because of the feedback, participants probably engage in active rule searching as opposed to passive incidental learning. The last stage tests generalization of the rule to longer sentences. It is usually obvious from the instructions given to participants that the rule is the same throughout, so at this stage participants have to learn how to apply the previously learned rule effectively.

Even in these experiments, where learning was successful, extensive training was needed to reach the desired performance. This is quite surprising, seeing that CER is present in all known human languages (but see Everett, 2005) and the ability to parse it was supposed to be a natural and straightforward human ability. Simplifying natural language to syntax plus semantics and comparing it to artificial languages that lack the latter lead to the idea that it is indeed the absence of semantics that makes it so difficult to recognize the center-embedded structure in artificial sentences. We designed an experiment to test this hypothesis, in which we trained participants on artificial sentences that involved different degrees of semanticity. We predicted that learning is made easier at all stages by artificial sentences with semantic content than by sentences with no semantic content.

Method

Participants

Sixty-seven Hungarian native speaker participants (two participants were bilingual), mainly university students, participated in this study ($M = 22.1$ years, $SD = 3.8$; 30 female and 37 male). They were randomly assigned to four groups: There were 18 participants in Group WS, 16 participants in Group WR, 16 participants in Group NR1, and 17 participants in Group NR2. Groups were named after the vocabulary types they were trained on (see below).

Participants had no known disorder and had not taken any drugs that might have influenced memory or attentional abilities. They had normal or corrected to normal vision. They received course credit or light refreshments (chocolate or beer) for their participation.

Stimuli

Vocabularies from which sentences were composed contained six pairs of words; there were six words in Class A and six words in Class B. Every word had exactly one pair from the other class. There were four distinct vocabularies, one for each group of participants (see Table 1). The first vocabulary consisted of two- T1
letter Hungarian words that were selected during a previous short study. In this study, participants (different from those in the present experiment) had to make pairs from a pool of 21 Hungarian two-letter nouns based on free association. Those pairs that were chosen most often made up the first vocabulary. Pairing was mainly based on the semantic relationship of the words; that is why we labeled these words Vocabulary WS (words with semantic relatedness). The second vocabulary consisted partly of words AQ: 2
from Vocabulary WS: Class A was the same as in Vocabulary WS,

Table 1
Four Different Vocabularies From Which Sentences of the Artificial Language Were Generated for the Four Distinct Groups of Participants

Vocabulary WS		Vocabulary WR		Vocabulary NR1		Vocabulary NR2	
Class A	Class B	Class A	Class B	Class A	Class B	Class A	Class B
eb [dog]	*ól* [kennel]	*eb* [dog]	*ón* [tin]	*ev*	*ób*	*nu*	*zi*
én [me]	*te* [you]	*én* [me]	*tó* [lake]	*éz*	*ta*	gi	*pe*
év [year]	*ösz* [autumn]	*év* [year]	*ös* [ancestor]	*ögy*	*fe*	ru	*ve*
fü [grass]	*fa* [tree]	*fü* [grass]	*ma* [today]	*fé*	*ísz*	*fe*	*ko*
íny [gum]	*íz* [flavor]	*íny* [gum]	*ív* [arc]	*ít*	*ön*	*bi*	*mo*
kö [stone]	*út* [road]	*kö* [stone]	*úr* [gentleman]	*kü*	*úl*	*lu*	*co*

Note. In all of the vocabularies, each word from Class A had exactly one pair from Class B (shown in the same row). Vocabulary WS: Hungarian two-letter words, paired mainly semantically, according to a previous study. The English translation of words is given in brackets. Vocabulary WR: Hungarian two-letter words paired randomly. Vocabulary NR1: Nonwords paired randomly and composed mainly from the letters of the words of Vocabulary WS. Vocabulary NR2: Nonwords paired randomly, similar to those used in other studies (e.g., Bahlmann et al. (2008); Friederici et al. (2006).

but Class B contained different words that were chosen so there was no semantic relatedness between A and B words. Moreover, we chose words that had one letter in common with a Class B word in Vocabulary WS; hence, the two vocabularies were phonologically as similar as possible. We labeled the second group of words Vocabulary WR (words randomly paired).

We chose these vocabularies to test whether semantic relationship between words has an effect on learning. We could have generated Vocabulary WR from the words of Vocabulary WS by randomizing the pairs; however, it would have resulted in a vocabulary where there were obviously related words that were not treated as pairs, which could have made the task more difficult. Therefore, we composed Vocabulary WR partly from Vocabulary WS (Class A) and partly from new words (Class B), so there is no obvious semantic relationship between any two words.

The third and fourth vocabularies contained nonwords (CV syllables) randomly paired, so we labeled them NR1 and NR2. Vocabulary NR1 was generated mainly from the letters of words in Vocabulary WS in such a way that no word had a meaning, not even if read backwards (we had to change some of the letters to meet this criterion). Care was taken that words had no meaning in most other languages that Hungarian students usually learn and that word pairs (read together as one word) did not make sense either. As much as possible, the position of letters in words was kept as in Vocabulary WS. In this way, this vocabulary was phonologically similar to Vocabularies WS and WR, but the words had no meaning. Last, Vocabulary NR2 consisted of nonwords that were similar to vocabularies of other studies that were conducted with German-speaking participants (e.g., Bahlmann et al., 2008; Friederici, Bahlmann, Heim, Schubotz, & Anwander, 2006). There were no long vowels, which are very common in Hungarian words, in this vocabulary.

Our motivation to test participants on two different nonword vocabularies was that we realized that Hungarian students learned much more slowly in our previous study (Fedor & Szathmáry, unpublished results) than did German students in Bahlmann et al.'s study; however, the circumstances were quite similar. We thought that the vocabulary that was used in both studies could be more familiar for German native speakers than for Hungarians (even though the vocabulary was phonotactically legal in Hungarian, too). To test this effect, we constructed Vocabulary NR1 using Hungarian-specific vowels. Thus, it sounded more "Hungarian-like" than Vocabulary NR2.

Sentences composed from these vocabularies represent four T2 different levels of diversion from natural language (see Table 2) according to three criteria: phonetic familiarity, words with meanings, and semantic associations between words. Vocabulary NR2 is the least natural; it does not meet any of the above mentioned criteria. All the other vocabularies sound familiar to Hungarian participants. Vocabulary NR1 is composed of nonwords that have no meaning, whereas the remaining two vocabularies are composed of natural words with meaning. Only Vocabulary WS meets all three criteria; however, there are still a lot of differences from natural language.

The rule of CER was used to compose sentences from these vocabularies. In case of two-, four-, and six-word-long sentences, the rules were A_1B_1, $A_1A_2B_2B_1$, and $A_1A_2A_3B_3B_2B_1$, respectively. Indices denote dependencies between words (i.e., an A word and a B word with the same index make up a word pair). In this way, 6 two-word-long, 30 four-word-long, and 120 six-word-long grammatical sentences were composed with each vocabulary.

Ungrammatical sentences were generated by randomly replacing one of the words in the second half of a grammatical sentence by another B word. This violated the structure of word pairs but not the structure of word classes (As and Bs) in sentences, thus ensuring that the error was detectable provided that one was aware of the center-embedded structure of word pairs. B words that were already in the sentence were not excluded from being replacements; thus, word repetitions could occur in four-word-long and six-word-long ungrammatical sentences. This decision was made in accordance with Bahlmann et al. (2008), where such repetitions were also allowed, because we wanted to compare the performance of our participants on Vocabulary NR2 with the performance of participants in the above mentioned study.[1] Replacements were Fn1 performed in all possible positions (but only in one position in a sentence) and thus occurred in the second position of two-word-long sentences; in the third or fourth position of four-word-long sentences; and in the fourth, fifth, or sixth positions of six-word-long sentences.[2] Fn2

Procedure

The procedure followed the schema of the learning period of Bahlmann et al. (2008). In the beginning of the training, participants were given the instructions that they would read the sentences of an artificial language, and their task was to find out the rule according to which the sentences were composed. Training of participants was performed according to the starting small paradigm with staged input (Conway, Ellefson, & Christiansen, 2003): It started with two-word-long sentences (Level 1) and continued with four-word-long (Level 2) and then six-word-long sentences (Level 3).

A training block consisted of a set of 10 familiarization sentences and a set of 10 test sentences. The familiarization set started with an instructional sentence (the whole sentence presented all at once): "Please read carefully the following sentences corresponding to the rule!" During familiarization, sentences followed each other, separated only by a fixation cross in the middle of the screen. All sentences were grammatical. Test sets were also anticipated by an instructional sentence ("Please decide whether the following sentences correspond to the rule or not!"). Test sets were compiled from five grammatical and five ungrammatical sentences, randomly ordered. There was a fixation cross before and a choice of "Yes" or "No" after each sentence. Participants had 3 s to answer and then feedback was given: The right answer flashed on the screen for 250 ms.

Familiarization and test sentences were randomly chosen from the pool of grammatical and ungrammatical sentences without

[1] It can be argued that repetitions make it possible to detect ungrammaticality without learning the grammar of sentences; however, it is very unlikely that participants could pass the test if their decisions had been based solely on repetitions (see calculations for this probability in the Results section).

[2] As an example, see supporting online material for the entire pool of grammatical and ungrammatical sentences for Vocabulary WS, from which training and test sentences were randomly chosen for each participant in Group WS. AQ: 3

Table 2
Similarity of Vocabularies to Natural Language According to Three Criteria

Vocabulary	WS	WR	NR1	NR2
Does the vocabulary sound phonetically familiar?	Yes	Yes	Yes	No
Do the items in the vocabulary have meaning?	Yes	Yes	No	No
Is there semantic relationship between items?	Yes	No	No	No

Note. WS = words with semantic relatedness; WR = words randomly paired; NR = nonwords randomly paired.

replacement until all sentences were used. After that, all sentences were placed back in the pool and the same procedure was applied again. Sentences were visually presented on a computer screen, one word at a time. The first word of sentences started with a capital letter, and sentences were closed by a full stop. Words were shown for 800 ms followed by a 200-ms gap. The fixation cross was shown for 1,000 ms before every sentence.

If a participant had reached nine or 10 correct answers in two consecutive training blocks, the next level with longer sentences followed. Each level consisted of as many blocks as the participant needed to reach the required performance. If a participant had not mastered a level during 20 blocks, the test was finished without proceeding to the higher levels. After the test was finished, participants were asked to write down the rule that they deduced from the sentences.

Results

To find out whether the difficulty of the task was different in the four groups, we performed two kinds of analyses. First we compared the success rate of participants in the four groups (whether they reached the required performance on the different levels and the correctness of their written formulation of the rule), and then we compared the number of training blocks they needed to finish the training.

Whether passing the 90% performance criterion means that the participant understands the rule can be questioned. Because there is a relatively low number of grammatical sentences in Level 1 (6 sentences) and Level 2 (30 sentences), participants could memorize the sentences rather than learn the rule (in fact, sentences—word pairs—had to be memorized in Level 1). However, participants who memorized four-word-long sentences without understanding the rule would not be able to pass the criterion on six-word-long sentences (unless they memorized six-word-long sentences too, which is unlikely). Because there were no participants who passed Level 2 but did not pass Level 3, we can exclude this possibility.

Participants could have passed the 90% performance criterion by basing their decisions solely on detecting word repetition in ungrammatical sentences if there had been four or five ungrammatical sentences with word repetition in two consecutive blocks. This means 8–10 sentences with word repetition in sum out of 10 ungrammatical sentences in two consecutive blocks: If participants categorize sentences with repetition as ungrammatical and sentences without repetition as grammatical, they could have 18–20 correct answers in two blocks and could pass the test. This is obviously undesirable, because we do not want to confound this simple strategy with true understanding of the grammar. However, we did not worry about this, because the probability, according to the binomial distribution, is very small: It is $3.5006^{*}10^{-5}$ and .0202 in the case of four- and six-word long sentences, respectively (calculated from the average percentage of ungrammatical sentences with word repetition across vocabularies: 18% and 43%). In fact, we checked the last two blocks in Level 3 for successful participants, and we found only three cases where more than seven ungrammatical sentences occurred with word repetition. None of these participants mentioned word repetition in their written formulation of the rule. There was only one participant in the four groups who mentioned that sentences with word repetitions were not correct, and he was not successful in passing Level 2.

One participant in Group WR and one participant in Group NR2 did not learn the word pairs and thus were excluded from all further analyses. All other participants reached the 90% criterion on word pairs (Level 1) and proceeded to Level 2. Two participants in Group NR1 and six participants in Group NR2 did not learn the recursive rule in four-word-long sentences during the 20 training blocks provided (400 sentences) and thus did not proceed to Level 3. All successful participants on Level 2 were able to reach the 90% criterion on Level 3, too. According to the chi-square test, the success rate of participants on Level 2 and their group membership were related, $\chi^2(3, N = 65) = 14.04$, $p = .003$, which implies that the success rate (which was influenced by the difficulty of the task) was significantly different in the four groups. Note that this difference results only from participants' performance on Level 2.

An independent colleague analyzed participants' written formulation of the rule. Answers were regarded as correct if they expressed somehow the center-embedded structure of sentences. Most correct answers included the words *symmetrical, mirrored,* or *embedded* or an explicit formula of the sentences (e.g., "abccba" or "123321"). The overlap between success according to the 90% criterion and correctness of the written rule was not perfect: Eight participants who were successful according to the 90% criterion were unable to write down the rule (3 from Group WS, 1 from Group WR, 1 from Group NR1, and 3 from Group NR2). Although it can be a far-reaching question what these participants really learned, the true understanding of the rule by those participants who passed both criteria cannot be questioned. According to the chi-square test, the success rate of participants on the formulation of the rule and their group membership were related, $\chi^2(3, N = 65) = 12.143$, $p = .007$, which enforces the previous finding.

For comparing the number of training blocks needed in the four groups we included the data of unsuccessful participants (i.e., we used 20 blocks as their measure of performance on Level 2 in the analysis). Note that we do not know the accurate number of training blocks they would have needed to reach the criterion on Level 2; the only thing we know is that it would be more than 20. Fortunately, this decision did not affect our statistics (see Footnote 3). Also, we note that the number of training blocks to reach criterion on Level 3 is missing from the analysis for these participants.

The average number of blocks needed to finish all three levels in Group WS was 7.28 (SD = 3.03). Most of the participants needed only two blocks per level (note that this is the least possible according to the training regime), which means that their performance was 90% or above after reading only 10 sentences. Group WR needed 12.27 blocks (SD = 3.788), Group NR1 needed 16.94 blocks (SD = 5.260), and Group NR2 needed 20.25 blocks (SD = 7.646) to finish all levels on average, and the difference was significant between each pair of groups except for Group NR1 and NR2: Kruskal–Wallis test, $\chi^2(3, N = 65) = 38.877$, $p < .001$; Mann–Whitney U test for Groups NR1–NR2, U (N = 32) = 93.5, $p = .196$; in all other cases $p < .01$.

Figure 1 shows the mean number of blocks needed to finish different levels separately in each group. A similar pattern emerges: It seems that the task was easiest for Group WS, more difficult for Group WR and Group NR1, and the most difficult for Group NR2 on all levels. On Level 1 there is significant difference between Group WS and the other groups, but there is no significant difference between Groups WR, NR1, or NR2: Kruskal–Wallis test, $\chi^2(3, N = 65) = 36.674$, $p < .001$; see U and p values from the Mann–Whitney U test for pairwise comparisons in Table 3. This means that learning the word pairs was the easiest when words had a meaning and were semantically related, which is not surprising.

On Level 2 there was no significant difference between Groups WS and WR, Groups WR and NR1, and Groups NR1 and NR2,[3] but the difference was significant between Groups WS and NR1, WS and NR2, and WR and NR2: Kruskal–Wallis test, $\chi^2(3, N = 65) = 17.384$, $p = .001$; for the results of the Mann–Whitney U test, see Table 3. This means that learning the grammar was not facilitated by the semantic relationship between words alone (Vocabulary WS vs. WR), by using words instead of familiar-sounding nonwords (Vocabulary WR vs. NR1), or by the phonetic familiarity of nonwords (Vocabulary NR1 vs. NR2). In other words if two vocabularies were different along one criterion only (see Table 2), it did not make the task of learning CER significantly easier. However, difference along two or three criteria significantly decreased the number of training blocks participants needed to learn the rule.

Pairwise comparison of groups on Level 3—Kruskal–Wallis test, $\chi^2(3, N = 65) = 12.296$, $p = .006$; for the results of the Mann–Whitney U test, see Table 3—yielded similar results as on Level 2. This means that the same factors that helped in recognizing the rule also helped in generalizing and applying it to longer sentences.

An additional analysis was performed to compare the words and nonwords conditions, which divided the participants along one dimension (whether the vocabulary was composed of natural words) into two almost equal groups. The words condition included participants from Groups WS and WR, and the nonwords condition included participants from Groups NR1 and NR2. The difference between the two groups was extremely significant on all levels: Level 1, U (N = 65) = 224.5, $p < .001$; Level 2, U (N = 65) = 266.5, $p < .001$; Level 3, U (N = 57) = 204.0, $p = .001$.

Discussion

The present study investigated the effects of different vocabularies on the speed of learning CER in an AGL task. Sentences composed from these vocabularies represented four different degrees of diversion from natural language according to three factors: familiarity of sounds, meaning of words, and semantic relationship between words (see Tables 1 and 2). We predicted that participants trained with more realistic vocabularies would learn faster than participants trained with vocabularies less similar to natural language.

The most similar to natural language is Vocabulary WS (words semantically paired); however, there are still a lot of differences. For example, in Vocabulary WS both classes of words are nouns, whereas in natural language members of word pairs in center-embedded sentences are from different grammatical categories (e.g., in the sentence “The rat that the cat chased squeaked,” *cat–chased* and *rat–squeaked* form word pairs). Moreover, in natural language, words can have more than one pair from a different class (e.g., *cat–ate* would also be a valid word pair in the above mentioned sentence). Also, sentences composed from Vocabulary WS lack the dependencies between phrases present in natural sentences. On the other hand, these sentences are closer to natural language than those in other experiments in the second generation of AGL studies, because the within-phrase dependencies connecting word pairs are semantic in nature and not the phonological cues used elsewhere.

Stimuli were staged according to the length of the sentences. On Level 1 of training, two-word-long sentences were presented and required associative learning of word pairs. It can be thought of as a simple memory task. Our analysis showed that preexisting semantic relationships between words helped establishing these associations, but none of the other factors present in the vocabularies made a difference.

Level 2 (four-word-long sentences) involved learning or recognizing the center-embedded structure of sentences. The instructions given to participants and the feedback presumably encouraged active rule searching as opposed to passive, incidental learning. As shown by their written formulation of the rule, half of the unsuccessful participants were indeed involved in active rule searching because they mentioned different incorrect rules that they investigated. On this level, there was no significant difference between the learning speed of participants who were trained with vocabularies differing in only one criterion (see Table 2). However, there was significant difference between all other groups, which means that the combined effect of these criteria can help in learning the grammar. The comparison of the words and nonwords conditions, which yielded highly significant differences between these two groups on all levels, supports this hypothesis.

Level 3 tested generalization of the rule to six-word-long sentences. Participants rarely scored under 80% in these blocks, which

[3] These results are not affected by the fact that we used 20 blocks as the measure of performance of unsuccessful participants on Level 2; these differences would not have been significant even if participants had continued their training for more than 20 blocks.

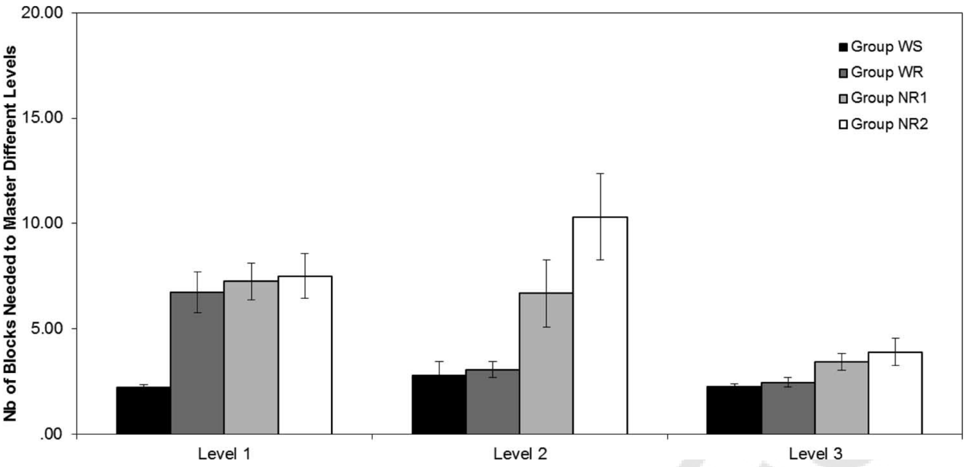

Figure 1. The mean (plus/minus *SE*) number of blocks needed to master Levels 1, 2, and 3 in the four groups of participants. On Level 1 there was significant difference between Group WS and all the other groups. On Level 2 and Level 3 the difference was not significant between Groups WS and WR, Groups WR and NR1, and Groups NR1 and NR2 (those groups whose performance is represented by columns next to each other), but all other pairwise comparisons showed significant differences. WS = words with semantic relatedness; WR = words randomly paired; NR = nonwords randomly paired.

means that generalization was relatively easy. We assume that the differences between groups arose mainly from differences in the difficulty of applying the rule to the sentences. At this level, remembering the first half of the sentence was required for being able to match the words with the second half of the sentence. Analyzing the number of training blocks needed to pass the criterion on this level gained similar results as on Level 2, which means that the same factors that help in recognizing the rule also help in generalizing and applying it.

Table 3
Results of the Mann–Whitney U Test on the Pairwise Analysis of the Performance of Groups on Different Levels of the Task

Group	Level 1	Level 2	Level 3	All levels
WS and WR				
U	6.000	93.000	126.500	17.000
p	**0.000**	0.135	0.762	**0.000**
N	33	33	33	33
WS and NR1				
U	3.000	81.000	62.500	11.000
p	**0.000**	**0.030**	**0.014**	**0.000**
N	34	34	32	34
WS and NR2				
U	6.000	41.500	41.000	9.500
p	**0.000**	**0.000**	**0.018**	**0.000**
N	34	34	28	34
WR and NR1				
U	105.000	88.000	61.500	55.000
p	0.572	0.216	0.057	**0.009**
N	31	31	29	31
WR and NR2				
U	110.500	56.000	39.000	41.000
p	0.711	**0.011**	**0.048**	**0.001**
N	31	31	25	31
NR1 and NR2				
U	122.500	91.500	65.000	93.500
p	0.838	0.171	0.796	0.196
N	32	32	24	32

Note. Significant differences are emphasized by bold numbers. WS = words with semantic relatedness; WR = words randomly paired; NR = nonwords randomly paired.

Vocabulary NR2 was very similar to the vocabulary in Bahlmann et al. (2008). German participants in that study needed 9.47 blocks on average to finish all three levels, and Hungarian participants in our study needed 20.25 blocks. The reason for this difference could be that the vocabulary sounded more familiar to German participants than to Hungarian participants. Some participants in our study reported that they tried to associate nonwords with similar-sounding words and thus giving meaning to nonwords. This strategy to remember the vocabulary is obviously easier when words are phonetically closer to the participants' mother tongue.

With this in mind, we can consider the three factors listed in Table 2 as different forms or levels of semanticity: semantic relationship between words, semantic content of words (real words vs. nonwords), and the ease with which nonwords can be associated with some meaning. This means that semanticity of vocabularies in general influences the speed of learning.

Human participants apparently have difficulties in recognizing CER in AGL tasks: 25% of our participants did not learn the rule after 400 training sentences, when these sentences were composed of nonwords with associative relationship between them. Our experiment is not the only one in which learning was unsuccessful (de Vries et al., 2008; Perruchet & Rey, 2005). This is quite contrary to the theory that CER, as an example of context-free grammar (Corballis, 2007b), is a crucial component of all human languages (Fitch & Hauser, 2004). This contradiction could be explained if there were different mechanisms at work when parsing CER in natural and in artificial languages. It may be that the factors present in natural language but absent from AGL tasks (e.g., the semantic content of sentences and the presence and nature of between and within-phrase dependencies) trigger those mechanisms that are responsible for parsing CER in language. This would mean that it is impossible to test the recursive component of language independently of language itself (or at least some features of language, such as semanticity). Another possibility is that CER is not parsed recursively: Because multiple embeddings are practically absent from natural language, it is indeed not necessary.

In sum, the type of vocabulary does have an effect on the learnability of CER. The more similar the vocabulary is to that of natural language, the easier it is to learn the rule. This makes the comparison of different studies that use different vocabularies and participants with different mother tongues problematic. It also raises the question of whether AGL tasks with artificial vocabularies are suitable for studying the learning and processing of linguistic-center-embedded recursion. A next step in AGL experiments would be to add dependency between phrases, which in turn would make artificial sentences more similar to natural language.

References

Bahlmann, J., Schubotz, R. I., & Friederici, A. D. (2008). Hierarchical artificial grammar processing engages Broca's area. *NeuroImage, 42,* 525–534. doi:10.1016/j.neuroimage.2008.04.249

Conway, C. M., Ellefson, M. R., & Christiansen, M. H. (2003, August). *When less is less and when less is more: Starting small with staged input.* Paper presented at the conference of the Cognitive Science Society, Boston, MA.

Corballis, M. C. (2007a). On phrase structure and brain responses: A comment on Bahlmann, Gunter, and Friederici (2006). *Journal of Cognitive Neuroscience, 19,* 1581–1583. doi:10.1162/jocn.2007.19.10.1581

Corballis, M. C. (2007b). Recursion, language, and starlings. *Cognitive Science, 31,* 697–704. doi:10.1080/15326900701399947

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,* 763–774. doi:10.1016/j.cognition.2007.09.002

Everett, D. L. (2005). Cultural constraints on grammar and cognition in Pirahã: Another look at the design features of human language. *Current Anthropology, 46,* 621–646. doi:10.1086/431525

Fedor, A., & Szathmáry, E. (2011). [Center-embedded recursion in artificial grammar learning tasks: The role of working memory]. Unpublished data.

Fitch, W. T., & Hauser, M. D. (2004, January 16). Computational constraints on syntactic processing in a nonhuman primate. *Science, 303,* 377–380. doi:10.1126/science.1089401

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, USA, 103,* 2458–2463. doi:10.1073/pnas.0509389103

Gentner, T. Q., Fenn, K. M., Margoliash, D., & Nusbaum, H. C. (2006, April 27). Recursive syntactic pattern learning by songbirds. *Nature, 440,* 1204–1207. doi:10.1038/nature04675

Hauser, M. D., Chomsky, N., & Fitch, W. T. (2002, November 22). The faculty of language: What is it, who has it, and how did it evolve? *Science, 298,* 1569–1579. doi:10.1126/science.298.5598.1569

Lai, J., & Poletiek, F. H. (2011). The impact of adjacent-dependencies and staged-input on the learnability of center-embedded hierarchical structures. *Cognition, 118,* 265–273. doi:10.1016/j.cognition.2010.11.011

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

Received July 26, 2010
Revision received November 28, 2011
Accepted November 29, 2011 ■

AUTHOR QUERIES

AUTHOR PLEASE ANSWER ALL QUERIES

AQ1: Author: If this is not the correct affiliation for Mate Varga, please supply the company and/or location.

AQ2: Author: One word in the Vocabulary WS section of Table 1 has three letters. Is this okay?

AQ3: Author: Footnote mentions supporting online material. Should there be online supplemental material for this article, or is a URL needed?