Published in final edited form as:
Nature. 2006 April 27; 440(7088): 1204–1207. doi:10.1038/nature04675.

Recursive syntactic pattern learning by songbirds

Timothy Q. Gentner[1,†], **Kimberly M. Fenn**[2], **Daniel Margoliash**[1,2], and **Howard C. Nusbaum**[2]

1*Department of Organismal Biology and Anatomy, University of Chicago, Chicago, Illinois 60637, USA*

2*Department of Psychology, University of Chicago, Chicago, Illinois 60637, USA*

Abstract

Humans regularly produce new utterances that are understood by other members of the same language community[1]. Linguistic theories account for this ability through the use of syntactic rules (or generative grammars) that describe the acceptable structure of utterances[2]. The recursive, hierarchical embedding of language units (for example, words or phrases within shorter sentences) that is part of the ability to construct new utterances minimally requires a 'context-free' grammar[2,3] that is more complex than the 'finite-state' grammars thought sufficient to specify the structure of all non-human communication signals. Recent hypotheses make the central claim that the capacity for syntactic recursion forms the computational core of a uniquely human language faculty[4,5]. Here we show that European starlings (*Sturnus vulgaris*) accurately recognize acoustic patterns defined by a recursive, self-embedding, context-free grammar. They are also able to classify new patterns defined by the grammar and reliably exclude agrammatical patterns. Thus, the capacity to classify sequences from recursive, centre-embedded grammars is not uniquely human. This finding opens a new range of complex syntactic processing mechanisms to physiological investigation.

The computational complexity of generative grammars is formally defined[3] such that certain classes of temporally patterned strings can only be produced (or recognized) by specific classes of grammars (Fig. 1). Starlings sing long songs composed of iterated motifs (smaller acoustic units)[6] that form the basic perceptual units of individual song recognition[7-9]. Here we used eight 'rattle' and eight 'warble' motifs (see Methods) to create complete 'languages' (4,096 sequences) for two distinct grammars: a context-free grammar (CFG) of the form A^2B^2 that entails recursive centre-embedding, and a finite-state grammar (FSG) of the form $(AB)^2$ that does not (Fig. 2a, b; '*A*' refers to rattles and '*B*' to warbles).

We trained 11 European starlings, using a go/nogo operant conditioning procedure, to classify subsets of sequences from these languages (see Methods and Supplementary Information). Nine out of eleven starlings learned to classify the FSG and CFG sequences accurately (as assessed by *d*', which provides an unbiased measure of sensitivity to differentiating between two classes of patterns), but this task was difficult (Fig. 2c). The rate of acquisition varied widely among the starlings that learned the task (303.44 ± 57.11 blocks to reach criterion (mean ± s.e.m.), range 94-562 blocks with 100 trials per block), and was slow by comparison to other operant song-recognition tasks[7].

To assess the possibility that starlings learned to classify correctly the motif patterns described by the CFG and FSG grammars through rote memorization of the training exemplars, we further

Correspondence and requests for materials should be addressed to T.Q.G. (tgentner@ucsd.edu)..
†Present address: Department of Psychology, University of California, San Diego, La Jolla, California 92093, USA.

tested the first four birds to reach stable asymptotic performance on the initial classification training (mean *d*' ± s.e.m. at asymptote 2.52 ± 0.40, Fig. 2d). We transferred the birds abruptly from the 16 baseline training stimuli to 16 new sequences from the same two grammars (A^2B^2 and $(AB)^2$, eight sequences from each) while maintaining the same operant contingencies used during baseline training. Starlings correctly classified the new CFG and FSG sequences during the first transfer session (Fig. 3a). The mean *d*' over the first 100 trials with new stimuli (roughly six responses to each exemplar) was 1.08 ± 0.50, which is significantly better than chance performance (*d*' = 0). Over the first five 100-trial blocks of the transfer session, the mean *d*' was 1.14 ± 0.20 (Fig. 3a), and the lower bound of the 95% confidence interval (CI) around *d*' was above zero for all birds (range 0.34-1.18), with subsequent performance continuing to improve. Thus, the birds did not simply memorize the 16 baseline training stimuli, but instead acquired general knowledge about features diagnostic of the two grammars, and applied this knowledge to classify the new stimuli correctly. Given that the same elements (motifs) composed the sequences in each class, this knowledge must be related to the differential patterning of these elements by each grammar. Additional generalization tests using 'probe' procedures that test for learning during exposure to the new grammatical stimuli (see Methods and Fig. 3b) also reject the rote memorization hypothesis, and support the conclusion that the birds acquired information about the patterning of motifs in the CFG and FSG classes.

One possibility consistent with interpretations of experiments on syntactic processing in cotton-top tamarins[10-12] is that the birds learned only the FSG, and treated the grammatical CFG sequences as the complement set. However, the results of further probe tests rule out this possibility. We constructed 16 motif sequences based on four different agrammatical patterns (*AAAA*, *BBBB*, *ABBA* and *BAAB*, with four exemplars per pattern, using the same *A* and *B* motifs as for the FSG and CFG grammars) and presented them along with new A^2B^2 and $(AB)^2$ patterns as probe stimuli (Methods). The response patterns for the agrammatical probe stimuli differed significantly from those for new $(AB)^2$ stimuli for all four birds, and from those for new A^2B^2 stimuli for three of the four birds (X^2 calculated separately for each bird and stimulus class, $P < 0.05$ in 7 of 8 cases; $X^2 = 1.74$ (not significant) for one case). It may be relevant that the nonconforming bird was the only one of the four tested for which the FSG sequences served as the S^+ stimuli (see Methods). Regardless, three of the four birds clearly learned to classify both the CFG and FSG stimuli during training, suggesting that they learned both FSG and the CFG patterning rules.

Time and memory capacity both constrain the functional length of any grammatical string, yet part of the power of a generative grammar is its capacity to describe strings of arbitrary length. To test whether our birds generalized from A^2B^2 and $(AB)^2$ to higher orders of grammatical structure, we probed the birds with $n = 3$ (that is, A^3B^3 and $(AB)^3$) and $n = 4$ motif sequences while they maintained baseline ($n = 2$) classification (see Methods and Fig. 3b). All birds accurately classified the $n = 3$ CFG and FSG sequences (mean *d*' 1.37 ± 0.54; range for lower bound of 95% CI 0.03-2.23) and the $n = 4$ CFG and FSG sequences (mean *d*' 1.27 ± 0.22; range for lower bound of 95% CI 0.23-1.54). Thus, classification training with $n = 2$ sequences can induce behaviour consistent with higher-order, generative grammars, including those using recursive centre-embedding.

An alternative explanation for these results is that the birds learned a 'simpler' grammar that approximates the recursive structure in the A^nB^n sequences. Sequences that follow the pattern A^nB^n constitute a subset of those that follow the more general pattern $A\bigstar B\bigstar$, in which the number of *a*'s and *b*'s can vary independently. Although a CFG is required to produce sequences in which the number of *a*'s and *b*'s are matched, (as in A^nB^n), the whole of $A\bigstar B\bigstar$ can be generated by a finite-state grammar. We tested whether the birds learned an $A\bigstar B\bigstar$ finite-state approximation to A^nB^n by examining their responses to the following $A\bigstar B\bigstar$ patterns: A^1B^3, A^3B^1, A^2B^3 and A^3B^2 (four randomly chosen sequences per pattern, same *A/*

B motif vocabularies as all FSG and CFG stimuli). We presented the $A\bigstar B\bigstar$ stimuli along with A^nB^n and $(AB)^n$ (n = 2, 3, 4) sequences as probes during the same sessions.

If birds learned $A\bigstar B\bigstar$ as a finite-state approximation to A^nB^n, then the pattern of response to each $A\bigstar B\bigstar$ stimulus should match the response to A^nB^n reference stimuli. Our results reject this hypothesis. All birds showed a strong bias to treat the $A\bigstar B\bigstar$ patterns differently from the A^nB^n reference stimuli, while maintaining accurate classification of the A^nB^n and $(AB)^n$ reference and training stimuli (mean $d' \geq 1.19$ in all four cases). Responses to all four $A\bigstar B\bigstar$ patterns were significantly different from responses to the A^2B^2 and A^4B^4 reference stimuli ($X^2 < 0.001$ for all 8 cases, d.f. = 3). Responses to the A^1B^3, A^3B^2 and A^3B^1 patterns were significantly different from responses to the A^3B^3 patterns ($X^2 < 0.001$ in all cases, d.f. = 3). In addition, all six pair-wise comparisons between responses to individual $A\bigstar B\bigstar$ patterns were significantly different ($X^2 < 0.0005$ for all cases, d.f. = 3) suggesting that the birds did not treat the $A\bigstar B\bigstar$ patterns as a single stimulus class. These results suggest that the birds did not solve the recursive classification tasks by learning a finite-state approximation to the CFG. Rather, it seems that they learned A^nB^n, or a functionally equivalent rule.

We then used the pattern of responses to the various agrammatical probe stimuli to test alternative hypotheses for the starlings' accurate classification of A^nB^n and $(AB)^n$ patterns. For example, the task could reduce to the classification '*AA*' and '*AB*', or '*BB*' and '*AB*', if only the initial (primacy) or terminal (recency) motif pairs are attended to, respectively. Similarly, birds could correctly classify A^nB^n and $(AB)^n$ motif sequences by listening for *B/A* transitions (A^nB^n patterns have none), counting the number of *A/B* transitions (A^nB^n patterns have only one), or listening for *AA* (or *BB*) motif pairs. Each of these potential solution strategies can be and were ruled out by considering specific comparisons among the various agrammatical probe stimuli (Fig. 4 and Supplementary Information). In all cases, classification of the agrammatical patterns was significantly worse (lower d') than grammatical probe stimuli, suggesting that these alternative strategies were not the basis for generalization performance. Instead, starlings seem to have learned the patterning rules defined by each grammar.

We thus demonstrate that starlings can recognize syntactically well-formed strings, including those that use a recursive centre-embedding rule. At least a simple level of recursive syntactic pattern processing is therefore shared with other animals. These results challenge the recent claim that recursion forms the computational core of a uniquely human narrow faculty for language (FLN)[4]. We attempted to rule out the most plausible finite-state solution strategies that could account for accurate classification of A^nB^n patterns (Supplementary Information), suggesting that the learned patterning rule conforms to a stochastic context-free grammar. In practice, however, the stimulus sets used to test such claims must be finite. Thus, the theoretical possibility remains that a finite-state grammar, however heavily contrived, may account for the observed behaviour (see Supplementary Information). Of course, theoretical difficulties in proving the use of context-free rather than finite-state grammars extend to studies of grammatical competence in humans as well, and therefore call into question the falsifiability of claims regarding CFGs in humans compared to non-humans.

Although uniquely human syntactic processing capabilities, if any, may reflect more complex context-free grammars or higher levels in the Chomsky grammatical hierarchy, it may prove more useful to consider species differences as quantitative rather than qualitative distinctions in cognitive mechanisms. Such mechanisms (for example, memory capacity) need not map precisely onto strict formal grammars and automata theories. There might be no single property or processing capacity that marks the many ways in which the complexity and detail of human language differs from non-human communication systems[13]. Future studies can gauge the extent of the recursive syntactic abilities demonstrated here, by examining the processing of

right-embedded structures more common in human languages (and more easily understood), and the interface between generalized syntactic and semantic knowledge.

METHODS

Baseline training

Motifs can be classified into four spectro-temporally distinct categories: whistles, warbles, rattles and high-frequency motifs[6] (Supplementary Information). We used eight 'rattle' and eight 'warble' motifs from the repertoire of a single male starling (sets a_i and b_i, respectively, i = 1-8) as the vocabulary for two distinct grammars (Fig. 2a, b). One grammar defined sequences of the form A^2B^2, and the other defined sequences of the form $(AB)^2$. For example, the explicit sequence of sound patterns rattle_i-rattle_j-warble_k-warble_l is defined by A^2B^2, but rattle_i-warble_k-rattle_j-warble_l, using the same four elements, is defined by $(AB)^2$. Because the song stimuli were created from a common vocabulary, only the patterning of motifs within each sequence varied between the classes defined by each grammar.

For each of the CFG (A^2B^2) and FSG ($(AB)^2$) grammars, we generated all possible motif sequences consisting of four elements. From each of these complete 'languages', we randomly selected three subsets of eight sequences such that within each subset, each motif appeared exactly once in each possible position, and no motif appeared twice in the same sequence. We chose one subset from each of the two grammars (sixteen sequences, eight per grammar) as the stimuli for initial operant classification training (Fig. 2 and Supplementary Information).

We trained starlings to classify sequences defined by the FSG and CFG using well-established operant procedures[9] (Supplementary Information). Starlings learned to respond regularly to the sequences defined by one grammar (S^+ stimuli) and to withhold responses to sequences defined by the other grammar (S^- stimuli). For half of the birds the CFG sequences served as S^+ stimuli and the FSG sequences as S^- stimuli, and for the other half these associations were reversed. Both baseline training groups (CFG sequences as S^+, n = 5; or CFG sequences as S^-, n = 4) learned at roughly equivalent rates. (There was no significant difference between groups in the mean number of trials required to reach criterion, P = 0.46, Mann-Whitney U-test.)

d' measure of classification

We used a learning criterion of $d' > 1.0$ and a lower bound of 95% confidence interval around d' above zero for five consecutive blocks (100 trials per block). The measure of d' indexes the subject's sensitivity to differentiating between the two classes of patterns presented as stimuli, uncontaminated by response bias (Supplementary Information).

Probe-session reinforcement contingencies

During probe sessions we presented familiar CFG and FSG sequences on 80% of all trials, and various new test ('probe') stimuli on the remaining 20% of trials. We reinforced responses to the familiar stimuli in the normal manner (but at a reduced rate), and randomly reinforced all responses to the probe stimuli, regardless of accuracy, with equal rates of food and a short time-out (see Supplementary Information). As probe stimuli, we presented new A^nB^n and $(AB)^n$ patterns (n = 2, 3 or 4), agrammatical patterns of the form *AAAA*, *BBBB*, *ABBA*, *BAAB*, and patterns defined by the FSG $A\bigstar B\bigstar$.

Higher-order stimuli

We used the CFG and FSG grammars to randomly generate 16 motif sequences with n = 3 (that is, A^3B^3 and $(AB)^3$, eight per grammar) and 16 sequences with n = 4 (eight per grammar). As for n = 2, for each set of eight exemplars from a given grammar at a given order, each motif

appeared in every position once and only once, and the same motif never appeared more than once in the same exemplar sequence. We then presented the $n = 3$ and $n = 4$ sequences as probe stimuli (in separate sessions for each n) while birds continued classification of the $n = 2$ baseline training stimuli.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We thank T. Brawn for help in conducting some of these experiments, and A. Henly, P. Visser and L. Kay for comments on an earlier draft. This research was supported by an NIH grant to D.M.

References

1. Hockett CF. The origin of speech. Sci. Am 1960;203:89–96. [PubMed: 14402211]
2. Chomsky, N. Syntactic Structures. Mouton; The Hague: 1957.
3. Hopcroft, J.; Ullman, J. Introduction to Automata Theory, Languages, and Computation. Addison-Wesley; Reading, Massachusetts: 1979.
4. Hauser MD, Chomsky N, Fitch WT. The faculty of language: what is it, who has it, and how did it evolve? Science 2002;298:1569–1579. [PubMed: 12446899]
5. Fitch WT, Hauser MD, Chomsky N. The evolution of the language faculty: clarifications and implications. Cognition 2005;97:179–210. [PubMed: 16112662]
6. Eens M. Understanding the complex song of the European starling: An integrated approach. Adv. Study Behav 1997;26:355–434.
7. Gentner TQ, Hulse SH. Perceptual mechanisms for individual vocal recognition in European starlings, *Sturnus vulgaris*. Anim. Behav 1998;56:579–594. [PubMed: 9784206]
8. Gentner TQ, Hulse SH. Perceptual classification based on the component structure of song in European starlings. J. Acoust. Soc. Am 2000;107:3369–3381. [PubMed: 10875382]
9. Gentner TQ, Margoliash D. Neuronal populations and single cells representing learned auditory objects. Nature 2003;424:669–674. [PubMed: 12904792]
10. Hauser MD, Newport EL, Aslin RN. Segmentation of the speech stream in a non-human primate: statistical learning in cotton-top tamarins. Cognition 2001;78:B53–B64. [PubMed: 11124355]
11. Hauser MD, Weiss D, Marcus G. Rule learning by cotton-top tamarins. Cognition 2002;86:B15–B22. [PubMed: 12208654]
12. Fitch WT, Hauser MD. Computational constraints on syntactic processing in a nonhuman primate. Science 2004;303:377–380. [PubMed: 14726592]
13. Pinker S, Jackendoff R. The faculty of language: what's special about it? Cognition 2005;95:201–236. [PubMed: 15694646]

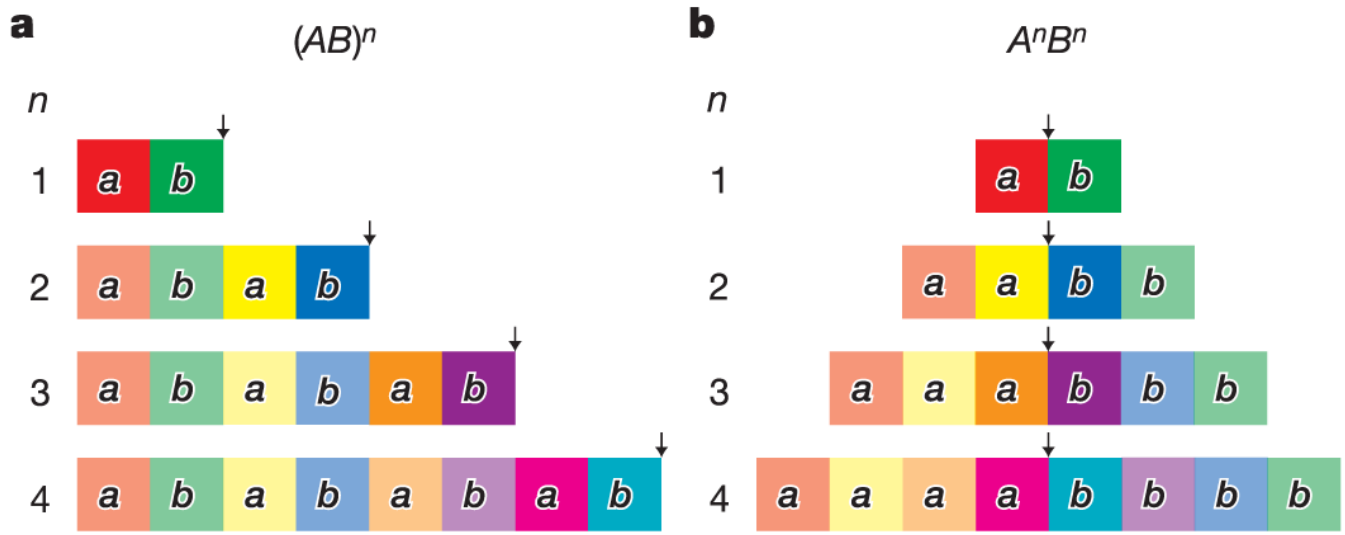

Figure 1. Grammatical forms

a, Finite-state form $(AB)^n$. **b,** Context-free form A^nB^n. Both grammars describe patterned sequences of elements (lower-case letters) of the sets '*A*' and '*B*'. Longer strings of the form $(AB)^n$, where *n* gives the number of *AB* iterations, are produced by appending elements to the end of an *n* - 1 sequence. Longer strings with the form A^nB^n are produced by embedding elements into the centre of an *n* - 1 sequence. Learning of and generalization to an A^nB^n pattern implies the capacity to process syntactic structures generated through recursive centre-embedding. Black arrows denote insertion points for higher-order sequences. Brightly coloured squares mark the '*AB*' phrase inserted at each order. Different hues denote different elements.

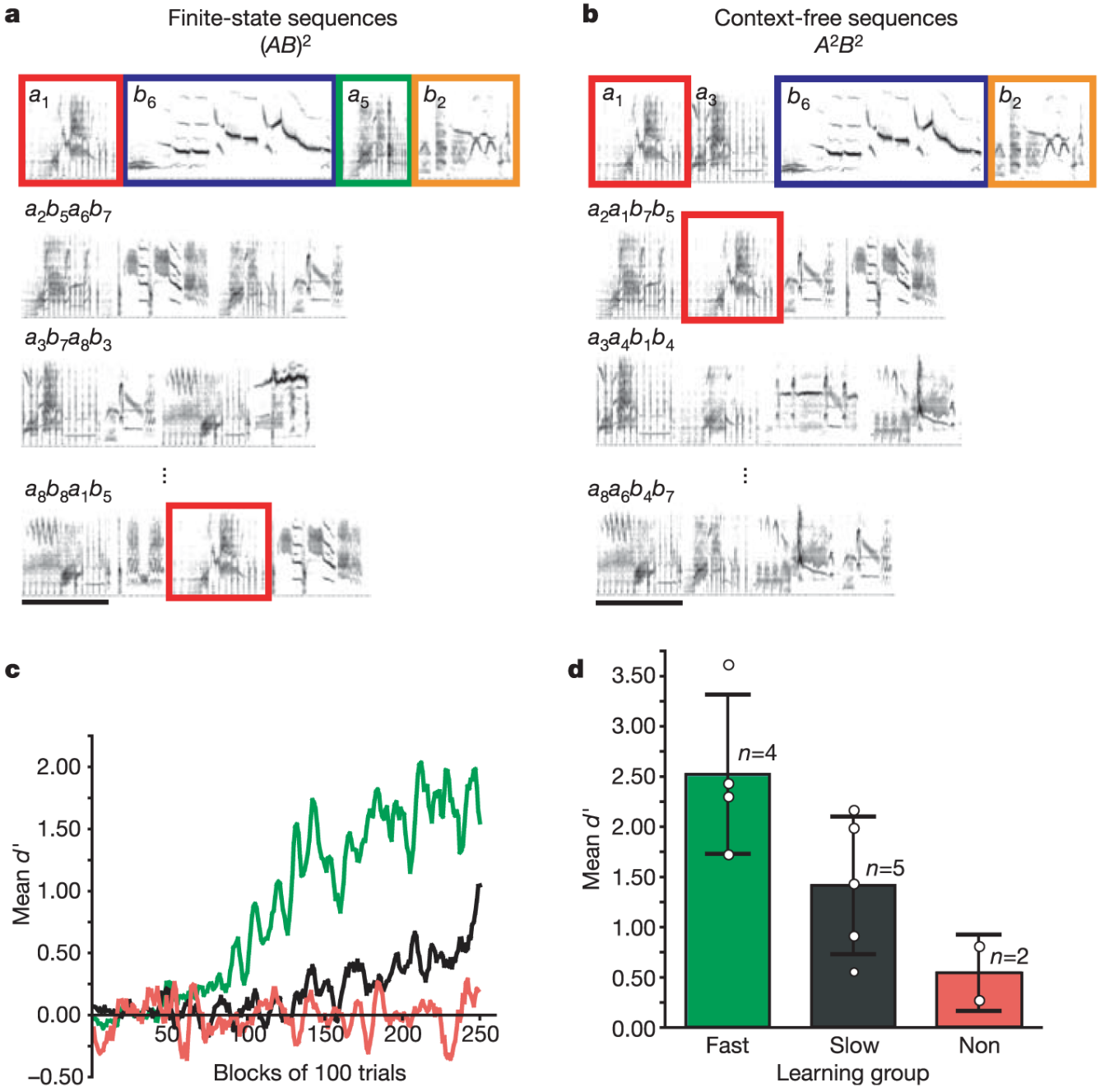

Figure 2. Classification of grammatical pattern stimuli

a, b, Sonograms (frequency range 0.2-10.0 kHz; scale bars, 1 s) showing four of the eight sequences constructed from the finite-state grammar $(AB)^n$ (**a**) and the context-free grammar A^nB^n (**b**) used in the initial FSG versus CFG pattern classification training with $n = 2$. Similarly coloured boxes mark the same motifs in multiple sequences. The position of a motif within a sequence is arbitrary with respect to its subscript label. See Supplementary Information for complete stimulus patterns and sonograms. **c,** Acquisition curves for the baseline FSG/CFG classification, showing mean *d*' over the first 250 blocks (100 trials per block) for birds that learned quickly and were subjected to further testing (green), birds that learned slowly (black) and birds that did not reach the accuracy criterion (red; see Methods). **d,** Mean *d*' (±s.d.) on

the baseline CFG versus FSG classification task at asymptote. Open circles show means from individual birds. Colours and groups as in **c**.

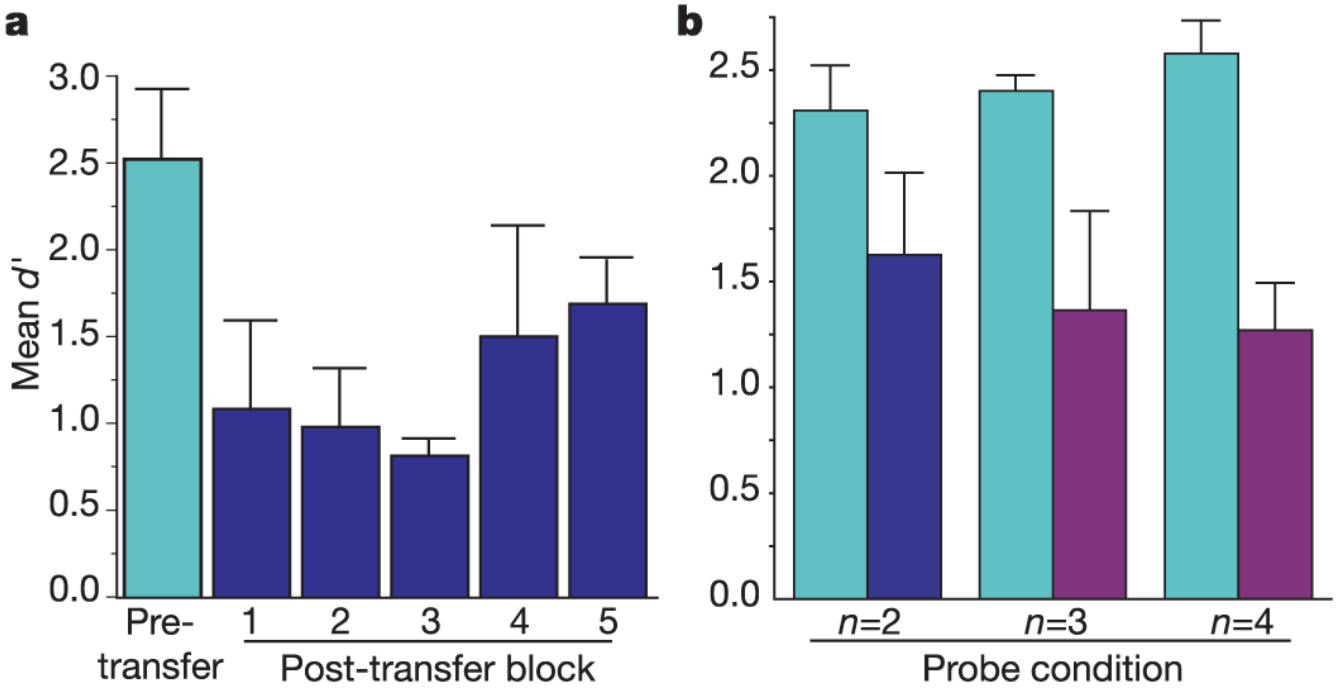

Figure 3. Generalization to new FSG and CFG sequences

a, Mean *d*' (±s.e.m.) for transfer from the training to new FSG and CFG stimuli (turquoise, mean performance over the five blocks of trials preceding transfer; blue, performance in the first five blocks after transfer; 100 trials per block). Performance was stable across these post-transfer blocks ($F_{3,4}$ = 1.15, P = 0.35, repeated measures ANOVA), then increased gradually to pre-transfer levels (not shown). All mean *d*' values shown are significantly greater than zero (see text). Acquisition of the transfer stimuli was much faster than for the original training sets (12.50 ± 3.11 blocks to criterion (mean ± s.e.m.), range 8-15 blocks; 100 trials per block), which can be attributed partially to generalization across the CFG and FSG classes. **b,** Mean *d*' (±s.e.m.) during grammatical probe sessions. Birds correctly classified new A^nB^n and $(AB)^n$ sequences when n = 2 (blue), n = 3 or n = 4 (purple). Classification accuracy was significantly above chance for all three types of probe sequences (mean *d*' for n = 2, 1.63 ± 0.39; see text for n = 3, n = 4). Classification of the baseline training stimuli (turquoise) was well above chance for all three conditions (mean *d*' ≥ 2.39, s.d. ≤ 0.25). The drop between training and probe stimulus classification was significant in only the n = 4 condition (P < 0.05, Mann-Whitney *U*-test), suggesting that these sequences were more difficult to classify correctly than the other grammatical test sequences (see Supplementary Information).

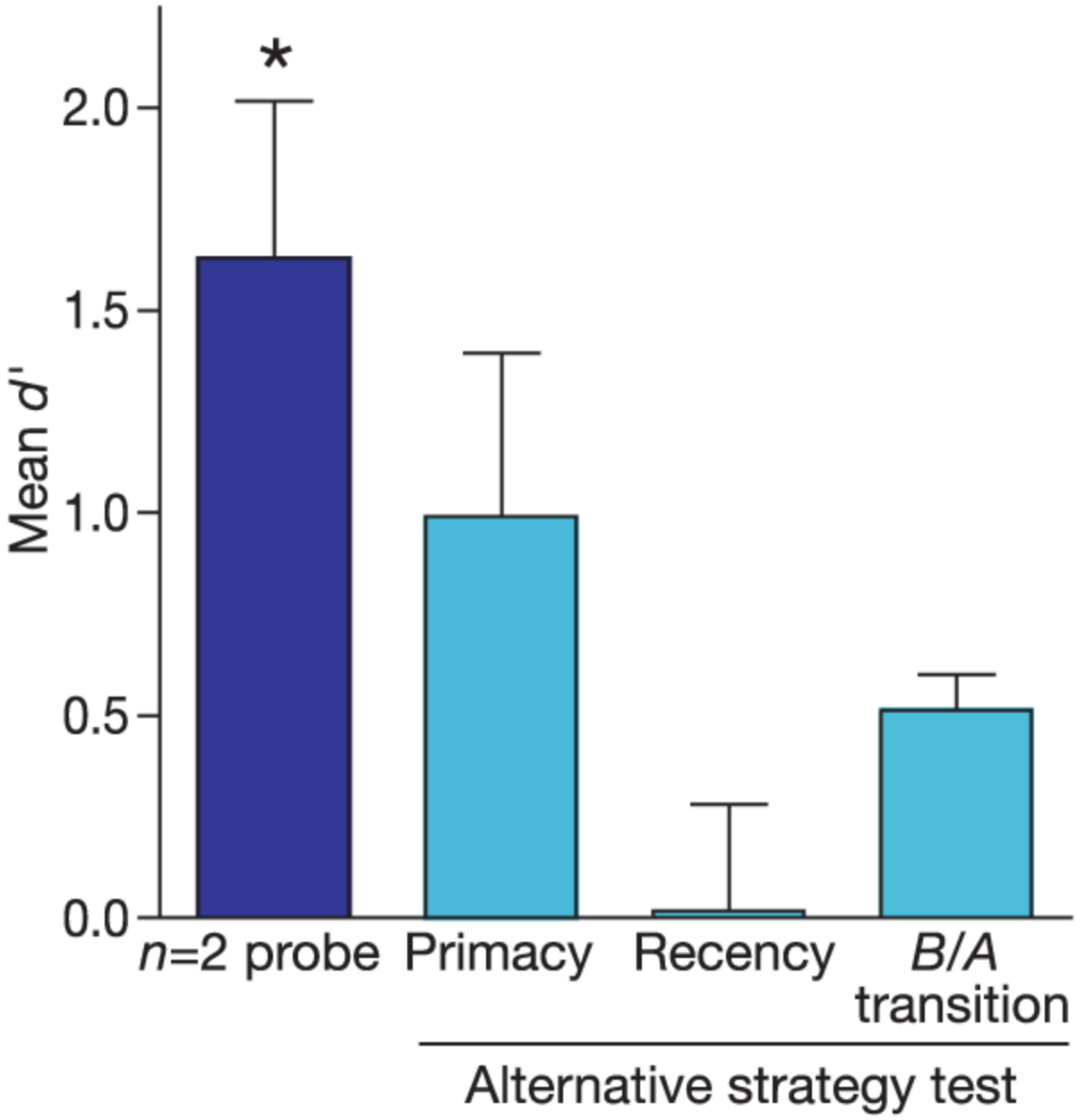

Figure 4. Agrammatical controls for alternative strategies

Mean *d*' (±s.e.m.) values for comparisons among the *AAAA*, *BBBB*, *ABBA* and *BAAB* agrammatical stimuli, to rule out the use of alternate solution strategies. For primacy (see main text), *AAAA* and *ABBA* should be classified similarly to new *n* = 2 CFG and FSG patterns, respectively, presented during the same probe sessions (Methods). For recency (see main text), *BBBB* and *BAAB* should be classified similarly to new *n* = 2 CFG and FSG patterns, respectively. If starlings are listening for the presence of a *B/A* motif transition (see text), then the *d*' value comparing *BAAB* and *ABBA* to *AAAA* and *BBBB* should be similar to that for new *n* = 2 CFG and FSG patterns. *d*' for the new *n* = 2 CFG and FSG probe stimuli (dark blue) was significantly higher than that for all three control comparisons (light blue; asterisk indicates $P < 0.05$ for all cases, paired *t*-test).