(in press, *Science*)

Computational constraints on syntactic processing in nonhuman primates

W. Tecumseh Fitch[1] and Marc D. Hauser[2]
[1]School of Psychology, University of St. Andrews
[2]Dept. of Psychology, Harvard University

The capacity to generate a limitless range of meaningful expressions from a finite set of elements differentiates human language from other animal communication systems. Rule systems capable of generating an infinite set of outputs ("grammars") vary in generative power. The weakest possess only local organizational principles, with regularities limited to neighboring units. We used a familiarization/discrimination paradigm to demonstrate that monkeys can spontaneously master such grammars. However, human language entails more sophisticated grammars, incorporating hierarchical structure. Monkeys tested with the same methods, syllables and sequence lengths were unable to master a grammar at this higher, "phrase structure grammar" level.

Syntax is one key component of human language, with no known equivalent in animal communication systems. The limitless expressive power of human language requires structures, termed phrases or sentences, above the word level (or by analogy above the single call level in animals). Linguistic syntax involves the rearrangement and permutation of such abstract hierarchical structures, often with concomitant changes in meaning. The production and perception of these hierarchical syntactic structures is a core capability underlying human linguistic competence. This level of organization goes far beyond the simple concatenation procedures sometimes called "syntax" in animal communication (*1-3*). However, the evolution of the language faculty presumably involved the incorporation of some ancestral primate cognitive capabilities. Thus, a critical question is whether hierarchical processing was one of these preexisting abilities, perhaps evolved to serve non-communicative functions (e.g. motor control, number or social cognition) (*4-10*)

Rule systems capable of generating infinite sets of sequences ("grammars") are arranged in a mathematical hierarchy of increasing generative power, termed the Chomsky hierarchy (*11, 12*). The weakest class in this hierarchy are finite state grammars (FSGs), which can be fully specified by transition probabilities between a finite number of "states" (e.g., corresponding to words or calls). Recent evidence suggests that parsing procedures at this superficial level of complexity are spontaneously available to both human infants and nonhuman primates (*13-17*). However, FSGs are inadequate to generate all the structures of any human language (*11, 18*), because all languages minimally require procedures at the next level of complexity, termed phrase structure grammars (or PSGs, see *29*). In addition to concatenating items like an FSG, a PSG can embed strings within other strings, thus creating complex hierarchical structures ("phrase structures"), and long-distance dependencies. For example, in English, the word "if" is typically followed by the word "then", but any arbitrary number of words or phrases can be inserted between them. Such constructions (and many others) demand more sophisticated parsing capabilities, including a perceptual ability to recognize these structures and an open-ended memory to store them. There is a broad consensus in linguistics and machine learning both that PSGs are more powerful than FSGs, and that grammars above the FSG level are, minimally, a crucial component of all human languages (*12, 19, 20*). While such abilities are available to all normal humans, it is currently unknown whether parsing abilities above the FSG level are available to nonhuman animals. We used a familiarization/discrimination procedure to address this issue in cotton-top tamarins (*Saguinus oedipus*), a New World primate species that has previously demonstrated successful discrimination of linguistic stimuli according to rhythmic class, along with a capacity to grasp transitional probabilities and abstract rules implicit in speech stimuli (*15, 16, 21*).

The infinite nature of grammars renders empirical tests of their comprehension problematic (*18, 22*). Because limited output from a PSG can always be approximated by a more complicated FSG (at the limit, a memorized list of exemplars), it is difficult to prove conclusively that subjects have learned the former. This is equally true for human or animal subjects. However, failure to master a grammar (as demonstrated by a failure to distinguish grammatical from ungrammatical strings) can be empirically confirmed. Of course, such a failure could occur for myriad reasons, and it is thus imperative to demonstrate success on a similar task, matched in all extraneous respects, before concluding that particular computational constraints are at work. Thus, based on Chomsky's original discussion (*11, 12*) we created two grammars, which were used to generate meaningless auditory strings consisting of sampled consonant-vowel (CV) speech syllables. Previous research demonstrates that such syllabic speech streams are readily attended to and processed by cotton-top tamarins without training (*15, 21*). The two grammars were designed to equate extraneous non-grammatical variables, and thus to differ specifically in their capacity to generate hierarchical phrase structure.

Each grammar created structures out of two classes of sounds, A & B, each of which was represented by 8 different CV syllables (see Table 1). The A and B classes were perceptually clearly distinguishable to both monkeys and humans: different syllables were spoken by a female (A) and a male (B), and differentiated by voice pitch (> 1 octave difference) phonetic identity, average formant frequencies, and various other aspects of the voice source. For any given string, the particular syllable from each class was chosen at random. Crucially, syllables for each class were sampled without replacement, since otherwise the possibility of exact acoustic repetitions in the PSG and not in the FSG would make the two grammars distinguishable on superficial grounds. The FSG was $(AB)^n$, in which a random "A" syllable was always followed by a single random "B" syllable, and such pairs were repeated *n* times. The corresponding phrase structure grammar, termed A^nB^n, generated strings with matched numbers of A and B syllables. In this grammar, *n* sequential "A" syllables must be followed by precisely *n* "B" syllables. We chose the A^nB^n grammar because it is the simplest PSG that cannot, in principle, be approximated with an FSG, but which can easily be brought into correspondence with a simple FSG in all non-grammatical respects as required for our experiment. Further, this grammar is trivially easy for humans to learn (see below). The A^nB^n grammar produces center-embedded constructions, which although less common in human language than other (e.g. right-branching) structures, are ubiquitous in mathematics (e.g. nested parentheses in formulas) or computer programming languages (e.g. BEGIN-END statements). Like any PSG, it requires additional computational machinery beyond a finite-state automaton. In computer science terminology, this addition would minimally be a push-down stack. In psychological terms, it requires some way to recognize a correspondence between either the groups formed by the As and Bs (e.g. counting) or between specific As and corresponding Bs (e.g. long-distance dependencies). This PSG thus provides the ideal grammar for the empirical issue addressed by this study by allowing us to focus on the generative power of the system without introducing extraneous performance variables (e.g., memory capacity or referentiality).

Although each of these grammars can theoretically generate infinite numbers of strings of infinite length, memory limitations will impose limits on subjects' practical ability to parse strings. Because previous work demonstrates that tamarins can readily remember, and precisely discriminate between, strings up to three syllables in length (*23*), we restricted *n* to be two or three in both of the above grammars. 64 random strings were generated by each grammar, with 60 used for exposure and 4 different strings for testing (*29*).

Our testing method has been previously described in detail (*15*). Briefly, the tamarin colony was pseudorandomly divided into two groups, one per grammar. Each group included a mixture of sexes and ages (all adult). All of the monkeys in a particular group were simultaneously exposed in their home cages to 20 minutes of repeated playback of 60 different grammar-consistent strings, in random order, during the evening. They were then tested individually the next morning in a sound chamber. Testing started with a re-familiarization phase, when random stimuli from the previous evening's session were again played back for two minutes while the animal was fed treats (at a rate determined by the animal's feeding, and uncorrelated with stimulus presentation). We then closed the sound chamber door, started video monitoring and recording, and began playback of the test stimuli. No food was delivered during testing. Playback was initiated by the observer when the animal was looking down and away from the loudspeaker, and latency and duration of looking (orientation towards the loudspeaker Fig. 1a) were later scored blind to condition from the digitized video (>90% reliability). Each animal (regardless of the grammar on which they were trained) was tested with the same eight stimuli in random order. Four were novel stimuli consistent with the training grammar, while the other four were violations (but consistent with the other grammar).

Tamarins easily mastered the FSG, as demonstrated by a significant increase in looking to stimuli which violated the rules of the grammar (N = 10 monkeys, mean of 72% looking to violations but 34% looking to grammatically consistent novel stimuli, Wilcoxon signed rank test, $p < 0.007$; Fig 1b). At an individual level, 9 of 10 monkeys looked more to violations than consistent stimuli. Thus, the simple alternating sequential pattern embodied in this grammar was spontaneously perceived and remembered, and novel stimuli following the familiar pattern elicited less attention than novel stimuli violating it. This success demonstrates that the acoustic cues differentiating the two syllable classes were salient to our tamarin subjects. More importantly, the ability to learn the rule governing the construction of an acoustic sequence, without any explicit training, indicates that tamarins are sensitive to regularities in an acoustic stream, and can recognize novel strings as consistent with past inputs. This finding is consistent with previous research suggesting that monkeys are able, either with or without training, to discover the rules governing sequential patterns in auditory and visual stimuli (*15, 24, 25*).

In contrast, tamarins failed to master the PSG, displaying an equivalent rate of looking to both consistent and inconsistent strings (N = 10 monkeys, 29% looks to inconsistent and 31% looks to consistent stimuli; Fig 1c). No monkey looked at more than half of the violations. The failure to master the PSG cannot be due to extraneous factors such as stimulus length, loudness or other acoustic factors, inability to perceive the A and B classes, or differences in exposure, testing or evaluation procedures, all of which were consistent between the two grammars. All of the test subjects had equivalent experience in this testing situation, and successfully mastered many other tasks in this laboratory. The pattern of results is what one would expect if tamarins attempted to parse the PSG strings by building an FSG structure (based on simple transitional probabilities, an ability of tamarins documented both here and elsewhere (*15, 17*)). Furthermore, in two other attempts to test tamarins on this PSG, using slight modifications of stimulus type and/or testing procedures, we have similarly found no ability to master this rule (*30*). Thus, it appears that cotton-top tamarins have difficulties in spontaneously learning a rule of this type, despite their demonstrated ability to master FSGs equivalent in every respect except for hierarchical structure.

An alternative explanation for these results might be that tamarins fail the PSG because their ability to differentiate successive items is limited to runs of two. If this were true, it would account for the asymmetric results we obtained because they would be able to encode AB AB AB patterns but unable to process the longer runs of AAA BBB. However, a subanalysis gave the same pattern of results even when *n* was limited to two (ABAB vs. AABB): tamarins clearly discriminated violations from consistent stimuli in the FSG grammar (Wilcoxon signed rank, $p < 0.02$) but failed to discriminate these in the PSG (Wilcoxon signed rank, $p = 0.68$). The data are thus inconsistent with this alternative hypothesis.

In sharp contrast to the monkeys, adult humans tested with these same grammars showed rapid learning of either grammar (with only three minutes of exposure), and were easily able to discriminate grammatical from non-grammatical stimuli for both grammars (Fig 2). Undergraduate subjects were passively exposed to the same training stimuli as the tamarins, and then were tested on the same test stimuli (*29*). Subjects scored 93% correct on the FSG and 85% on the PSG, indicating that adult humans can easily distinguish between, and master, either grammar under the same experimental conditions where the monkeys failed on the PSG. These data are consistent with other experimental findings that humans can learn a PSG and appear to prefer phrase-structured input (*18, 26, 27*), and with the widelyaccepted theoretical claim that human languages demand acquisition of rule systems at the PSG level (*11*).

These results suggest that, despite a clear ability to process sequential regularities in acoustic strings, tamarins are unable to process a simple phrase structure, where components at one portion of a string are related to other components some distance away. Because earlier work with this species using the same paradigm demonstrates that these animals are perfectly capable of storing and recalling at least three separate stimuli and comparing them with subsequent strings, this computational limitation does not result from some lower-level limitation on memory, attention, or number discrimination. Further work will be necessary using other methods (e.g. training and reinforcement), different grammars, and other species (e.g. apes), before any broad conclusions can be drawn about nonhuman primate limitations in general. It is also possible that nonprimates such as songbirds, which have some rule-based structure in their songs, would fare better at the task developed here. However, the current findings suggest that tamarins suffer from a specific and fundamental computational limitation on their ability to spontaneously recognize or remember hierarchically-organized acoustic structures. Put differently, the limitation we have demonstrated might indicate an over-reliance on superficial aspects of stimuli, which prevents tamarins from perceiving more abstract relationships available in the signal, as has been suggested by previous work on primate auditory perception (*28*). If nonhumans are "stuck" trying to interpret PSG-generated stimuli at the FSG level, it would make PSG stimuli seem much more complex to them, and perhaps even unlearnable in finite time. While the evolution of well-developed hierarchical processing abilities in humans might have benefited many aspects of cognition (e.g., spatial navigation, tool use or social cognition), this capability is one of the crucial requirements for mastering any human language. Thus, the acquisition of hierarchical processing ability may have represented a critical juncture in the evolution of the human language faculty.

Methods (for online Supplementary Online Material publication)

Speech syllables were recorded with a Sennheiser MKH-60 microphone and Tascam DA-P1 recorder in an IAC acoustic isolation chamber (Industrial Acoustics Company, Bronx, New York), and digitally transferred to computer via an Audiomedia III card (Digidesign, www.digidesign.com) at 16-bit quantization and 44.1 kHz sampling rate. Stimuli were generated using custom software programmed using the Supercollider programming language (www.audiosynth.com). Playback was controlled by custom software written in Hypercard (Apple Computer, www.apple.com). Playback was at the same sampling rate and quantization via the Audiomedia card and an Alesis Studio Monitor loudspeaker and amplifier in an IAC 400-A ventilated sound chamber (www.industrialacoustics.com). Videos were digitized for blind coding using Adobe Premiere (www.adobe.com). Each trial was saved as a separate file, the time of the test sound was marked visually, and each file was then coded blind to stimulus identity by two experienced observers. Inter-observer reliabilities were 90% or more for both grammars (90% FSG, 92% PSG).

Each of ten undergraduate subjects (all female, age 19-26) were tested on both grammars, using the same stimuli as for the monkeys, in balanced pseudo-random order. Playback was controlled by custom software written in Hypercard 2.4.1 (Apple Computer, www.apple.com) at the same sampling rate and quantization, through the built-in speakers on an Apple Powerbook G3 Macintosh. Each student was exposed to 30 strings randomly selected from the random strings presented to the monkeys. Students were asked simply to listen to the sounds for three minutes, and then were asked to rate a set of novel sounds, stating simply whether the pattern of each novel sound was the same as or different from the previous set, by pressing a button on the computer screen. No feedback was given. For strict comparability with the tamarin results, students always first rated precisely the eight test stimuli used for the monkeys (in random order) and then were tested with the remaining 25-30 stimuli (in random order). Order of testing for the FSG and PSG conditions was counterbalanced. After finishing the monkey-identical test trials, students listened to additional randomized trials to determine more precisely what they had learned. Of the additional trials, 15 were simply novel strings generated by the target grammar ("same" trials). Of the remaining "different" trials five were from the alternate grammar, and five reversed the order (so B's preceded A's in both grammars). Overall, students scored 90% correct (N = 330, 297 correct) on the FSG, and 89% correct (N = 367, 326 correct) on the PSG. A final category of five "same" trials extended the PSG grammar only, by using an *n* of 4 (i.e. AAAA BBBB). "Same" responses to this particular manipulation indicate generalization beyond n = 2 or 3. 41 of 49 answers (84%) were "same" for these trials, indicating that human adults generalized beyond the *n* given in the stimuli (suggesting that the subjects had generalized to A^nB^n rather than (A^2B^2 OR A^3B^3)). Subjects' performance for these, and other (Fitch, unpublished data), additional test stimuli was consistent with their having educed the PSG rule, as opposed to any FSG-level approximation thereof.

Technical terminological note:
Within finite state grammars there are multiple levels of power. Here, as is typical in both machine learning and linguistics, we restrict the term finite-state grammar to the simplest of these (synonymous with a "regular grammar", e.g. implemented as a first-order stationary Markov process). Other tests would be necessary to determine if monkeys can master more powerful FSGs, for example "*n*th order Markov processes" which can take account of their past *n* states in determining the current output. "Phrase structure grammars" (types 0, 1 and 2 of the Chomsky hierarchy) are also subdivided into subclasses of increasing generative power. All have in common the ability to generate recursively-embedded, hierarchically-structured strings. The A^nB^n grammar described here is a "context-free grammar", the weakest (most constrained) member of this class. Although there is debate concerning the necessity of more powerful classes of PSG (e.g. context-sensitive or transformational grammars), there is a broad consensus in linguistics and machine learning both that PSGs are more powerful than FSGs, and that grammars at the context-free level are, minimally, a crucial component of all human languages. We adopt the somewhat old-fashioned term "phrase structure grammar" here both because it avoids these technical debates, which are immaterial to this study, and because it is the term used for such grammars in Chomsky's original publications. The term "phrase structure grammar" also has the virtue of saying what it is, unlike "context-free grammar" which says what it is not.

Figure Legends:

Fig. 1. Familiarization/Discrimination paradigm:

A. Examples of the stimuli for the finite-state grammar and phrase structure grammar used in this study. Grammars were matched for length, composition, loudness and other acoustic features, and testing and evaluation procedures were identical for the two grammars. A and B stimulus classes were spoken by different speakers, a female (boldface) and male (normal) and thus differed considerably in pitch, as well as phonetic identity and other acoustic variables.

B. A familiarization/discrimination paradigm was used. We quantified a cotton-top tamarin's propensity to orient towards a stimuli using offline blind scoring of videos. The stimuli were either consistent with, or violated, the rules implicit in a previous set of familiarization strings.

Fig. 2. Experimental evidence that monkeys can master finite-state but not phrase-structure grammars.

Right: Humans exposed to a finite-state grammar with only local sequential structure (top), or a phrase structure grammar with hierarchical structure (bottom), rejected violations as "different" and accepted consistent stimuli as "same".

Left: Monkeys exposed to the same finite state grammar (top) oriented significantly more often to violations, and did not orient to novel strings consistent in structure with the familiar strings. However, when exposed to the phrase-structure grammar (bottom), monkeys failed to discriminate between consistent and inconsistent strings. looking at a similar (random baseline) level to both sets of stimuli.

---

1. J. P. Hailman, M. S. Ficken, *Animal Behaviour* 34, 1899-1901 (1987).
2. J. G. Robinson, *Behaviour* 90, 46-79 (1984).
3. K. Zuberbühler, *Animal Behaviour* 63, 293-299 (2002).
4. M. Hauser, N. Chomsky, W. T. Fitch, *Science* 298, 1569-1579 (Nov 22, 2002, 2002).
5. P. M. Greenfield, K. Nelson, E. Saltzman, *Cognitive Psychology* 3, 291-310 (1972).
6. P. M. Greenfield, *Behavioral and Brain Sciences* 14, 531-595 (1991).
7. R. W. Byrne, A. E. Russon, *Behavioral and Brain Sciences* 21, 667-684 (1998).
8. D. Kimura, *Neuromotor Mechanisms in Human Communication* (Oxford University Press, Oxford, England, 1993).
9. P. Lieberman, *Journal of Human Evolution* 14, 657-668 (1998).
10. B. McGonigle, M. Chalmers, A. Dickinson *Animal Cognition* 6, 185-197 (2003).
11. N. Chomsky, *Syntactic Structures.* (Mouton, The Hague, 1957).
12. N. Chomsky, *Information and Control* 2, 137-167 (1959).
13. J. Saffran, D. Aslin, E. Newport, *Science* 274, 1926-1928 (1996).
14. R. L. Gomez, L. Gerken, *Cognition* 70, 109-135 (Mar 1, 1999).
15. M. D. Hauser, E. L. Newport, R. N. Aslin, *Cognition* 78, 53-64 (2001).
16. M. D. Hauser, D. Weiss, G. Marcus, *Cognition* 86, B15-B22 (2002).
17. E. L. Newport, M. D. Hauser, G. Spaepen, R. N. Aslin, *Cognitive Psychology* (in press.).
18. G. A. Miller, in *Psychology of Communication* G. A. Miller, Ed. (Basic Books, New York, 1967).
19. L. Haegeman, *Introduction to government & binding theory* (Blackwell, Oxford, 1991).
20. E. Charniak, D. McDermott, *Introduction to Artificial Intelligence* (Addison-Wesley, Reading, Massachusetts, 1985).
21. F. Ramus, M. D. Hauser, C. T. Miller, D. Morris, J. Mehler, *Science* 288, 349-351 (2000).
22. G. A. Miller, N. Chomsky, in *Handbook of Mathematical Psychology* R. D. Luce, R. R. Bush, E. Galanter, Eds. (John Wiley & Sons, New York, 1963), vol. II, pp. 419-492.
23. M. D. Hauser, S. Dehaene, G. Dehaene-Lambertz, A. L. Patalano, *Cognition* 86, B23-B32 (2002).

24. A. A. Wright, H. C. Santiago, S. F. Sands, D. F. Kendrick, R. G. Cook, *Science* 229, 287-289 (1985).

25. H. S. Terrace, L. K. Son, E. M. Brannon, *Psychological Science* 14, 66-73 (2003).

26. J. L. Morgan, E. L. Newport, *Journal of Verbal Learning and Verbal Behavior* 20, 67-85 (1981).

27. J. L. Morgan, R. P. Meier, E. L. Newport, *Journal of Memory and Language* 28, 360-374 (1989).

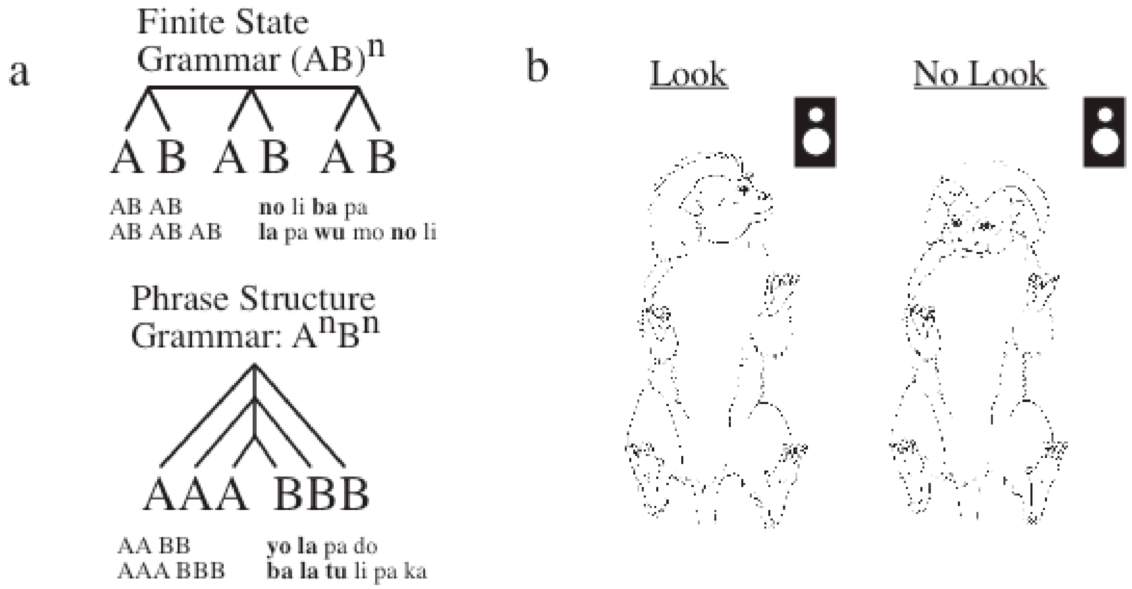
28. M. R. D'Amato, *Music Perception* 5, 452-480 (1988).

29, See Supplementary Online Material.

30. This is the third attempt we have made, over a period of several years, to test tamarins on this PSG, using slight modifications of stimulus type and/or testing procedures. All of these attempts have been complete failures, yielding no evidence that these monkeys were able to abstract the phrase structure rule. Briefly, the two previous experiments utilized the same A^nB^n grammar with stimuli and training procedures modeled on previous FSGs that tamarins had successfully acquired. The first was based on the techniques of (*15*) and used tonal stimuli differing in pitch (similar to the natural calls of cotton-top tamarins). The second was based on (*16*) , and used identical techniques as well as the same synthesized speech syllables that were used in that study. In each case, tamarins presented with the PSG version failed to show any differentiation, based on various possible measures of response, between novel grammatical and novel agrammatical stimuli.

31. We thank three anonymous reviewers, Richard Aslin, Noam Chomsky Ray Jackendoff, Mark Johnson, Elissa Newport, Steven Pinker and Jenny Saffran for useful discussions and/or comments on the manuscript, and Juergen Weissenborn and Barbara Hoehle for assistance in gathering human data This research was supported by an NSF ROLE and McDonnell grant to MDH and an NIH training grant to WTF.

Fitch & Hauser Figure 1



Fitch & Hauser Figure2