Cognitive Science 32 (2008) 1021–1036

ISSN: 0364-0213 print / 1551-6709 online
DOI: 10.1080/03640210801897849

Do Humans Really Learn A^nB^n Artificial Grammars From Exemplars?

Jean-Rémy Hochmann, Mahan Azadpour, Jacques Mehler

Department of Cognitive Neuroscience, International School of Advanced Studies

Received 5 April 2007; received in revised form 10 December 2007; accepted 11 December 2007

Abstract

An important topic in the evolution of language is the kinds of grammars that can be computed by humans and other animals. Fitch and Hauser (F&H; 2004) approached this question by assessing the ability of different species to learn 2 grammars, $(AB)^n$ and A^nB^n. A^nB^n was taken to indicate a *phrase structure grammar,* eliciting a center-embedded pattern. $(AB)^n$ indicates a grammar whose strings entail only local relations between the categories of constituents. F&H's data suggest that humans, but not tamarin monkeys, learn an A^nB^n grammar, whereas both learn a simpler $(AB)^n$ grammar (Fitch & Hauser, 2004). In their experiments, the A constituents were syllables pronounced by a female voice, whereas the B constituents were syllables pronounced by a male voice. This study proposes that what characterizes the A^nB^n exemplars is the distributional regularities of the syllables pronounced by either a male or a female rather than the underlying, more abstract patterns. This article replicates F&H's data and reports new controls using either categories similar to those in F&H or less salient ones. This article shows that distributional regularities explain the data better than grammar learning. Indeed, when familiarized with A^nB^n exemplars, participants failed to discriminate A^3B^2 and A^2B^3 from A^nB^n items, missing the crucial feature that the number of As must equal the number of Bs. Therefore, contrary to F&H, this study concludes that no syntactic rules implementing embedded nonadjacent dependencies were learned in these experiments. The difference between human linguistic abilities and the putative precursors in monkeys deserves further exploration.

Keywords: Language; Artificial grammar; Evolution; Center embedded; Nonadjacent dependencies

1. Introduction

Research and discussion about the origins of linguistic abilities were very lively until 1866 when all further discussion was banned by the Société de Linguistique de Paris. The subject was revitalized in later years by Pinker and Bloom (1990). In a recent theoretical article,

Correspondence should be sent to Jean-Rémy Hochmann, Department of Cognitive Neuroscience, International School of Advanced Studies, via Beirut 4, I–34014, Trieste, Italy. E-mail: hochmann@sissa.it

Hauser, Chomsky, and Fitch (HC&F; 2002) revisited the issue again. They suggested that it would be useful to distinguish between two conceptions of the language faculty: the language faculty in a broad sense (FLB), which contains parts that we share with other organisms; and the language faculty in a narrow sense (FLN), which is restricted only to language and solely to humans. Whereas FLB involves many components ranging from the sensory-motor system to the organisms' social abilities, the FLN is likely to be limited to one component.

HC&F (2002) argued that each component of FLB may have evolved at different times. In the behavioral sciences, in the absence of fossils, evolutionary arguments are generally warranted by studying what is common or similar in the behaviors of still-existing species. In fact, many features once claimed to be uniquely human have later been found in other species, particularly in nonhuman primates or birds. For instance, categorical perception of phonemes (Kuhl, 1989) were found in birds, and statistical computing of transitional probabilities (TPs) between syllables (Hauser, Newport, & Aslin, 2001) were found in monkeys. The development of songbirds' songs also shows similarities with the acquisition of human language such as the existence of dialects and a critical period for acquisition (Gardner, Naef, & Nottebohm, 2005; Marler & Tamura, 1962; Nottebohm, 1993).

Some features of the language faculty, however, remain good candidates for FLN. Linguists consider that syntax is one of the central components of the language faculty, and much of the current theoretical research continues to highlight its uniqueness. Although members of some nonhuman species combine different calls or gestures to communicate with their peers, none of them can create an infinite number of meaningful utterances, by rearranging the basic units of their language in a lawful manner. HC&F (2002) framed their article as a quest for the crucial evolutionary step that allowed our species to acquire the complex syntax of human languages. They proposed that recursion yielding discrete infinity might be the only component of FLN. Some experiments conducted by Fitch and Hauser (F&H; 2004) are often presented as support for the recursion-only hypothesis; so much so that other researchers have taken these data as the basis of their experimental work and theoretical hypothesis, conducting brain imaging studies (Bahlmann, Gunter, & Friederici, 2006; Friederici, Bahlmann, Helm, Schubotz, & Anwander, 2006), or extending the animal database to avian species (Gentner, Fenn, Margoliash, & Nusbaum, 2006).

Theoretically, the validity of the HC&F (2002) hypothesis is still debated. Particularly, Pinker and Jackendoff (2005) argued that it underestimates the importance and uniqueness of non-syntactic or non-recursive grammatical features, especially morphology and phonology (see also Fitch, Hauser, & Chomsky, 2005; Jackendoff & Pinker, 2005). However, in this article, our purpose is not to intervene in that debate. Our experiments do not aim to evaluate the recursion-only hypothesis. Rather, we explore its putative experimental support, as offered in F&H (2004). We first replicated their data with human participants and then conducted new experiments, showing that F&H's empirical findings are not as persuasive as the authors claim. Nevertheless, we agree with F&H that exploring evolutionary scenarios by comparing different species using artificial grammars is a helpful approach.

In recent years, many researchers successfully used artificial grammar to explore language acquisition in young infants and processing in adults. Saffran, Aslin, and Newport (1996) demonstrated the power of statistical computations in 8-month-olds using an artificial speech stream. Using a different speech stream, Peña, Bonatti, Nespor, and Mehler (2002) studied the

conditions under which nonadjacent dependencies yield structural generalizations in adults. Marcus, Vijayan, Rao, and Vishton (1999) studied the manipulation of symbols in 7-month-olds, using an artificial grammar consisting of items having repetitions like those in an ABB or ABA grammar (where As and Bs are syllables). Endress, Scholl, and Mehler (2005) further studied the acquisition of similar rules in adults and the influence of the position of the repetition. Furthermore, artificial grammars have been used to study morphological categories (Endress & Bonatti, 2007; Mintz, 2002), whereas Gomez and Lakusta (2004) used such grammars to explore the conditions under which 1-year-old infants can extract structural regularities. Similarly, Chambers, Onishi, and Fisher (2003) studied the acquisition of phonotactic regularities in 16.5-month-olds. All these studies using artificial grammar learning methods have provided the scientific community with new insights on the mechanisms that sustain the human language faculty. Therefore, it was a wise step to bring this technique to the study of language evolution.

Fifty years ago, Chomsky (1963) elaborated a hierarchy of formal grammars on the basis of their complexity. In this classification, the position of grammars describing natural languages remains discussed (see Kornai, 1985; Pullum & Gazdar, 1982; Shieber, 1985). Nevertheless, undeniably, some properties of natural languages require a grammar of a certain complexity. Particularly, the description of human languages must account for relations among nonadjacent constituents. For example, in "*the cat* that the dog chased *hides,*" the subject of "hides" is "the cat" and not the nearer element "the dog." Such long-range dependencies, if their number is not limited, require recursive or centrally embedded structures, where one sentence is embedded inside another one. Importantly, the dependency between "the cat" and "hides" is not only a relation between those specific words, but also a relation between the syntactic categories those items belong to. The relation would remain syntactically (although not semantically) correct for a subclass of nouns replacing "the cat" and a subclass of verbs replacing "hides" with the correct agreement and morphology. Although these sentences are rarely produced, the number of hierarchical levels is theoretically infinite (i.e., the sentence, "*the cat* the dog **the mouse loved** chased *hides,*" is grammatically correct with two levels of embedding). However, performance factors severely limit the number of embeddings with which language users can cope.

According to Chomsky's (1963) classification, context-free grammars are more complex than finite-state grammars. If A and B are two categories of constituents, the language $A^n B^n$ (containing items with structures *AB, AABB, AAABBB* ...) was claimed to necessitate a context-free grammar[1] to be described (Fitch & Hauser, 2004; but see Perruchet & Rey, 2005), whereas the language $(AB)^n$ (containing items with structures *AB, ABAB, ABABAB*) can be described by a finite–state grammar.[2] Both languages can be generated with the same vocabulary, with only the pattern of word ordering differing from one language to the other. Whereas learning the $(AB)^n$ structure only requires learning local relations, learning the $A^n B^n$ structure requires building long-range dependencies between As and Bs or counting the number of constituents in each category. Therefore, to master the $A^n B^n$ language, a learner needs to build a hierarchical structure or invoke a memory mechanism for storing the syllable counts. In the case where learners build nonadjacent relations between As and Bs, the process mirrors the subject–verb relations of natural sentences with center-embedded structures. Here again, the relations or dependencies are understood to hold between categories rather than specific items.

In order to test the evolutionary hypothesis about the uniqueness of recursion to human language, F&H (2004) habituated humans and a related species, cotton-top tamarin monkeys, to one or the other language using items potentially generated by one or the other grammar. The As and Bs were consonant–vowel syllables. More important, *A*s were pronounced by a female voice, whereas *B*s were pronounced by a male voice. Humans were exposed to 60 items generated by the grammar of habituation—$(AB)^n$ or $A^n B^n$, depending on the group—and subsequently judged a series of items congruent or incongruent with the grammar of habituation. Monkeys were exposed to 60 grammatical items repeated during 20 min one day prior to testing. Following a re-familiarization of 2 min, they were tested in a head-turning procedure. The results suggest that humans behaved as if they had learned both grammars, whereas monkeys behaved as if they had learned only the simpler $(AB)^n$ grammar, implying that only humans can master the higher grammatical complexity.

A recent and elegant experiment challenged the conclusion that only humans can learn recursive structures. Indeed, some European starlings, a songbird species more distant from humans than monkeys, can learn to discriminate in a go/no-go procedure the $(AB)^n$ and $A^n B^n$ patterns, after thousands of reinforced trials (Gentner et al., 2006). Here, As and Bs were two specific categories of sounds produced by these birds. The methodology differed from the brief non-reinforced exposure used by F&H (2004) to test monkeys and humans. Nevertheless, beyond these differences, this pattern of results across species is striking. The ease with which humans pass F&H's tests, and the ability of songbirds to learn both patterns, contrasts with the failure of monkeys on $A^n B^n$.

From a different perspective, Perruchet and Rey (2005) explored F&H's (2004) conclusions in an experiment that is a modified version of F&H's. The authors argued that human participants in F&H did not need to build nonadjacent dependencies and, therefore, did not establish center-embedded structures while learning the $A^n B^n$ language. They habituated human participants to exemplars of an $A^n B^n$ grammar where, as in F&H, As were high-pitch syllables and Bs were low-pitch syllables. The acoustic pattern was, therefore, *n* syllables pronounced with a high pitch, followed by *n* syllables pronounced with a low pitch. Unlike F&H, embedded dependencies between specific syllable pairs were implemented as follows: If A_1 appeared in the first position, B_1 appeared in the last position; if A_2 appeared in the second position, B_2 appeared in the penultimate position, and so forth. Test items subsequently respected or violated the acoustic pattern, the dependencies between specific syllables, or both. The results suggested that participants were guided by the acoustic pattern and ignored the dependencies between specific syllables.

Although Perruchet and Rey (2005) raised some interesting issues, there are several other points that they do not mention. For instance, because the contrast between the high pitch and low pitch is so salient, the dependencies between syllables and the center-embedded pattern might have been unnoticed. We suspect that syllable categories may be the crucial element that guided participants' behavior during the test phase.

The ability of humans to learn the more complex grammar should be re-evaluated before one supports the evolutionary conjectures. In the original F&H (2004) experiment, participants did learn "something" because they were able to discriminate between $A^n B^n$ and $(AB)^n$ items. Our purpose is not to find out under what conditions an $A^n B^n$ grammar can be learned. Rather, our aim is to explore what human participants actually learned in F&H's experiment. We begin

by assessing whether the $A^n B^n$ grammar is learned using the same experimental arrangement as the one in F&H. As in F&H, and in contrast with Perruchet and Rey (2005), we do not add supplementary cues to signal dependencies between specific pairs of syllables. In Experiment 2, we evaluate if such a replication necessarily entails the extraction of an $A^n B^n$ grammar. Finally, Experiments 3 and 4 assess the influence of the category saliency, suppressing the female–male contrast, and assigning a different syllabic structure to each category.

2. Experiment 1

In Experiment 1, we attempt to replicate the results reported in F&H (2004). The description of the material and methods given in F&H, and the associated supplementary material, were our guidelines for designing our experiment.

2.1. Method

2.1.1. Participants and procedures

Thirty-two Italian students participated in this experiment. One half were familiarized with the $(AB)^n$ grammar, the other half were familiarized with the $A^n B^n$ grammar.

Each participant was asked to listen to 64 items (4 min) generated by their grammar of habituation. Consecutive items were separated by 1 sec of silence. In this phase, as well as in the test phase, one half of the items were four syllables long, and one half were six syllables long (e.g., ABAB or ABABAB for the $(AB)^n$ grammar and AABB or AAABBB for the $A^n B^n$ grammar). During the test phase, participants had to judge 32 novel items by pressing one of two buttons: 16 were novel items generated by the grammar they were habituated to (congruent items); 8 were items generated by the other grammar; and 8 were constructed according to the grammar they were habituated to, inverting A and B syllables (structure BABA or BABABA when habituated to the $(AB)^n$ grammar; BBAA or BBBAAA when habituated to the $A^n B^n$ grammar). Therefore, this experiment is a replication of the second series of tests in F&H's (2004, supplementary material) original experiment.

2.1.2. Stimuli

The syllables *bo, fe, ge, ku, pu, ri, vo,* and *zi* were employed for Category A; and the syllables *be, fi, gu, ko, pi, ro, vu,* and *ze* were employed for Category B. Each syllable was generated using the Mbrola Synthesizer Version 3.02b. The German female voice database, de5, was used for the syllables of Category A with a constant pitch of 250 Hz. The German male voice database, de6, was used for the syllables of Category B with a constant pitch of 100 Hz. All syllables were made up of a 200 msec long consonant and a 300 msec long vowel.

Ninety-six grammatical strings were generated by each grammar, one half of which were four syllables long, and one half were six syllables long. Each syllable appeared equally often in each of the positions allowed by its category. The corresponding sound stimuli

were generated with the PRAAT sound synthesizer Version 4.3.30. The inverted items were generated in a similar manner.

2.2. Results

We computed A′ coefficients, a non-parametric estimate of discriminability (Green & Swets, 1966), to evaluate the ability of participants to discriminate incongruent items from items congruent with their grammar of habituation. An A′ equal to 0.5 denotes chance performance; an A′ equal to 1 denotes perfect discrimination.

We conducted a 2×2 analysis of variance (ANOVA) with grammar—$A^n B^n$ versus $(AB)^n$—and incongruent type of tests (other grammar vs. inverted items) as factors. We found a main effect of incongruent type of tests, $F(1, 60) = 7.31$, $p < .01$. Neither the effect of the grammar factor nor the interaction was significant.

A *post-hoc* paired t test revealed that participants were better at rejecting the items generated by the other grammar than the inverted items produced by their grammar of habituation, $t(31) = 5.76$, $p < .00001$.

However, all discriminations were significantly above chance. In the $(AB)^n$ group, participants accepted $(AB)^n$ and rejected $A^n B^n$ items (mean $A' = 0.86$), $t(15) = 10.02$, $p < .0000001$. They also rejected significantly the structure $(BA)^n$ (mean $A' = 0.71$), $t(15) = 4.7$, $p = .001$. In the$A^n B^n$ group, participants accepted $A^n B^n$ and rejected $(AB)^n$ items (mean $A' = 0.77$), $t(15) = 4.85$, $p < .001$. They also rejected significantly the structure $B^n A^n$ (mean $A' = 0.66$), $t(15) = 2.91$, $p = .011$.

2.3. Discussion

Our results support the notion that human participants were able to correctly identify items congruent with the habituation grammar and to reject incongruent ones, regardless of whether they were habituated to strings generated by the $(AB)^n$ or the $A^n B^n$ grammar.

In the F&H (2004) study, as well as in our experiment, the discrimination of $A^n B^n$ and $(AB)^n$ strings was the crucial measure used to evaluate learning. However, successful discrimination does not license claims about grammar learning. Indeed, F&H used syllables pronounced by females (As) and males (Bs) to generate items. Female and male voices differ in many features. In particular, their pitch differs by more than one octave. Thus, the transition between constituents of different categories becomes very salient, which may help participants compute TPs between categories. Computing the TPs between categories[3] or focusing on the female–male alternation would thus suffice to discriminate between test items congruent and incongruent with the grammar of familiarization. Indeed, in the case of the items generated by the $(AB)^n$ grammar, an A syllable is always followed by a B syllable, and a B syllable is never followed by another B syllable. Therefore, the sequences of two or three syllables from the same category present in $A^n B^n$ items are unexpected and should be easily rejected. On the contrary, participants familiarized with $A^n B^n$ items may have learned that correct items are made up of two series of syllables from the same category. The $(AB)^n$ items not following that pattern can be rejected.

Indeed, in rejecting the inverted items B^nA^n, the performance of participants habituated to strings generated by the A^nB^n grammar is lower than the performance in rejecting $(AB)^n$ items. This could mean that participants mainly focus on the alternation between the two categories of syllables, and might not learn abstract grammar. In other words, in the $(AB)^n$ group, they know that familiarization items were made of an alternation of A syllables pronounced by a female voice and B syllables pronounced by a male voice; whereas in the A^nB^n group, they learned that familiarization items were made of two parts: one uninterrupted series of syllables pronounced by a female voice and one uninterrupted series of syllables pronounced by a male voice. In what order those two parts appear is not crucial to them.

In the next experiment, we evaluate, using a more stringent test, whether participants are capable of extracting abstract regularities from items like the ones F&H (2004) used.

3. Experiment 2

In this experiment, we asked if distributional properties, such as TPs between categories or the rate of the female–male alternation, suffice to explain the results in F&H (2004) and Experiment 1. We exposed a group of human participants to A^nB^n strings ($n = 2$ or 3) using female–male categories. As in Experiment 1, participants subsequently evaluated whether new strings of the $(AB)^n$ or A^nB^n languages belonged to the language of familiarization. Importantly, participants who learned the habituation grammar should only accept items with an equal number of As and Bs. Thus, participants were also tested with "ungrammatical" items A^2B^3 and A^3B^2 (hereafter, "odd items"), which are distributionally similar to A^nB^n strings, but whose numbers of As and Bs differ. Participants focusing on TPs between categories or voice alternation should accept such items, but not participants who learned the grammatical structure.

To provide our participants with more information and give them a better chance to learn the abstract grammar, we doubled the length of the familiarization phase compared to Experiment 1.

3.1. Method

3.1.1. Participants and procedures

Sixteen Italian students were exposed to 128 items (8 min) generated by the A^nB^n grammar. One half of the items were four syllables long ($n = 2$), and the other half were six syllables long ($n = 3$). Two consecutive items were separated by 1 sec of silence.

Participants were subsequently asked to judge 36 novel items as congruent or incongruent with the language they had been exposed to, by pressing one of two buttons. Twelve test items were generated by the A^nB^n grammar, and another 12 test items were generated by the simpler grammar $(AB)^n$. One half were four syllables long, and one half were six syllables long. Twelve test items were constructed with the structures A^2B^3 (6 items) or A^3B^2 (6 items) in such a way that the TPs between syllabic categories A and B remain the same as for items generated by the A^nB^n grammar. Test items were presented in random order.

3.1.2. Stimuli

The same syllables as in Experiment 1 were used. As in Experiment 1, we used the Mbrola Synthesizer Version 3.02b and the PRAAT sound synthesizer Version 4.3.30, to generate the stimuli.

One hundred forty grammatical strings were generated by the A^nB^n grammar, and 12 items were generated by the $(AB)^n$ grammar. The A^2B^3 and A^3B^2 items were generated by randomly selecting two or three A syllables and two or three B syllables. For each type of item, we counterbalanced the appearance of different syllables in each item position. We also ensured that none of the A^2B^3 and A^3B^2 items were contained in any of the grammatical items otherwise used.

3.2. Results

Participants accepted A^nB^n and rejected $(AB)^n$ items (mean $A' = 0.84$), $t(15) = 8.08$, $p < .000001$, replicating once more F&H (2004). In discriminating A^nB^n from the odd items, performance was bimodally distributed (Lilliefors test; Lilliefors, 1967), $p < .04$. Five participants rejected the odd items (mean $A' = 0.97$). They explicitly reported counting the number of As and Bs. The remaining 11 participants accepted the odd items as correct at the same rate as A^nB^n items (mean $A' = 0.50$), $t(10) = 0.06$, $p > .95$.

3.3. Discussion

The results of this experiment confirm that participants exposed to items congruent with the A^nB^n grammar reject items produced by the $(AB)^n$ grammar. However, the majority of participants were unable to reject odd items as ungrammatical. This shows that they did not learn the abstract grammar, missing the crucial feature that to each A syllable must correspond a B syllable; and, therefore, the number of A syllables must equal the number of B syllables.

Moreover, the few participants who successfully rejected the odd items reported explicitly counting the number of As and Bs. This strategy is unsuitable for acquiring language, as related items (words or phonemes) can routinely be separated by an arbitrary number of items (words or phonemes). Thus, this shows that even those participants did not build any putative long-distance dependencies between A and B constituents.

Therefore, most participants did not learn the underlying abstract grammar for A^nB^n items. Rather, they focused on the single alternation of categories. Even those who actually learned a rule to construct A^nB^n items did not use nonadjacent dependencies.

In sum, by focusing participants' attention on the A to B transition, the salient female–male categories may have facilitated the discrimination of $(AB)^n$ and A^nB^n strings, while impairing grammar extraction. To assess the influence of categorical saliency on the task of learning our two grammars, we conducted two additional experiments similar to the preceding ones, except that the *same voice* pronounced all syllables, reducing the phonological contrast between categories. Here, the A–B contrast was implemented using syllables with different structures. *A*s were consonant–vowel, and *B*s were consonant–vowel–consonant syllables. In Experiment 3, participants exposed to $(AB)^n$ items had to judge novel $(AB)^n$ and A^nB^n items.

In Experiment 4, as in Experiment 2, participants exposed to A^nB^n items had to judge novel A^nB^n and $(AB)^n$ items, as well as odd items (structures A^2B^3 and A^3B^2).

4. Experiment 3

4.1. Method

4.1.1. Participants and procedures

Thirty-two Italian students were familiarized with items produced by the $(AB)^n$ grammar. They were divided into two groups that differ only in the length of familiarization. In the 4-min familiarization group, participants listened to 64 items. In the 6-min familiarization group, they listened to 96 items. In both cases, one half of the items were four syllables long ($n = 2$), and one half were six syllables long ($n = 3$). Consecutive items were separated by 1 sec of silence.

In both groups, participants were subsequently asked to judge 32 novel items as congruent or incongruent with the language they had been exposed to, by pressing one of two buttons. Sixteen items were generated by the $(AB)^n$ grammar, and 16 were generated by the A^nB^n grammar. One half were four syllables long, and one half were six syllables long. Test items were presented in a random order.

4.1.2. Stimuli

We used the syllables *ta, ba, lo, do, nu, mu, gi,* and *ki* for Category A; the syllables *bod, dop, mit, lik, gat, pak, nup,* and *pud* were used for Category B. One hundred twelve items were generated by the $(AB)^n$ grammar and 16 by the A^nB^n grammar. We counterbalanced the appearance of different syllables in each item position. The stimuli were generated using a Mbrola Synthesizer Version 3.02b. The female German voice database, de5, was used. Every syllable lasted 500 msec with a pitch of 200Hz. In Category A, consonants lasted 200 msec and vowels 300 msec. In Category B, consonants lasted 200 msec and vowels 100 msec. Syllables were separated by a silence of 50 msec.

4.2. Results

Participants accepted $(AB)^n$ items and rejected A^nB^n items both in the 4-min familiarization group (mean $A' = 0.65$), $t(15) = 5.16$, $p < .001$; and in the 6-min familiarization group (mean $A' = 0.79$), $t(15) = 7.45$, $p < .00001$.

A one-way ANOVA further showed that performance in the 6-min familiarization group was significantly better than in the 4-min familiarization group, $F(1, 30) = 8.25$, $p < .01$.

Comparing the performance of participants in the 4-min group with participants habituated to $(AB)^n$ items in Experiment 1, a one-way ANOVA showed that performance was significantly better for participants habituated to the female–male voice contrast than for those habituated to the syllabic structure contrast, $F(1, 30) = 20.67$, $p < .0001$.

A one-way ANOVA showed that participants in the 6-min group and those habituated to $(AB)^n$ items in Experiment 1 did not differ significantly, $F(1, 30) = 1.78$, $p > .19$.

4.3. Discussion

In this experiment, we tested the influence of categorical saliency by replacing the female–male voice contrast by a contrast in the structure of syllables. Although the performance was lower than in Experiment 1, we showed that this less salient contrast did not prevent participants from learning the simple $(AB)^n$ grammar. The two categories were therefore identifiable in our stimuli.

Moreover, we showed that increasing the length of the familiarization from 4 to 6 min significantly increased the performance of participants. Participants familiarized longer reached a performance similar to that of participants in Experiment 1, familiarized with the female–male voice contrast. This confirms the influence of the categorical contrast in Experiment 1, having the same effect as a 50% increase in familiarization.

5. Experiment 4

In this experiment, as in Experiment 3, the *same* voice pronounced all syllables. *A*s and *B*s were consonant–vowel and consonant–vowel–consonant syllables, respectively. As in Experiment 2, participants exposed to A^nB^n items had to judge novel A^nB^n and $(AB)^n$ items, as well as odd items. To provide our participants with more information and give them a better chance to learn the abstract grammar, the familiarization phase lasted 8 min, as in Experiment 2. The results of the preceding experiment show that this is sufficient to learn the categories of syllables.

5.1. Methods

5.1.1. Participants and procedures

Sixteen Italian students were exposed to items produced by the A^nB^n grammar. The procedure for these participants was similar to Experiment 2, except for the stimuli used.

5.1.2. Stimuli

The same syllables, the same voice, and the same parameters as in Experiment 3 were used. One hundred forty items were generated by the A^nB^n grammar and 12 by the $(AB)^n$ grammar. Six A^2B^3 and six A^3B^2 items were generated by randomly selecting two or three A syllables and two or three B syllables. We counterbalanced the appearance of the different syllables in each item position.

5.2. Results

Participants' performance differed significantly from chance in discriminating A^nB^n from $(AB)^n$ strings (mean $A' = 0.66$), $t(15) = 2.71$, $p = .016$. However, this performance was

significantly lower than that of Experiment 2, with the female–male contrast, $t(30) = 2.43$, $p = .021$. In discriminating A^nB^n strings from odd items, performance was unimodally distributed (assessed by a Lilliefors test; Lilliefors, 1967) and at chance (mean $A' = 0.53$), $t(15) = 0.51$, $p = .62$.

5.3. Discussion

In Experiments 1 and 2, the perceptual contrast between the female and male voices permitted participants to judge items without taking into account the identity of the syllables. This may have impaired grammar extraction, as relying on one acoustical feature (i.e., pitch) was seemingly sufficient for participants to do their task.

In Experiment 4, the influence of categorical saliency was tested by replacing the female–male voice contrast with a contrast in syllable structure. The participants exposed to items produced by the A^nB^n grammar with this contrast were still unable to learn the abstract grammar. Their performance was even impaired compared to participants of Experiment 2 in rejecting $(AB)^n$ items, and they were unable to reject odd items.

When the categorical saliency was reduced, grammar extraction was still not elicited, and performance was impaired. This supports the notion that TPs between categories or rhythmic regularities, rather than grammar extraction, guided behavior in these experiments.

6. General discussion

In this study, we assessed if the grammar underlying A^nB^n patterns can be learned from a short passive exposure to exemplars. Our results suggest that under these conditions, adult participants do not extract the grammar. Although Experiment 1 replicated the findings of F&H (2004), Experiment 2 suggests that distributional regularities such as the TPs between As and Bs or rhythmic patterns are sufficient to account for the observed results. It could be that the presence of highly salient A and B categories in Experiments 1 and 2 disrupted grammar extraction by monopolizing the attention of participants. However, in Experiments 3 and 4, when the categories were made less salient, the performance of our participants was significantly impaired compared to that in Experiments 1 and 2. These results add further support to the view that no abstract grammar was learned in these experiments.

If our results, as well as those of F&H (2004), involved only the processing of surface features, we can wonder why monkeys do not perform as well as humans. As a matter of fact, Hauser et al. (2001) showed that tamarin monkeys are able to process distributional features, such as TPs between syllables. In our study, the results of Experiment 1 show that for humans distributional regularities are easier to learn from the $(AB)^n$ items than from the A^nB^n ones. This may explain why tamarin monkeys succeeded in learning the $(AB)^n$ pattern but failed with the A^nB^n, as reported by F&H. Indeed, a possible scenario is that when exposed to $(AB)^n$ items, monkeys habituated to the regular alternation of the A and B constituents, which was made salient by the use of the female and male voices. During the test phase, the succession of two syllables pronounced by the same voice as in A^nB^n tokens was unexpected and may have attracted the monkeys' attention, resulting in a novelty preference.

When habituated to A^nB^n items, the monkeys encountered transitions between one voice and the other, as well as the succession of syllables pronounced by the same voice. Thus, none of the test stimuli were particularly new, and monkeys showed no preference for any kind of test stimuli.

Chomsky's (1963) aim when using the A^nB^n grammar was to highlight the human ability to use recursivity, a crucial feature of natural languages. Since this original proposal, questions have arisen as to whether recursivity is specific to humans (and to language). The results of F&H (2004) lent credibility to the uniquely human hypothesis. However, our results suggest that the procedure used in F&H does not induce participants to spontaneously search for abstract regularities. Participants' analysis of the exemplars provided in the familiarization phase seems limited to surface features. Both humans and monkeys characterized, without difficulty, the finite-state grammar pattern $(AB)^n$, using the regularity of A to B and B to A transitions. In contrast, most humans, like the tamarin monkeys, failed to correctly characterize the context-free grammar pattern A^nB^n. Under certain conditions, participants may learn to count the number of elements from each category of components and respond accordingly. The few participants in Experiment 2 who correctly rejected all ungrammatical items reported comparing the number of consecutive elements from Category A and Category B. No participants in our experiments mentioned other structural properties as the motivation of their responses.

The data provided for songbirds (Gentner et al., 2006) are, however, more controversial. Trained with A^2B^2 and $(AB)^2$ items, songbirds were able to generalize their behavior to A^3B^3, $(AB)^3$, A^4B^4, and $(AB)^4$ items. Furthermore, the authors controlled for a series of alternative explanations to their results. Particularly, starlings successfully discriminated items similar to our "odd items" (A^3B^2 but not A^2B^3) from A^3B^3 items. The experimental procedures used with humans on the one hand, and with birds on the other, are quite different. Particularly, the go/no-go procedure may allow organisms to pick up unsuspected cues that correlate with one of the patterns. Nevertheless, it remains quite intriguing that songbirds display a behavior resembling the one F&H (2004) predicted for humans Although the only humans who performed at a ceiling level used a counting strategy, the question remains open whether starlings used a similar strategy or actually learned the center-embedded structure. Further research should clarify this issue.

Using extended familiarization or providing feedback may also improve the performance of monkey and human participants. Indeed, in a fMRI study, Friederici et al. (2006) used longer familiarization and feedback to teach an $(AB)^n$ or an A^nB^n grammar. Their participants achieved better performances than ours. In this experiment, all stimuli were visually presented. All item constituents were consonant–vowel syllables. The vowel was *i* or *e* for As and *o* or *u* for Bs. The mean percentage of correct answers was very high for both grammars. It was found that processing items of both $(AB)^n$ and A^nB^n grammars activated the left frontal operculum. Broca's area, which is phylogenetically a more recent area than the left frontal operculum, was only activated by items generated by the A^nB^n grammars. The authors proposed that Broca's area is recruited for processing the hierarchical structure necessary to describe A^nB^n items, whereas the frontal operculum is devoted to the processing of TPs.

However, the implementations of the A^nB^n grammar used by Friederici et al. (2006), by F&H (2004), and in our experiments do not contain cues for building relations between specific

As and Bs. Therefore, longer exposure to exemplars of correct and incorrect items, as well as feedback on their judgments, tends to lead participants toward a simpler characterization of the two patterns (i.e., the TPs in 1 case and to count the number of items of each category in the other case) rather than biasing participants to extract the putative underlying grammar. Therefore, the involvement of Broca's area in the participants in Friederici et al.'s experiment is unlikely to be related to the construction of hierarchical structures. Rather, the success of some participants in these experiments is likely to be based on a counting strategy. Although some linguistic rules may be described in terms of counting constituents, there is no known linguistic-like rule that relies on counting syllables or words in utterances. Such illegitimate counting rules and natural linguistic rules are actually processed by different neural systems in humans (Musso et al., 2003). Counting words or syllables is therefore not a suitable strategy for learning linguistic rules.

In our results, as well as in the other studies discussed here, we found no evidence that embedded nonadjacent dependencies between categories of syllables are spontaneously built. However, this does not contradict the hypothesis that humans differ from other animal species in their abilities to learn such structures. Humans are able to build complex structures containing nonadjacent dependencies, as is demonstrated by their ability to learn natural languages. In natural conditions, the sources of information that can highlight such dependencies range from prosodic to semantic cues. The use of artificial grammars provides a way to study some abilities or processes used in language acquisition with a controlled paradigm. Some investigations have attempted to isolate syntax from the multiple other properties that correlate with it. However, we suspect that under natural conditions, syntax would be impossible to learn if the speech signal were deprived of sentence prosody and possibly meaning. Syntactic cues may also be missing here and explain participants' failure to discover the putative center-embedded organization in the sentences. Indeed, the $A^n B^n$ structure mirrors only the organization of the subject–verb relations in natural sentences. It does not take into account other syntactic relations. Particularly, the verb–object relations may be crucial to explain the center-embedded pattern (Aravind Joshi, personal communication July 26, 2007). This in no way means that artificial grammar studies should be abandoned, given the important insights provided by this technique as documented in our introduction.

Statistical computations are ubiquitous and so is the detection of perceptual primitives, such as repetitions and edges. This has been demonstrated in numerous artificial grammar experiments (Endress et al., 2005; Saffran et al., 1996). In some artificial grammar experiments, participants established categories on the basis of edges (Endress & Bonatti, 2007) or exhibit behaviors that seem to rely on rules (Gomez, Gerken, & Schvaneveldt, 2000; Marcus et al., 1999; Peña et al., 2002). However, Shukla, Nespor, and Mehler (2007) studied how prosody interacts with the extraction of statistical information in an artificial speech stream. They found that words statistically highlighted are nevertheless filtered out if they straddle a prosodic boundary. This shows the importance of continuously trying to confront artificial grammar results to more realistic conditions like the ones infants encounter.

When exposed to more complex stimuli, human adults may engage in reasoning and build up rules counting the number of syllables or explicitly noting some configurations. However, once more, none of these strategies enters the natural process of language learning. Human babies, when learning their mother tongue, are unable to count and unlikely to possess metalinguistic

knowledge. For that reason, the processes at play in adults and pre-linguistic babies when learning both artificial and natural grammars may be quite different.

In summary, the seminal articles by Hauser et al. (2002) and F&H (2004) bolstered the study of language evolution, which had remained a taboo for too long. However, we showed in the present study that the recursion-only hypothesis is still not demonstrated. Indeed, the artificial grammar studies discussed above yield results that may arise from distributional proprieties such as TPs or alternation patterns. The strong correlation between deep structures and surface features renders the extraction of an abstract grammar unnecessary for participants. Observing a phenomenon of grammar extraction requires that explanations based on distributional analysis of surface features be excluded. Moreover, the process of implicit grammar learning in humans is still not understood. Passive exposure to examples may not be enough to trigger learning of complex structures. Our field may gain from more research on what mechanisms are operating in such conditions, especially in pre-linguistic babies, and how these can lead to the projection of generalization. Differences between humans and monkeys are certainly expected, but their nature remains to be experimentally uncovered. Combining comparative studies with the investigation of human competence should promote a better understanding of language emergence in humans.

Notes

1. Note that in Fitch and Hauser (2004), the grammar describing the language $A^n B^n$ is termed as *phrase structure grammar.* In this article, we used both terms. Beyond the terminology, the important point is to notice that a finite-state grammar does not suffice to describe the language $A^n B^n$. A higher level of complexity is necessary.
2. Note that to run their experiments, Fitch and Hauser (2004) used only a limited number of patterns generated by each grammar. From contrasting difficulties that monkeys and humans may show when familiarized with these patterns, they hoped to be able to draw inferences about the abilities of the species to learn these two grammars.
3. Note that whereas many studies refer to transitional probabilities (TPs) between *syllables* as the conditional probability of syllable y following syllable x (Aslin, Saffran, & Newport, 1998; Christophe, Dupoux, Bertoncini, & Mehler, 1994; Saffran, Aslin, & Newport, 1996), in this study we consider TPs between *categories of syllables*—that is, the conditional probability of a syllable in Category Y following a syllable in Category X.

Acknowledgment

This research has been supported by McDonnell Foundation Grant No. 21002089 and the European Commission Special Targeted Project CALACEI (Contract No. 12778 NEST).

References

Aslin, R. N., Saffran, J. R., & Newport, E. L. (1998). Computation of conditional probability statistics by 8-month-old infants. *Psychological Science, 27,* 321–324.

Bahlmann, J., Gunter, T. C., & Friederici, A. D. (2006). Hierarchical and linear sequence processing: An electrophisiological exploration of two different grammar types. *Journal of Cognitive Neuroscience, 18,* 1829–1842.

Chambers, K. E., Onishi, K. H., & Fisher, C. (2003). Infants learn phonotactic regularities from brief auditory experience. *Cognition, 87,* B69–B77.

Chomsky, N. (1963). Formal properties of grammars. In D. Luce, R. Bush, & E. Galanter (Eds.), *Handbook of Mathematical Psychology, Vol. II* (pp. 323–418). New York: Wiley & Sons.

Christophe, A., Dupoux, E., Bertoncini, J., & Mehler, J. (1994). Do infants perceive word boundaries? An empirical approach to the bootstrapping problem for lexical acquisition. *Journal of the Acoustical Society of America, 95,* 1570–1580.

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

Endress, A. D., Scholl, B. J., & Mehler, J. (2005). The role of salience in the extraction of algebraic rules. *Journal of Experimental Psychology: General, 134*, 406–419.

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

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

Friederici, A. D., Bahlmann, J., Helm, 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 of the United States of America, 103*, 2458–2463.

Gardner, T. J., Naef, F., & Nottebohm, F. (2005). Acquisition and reprogramming of a bird's learned song. *Science, 308,* 1046–1049.

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

Gomez, R. L., Gerken, L. A., & Schvaneveldt, R. W. (2000). The basis of transfer in artificial grammar learning. *Memory & Cognition, 28,* 253–263.

Gomez, R. L., & Lakusta, L. (2004). A first step in form-based category abstraction by 12-month-old infants. *Developmental Science, 7,* 567–580.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: Wiley.

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.

Hauser, M. D., Newport, E. L., & Aslin, R. N. (2001). Segmentation of the speech stream in a non-human primate: Statistical learning in cotton-top tamarins. *Cognition, 78,* B53–B64.

Jackendoff, R., & Pinker, S. (2005). The nature of the language faculty and its implications for evolution of language (reply to Fitch, Hauser, and Chomsky). *Cognition, 97,* 211–225.

Kornai, A. (1985). Natural languages and the Chomsky hierarchy. In M. King (Ed.), *Proceedings of the 2nd conference of the European chapter of the Association for Computational Linguistics* (pp. 1–7), March 27–29, Geneva, Switzerland.

Kuhl P. K. (1989). On babies, birds, modules, and mechanisms: A comparative approach to the acquisition of vocal communication. In R. J. Dooling & S. H. Hulse (Eds.), *The comparative psychology of audition* (pp. 379–422). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Lilliefors, H. (1967). On the Kolmogorov–Smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association, 62,* 399–402.

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

Marler, P., & Tamura M. (1962). Song "dialects" in three populations of white-crowned sparrow. *The Condor, 64,* 368–377.

Mintz, T. H. (2002). Category induction from distributional cues in an artificial language. *Memory & Cognition, 30,* 678–686.

Musso, M., Moro, A., Glauche, V., Rijntjes, M., Reichenbach, J., Büchel, C., et al. (2003). Broca's area and the language instinct. *Nature Neuroscience, 6,* 775.

Nottebohm, F. (1993). The search for neural mechanisms that define the sensitive period for song learning in birds. *Netherlands Journal of Zoology, 43,* 193–234.

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

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

Pinker, S., & Bloom, P. (1990). Natural language and natural selection. *Behavioral and Brain Sciences, 13*(4), 707–784.

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

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

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

Shieber, S. M. (1985). Evidence against the context-freeness of natural language. *Linguistics and Philosophy, 8,* 333–343.

Shukla, M., Nespor, M., & Mehler, J. (2007). An interaction between prosody and statistics in the segmentation of fluent speech. *Cognitive Psychology, 54,* 1–32.