Can squirrel monkeys learn an AB^nA grammar?
A re-evaluation of Ravignani et al. (2013)

Stefano Ghirlanda

Preprint of May 2, 2017

Abstract

Ravignani et al. [2013] habituated squirrel monkeys to sound sequences conforming to an AB^nA grammar, then tested them for the ability to identify novel grammatical sequences as well as non-grammatical ones. Although they conclude that the monkeys "consistently recognized and generalized the sequence AB^nA," the data indicate very poor generalization. Pattern grammaticality accounted for at most 6% of the variance in responding. In addition, the statistical significance of results depends on specific choices of data analysis (dichotomization of the response variable and omission of certain data points) which appear to have a weak rationale. I also suggest that the task used by Ravignani et al. [2013] may be fruitfully analyzed as an auditory sequence discrimination task that does not require specific proto-linguistic abilities.

1 Introduction

Within a wider program of research aimed at charting the evolution of linguistic abilities, Ravignani et al. [2013] tested whether squirrel monkeys (*Saimiri sciureus*) can detect violations of an AB^nA grammar. The grammar was instantiated using sequences of low and high pitch sounds. Grammatical sequences were composed of a variable number of sounds of very similar pitch (the *B*s), sandwiched between two other sounds (the *A*s), very similar in pitch to each other, but very different from the *B*'s (see section 3.3 for details). The authors first habituated the monkeys to sequences with structure *ABA*, *ABBA*, and *ABBBA*, by playing these sequences for a total of 360 times over two days. Let me indicate, collectively, sequences with structure *ABA*, *ABBA*, or *ABBBA* as $AB^{1,2,3}A$ sequences. Following habituation, the authors conducted two tests to ascertain whether the animals would:

1. Show habituation to *other* grammatical sequences.

2. Show lack of habituation to non-grammatical sequences, such as *BA* or *ABB*.

Such a pattern of behavior, if found, would indicate that the monkeys could generalize the AB^nA structure heard during habituation to novel sequences. In Test 1, the animals listened to grammatical sequences (both habituated and novel) and non-grammatical sequences (all novel) composed of the same sounds heard during habituation. In Test 2, the role of high and low pitch tones was

1

reversed. That is, whereas *A* and *B* had signified, respectively, low and high pitch sounds prior to Test 2, the opposite was true in Test 2.

According to Ravignani et al. [2013], both Test 1 and 2 indicated that squirrel monkeys can detect the presence or absence of AB^nA structure in novel sequences, including novel pitch patterns. Here, I show that this claim rests on several details of the authors' analysis, which are necessary to obtain conventionally "significant" results ($p < 0.05$) in both Test 1 and Test 2. Namely, significance is attained only when using a dichotomized response measure as well as specific criteria of data selection. Moreover, in Test 1, significance requires including responses to $AB^{1,2,3}A$ sequences that were very similar (sometimes indistinguishable) to sequences used during training. In Test 2, significance requires excluding responses to *AB* and *BA* sequences. The rationale for these choices, however, appears weak. I conclude with a Discussion that places the work of Ravignani et al. [2013] in the broader perspective of working memory research and proto-linguistic abilities in animals.

2 Materials and Methods

I acquired the data posted alongside the original article at http://rsbl.royalsocietypublishing.org/content/9/6/20130852 and recast them as a table with the format displayed in Table 1. All analyses were performed with R, version 3.3.1 [R Core Team, 2016]. The R code and the reformatted data are included as supplementary material.

Test	Subject	Sequence	Response	Grammatical
1	Pi	BBBA	1	No
1	An	BA	0	No
1	Mo	ABBBBA	2	Yes
1	Pi	ABBBA	0	Yes
1	An	ABBBB	0	No
1	Ti	BBBA	1	No
2	Ch	BBBBA	1	No
2	Mo	ABBB	1	No
2	Ch	ABB	1	No
2	Co	ABBA	1	Yes

Table 1: A sample of reformatted experimental data. The Response column indicates how many times a subject was observed orienting toward a speker that was playing the test sequence.

3 Results

3.1 Replication of results in Ravignani et al. [2013]

To ascertain that data acquisition did not introduce errors, this section reproduces the main analyses in Ravignani et al. [2013]. It also serves to summarize the original results, in order to understand how they are affected by the factors discussed later. Responding to grammatical and non-grammatical sequences in Tests 1 and 2 is displayed in Table 2. Results for Test 1 match the height of the bars in Figure 2 of Ravignani et al. [2013]. Results for non-grammatical sequences

in Test 2 are, however, are lower than in the original figure. The reason is that original analysis excluded responses to sequences *AB* and *BA* (see section 3.4 for furhter analysis of this point). Excluding these sequences, I obtain Table 3, whose content matches the height of the bars for Test 2 in the original Figure 2.

Test	Grammatical	Fraction of trials with > 0 responses
1	No	0.77
1	Yes	0.60
2	No	0.78
2	Yes	0.62

Table 2: Results of Tests 1 and 2, all data included.

Test	Grammatical	Mean
2	No	0.83
2	Yes	0.62

Table 3: Results of Test 2 after exclusions of stimuli *AB* and *BA*.

I also reproduced the original statistical results. The authors performed two paired *t*-tests, comparing for each subject the fraction of grammatical and non-grammatical sequences to which at least one response was recorded. For these tests, my analysis yields Table 4, which matches the results by Ravignani et al. [2013] apart from these authors reporting $t = 4.64$ for Test 2.

Lastly, Ravignani et al. [2013] conducted an ANOVA of the whole data set, with fraction of trials with a response as dependent variable and grammaticality and test as independent variables. The results showed a significant effect of grammaticality, which is reproduced in Table 5. Responses to stimuli *AB* and *BA* from both Tests 1 and 2 were excluded from this analysis (see section 3.4 below).

3.2 Dichotomization of the dependent variable

Ravignani et al. [2013, Supplementary Material] counted the number of times monkeys turned the head toward a speaker, within 7 s from stimulus onset. Before performing the analyses reproduced above, they dichotomized this measure so that 0 head turns was "no response" and ≥ 1 head turns was "response." Generally, dichotomization is not advised as it leads to loss of information [MacCallum et al., 2002, DeCoster et al., 2009]. In this case, it may be argued that more head turns indicate greater surprise, which appears relevant to infer whether the monkeys detected grammatical violations. Crucially, repeating the same analyses, but without dichotomizing responses, yields no significant differences between grammatical and non-grammatical sequences. The results for Test 1 and Test 2 are in Table 6. Results for the ANOVA of the whole data set also become non-significant when responses are not dichotomized (cf. Table 5 and Table 7).

Why does dichotmization influence results so markedly? Two features of the data make it so that small changes to data analysis have a large impact. First, sample size is small: 100 responses recorded in Test 1, 61 in Test 2. Given that the observed effect is small (see Discussion), data

Test	t	d.f.	p
1	3.16	5	0.025
2	4.63	3	0.019

Table 4: Results of t tests for Tests 1 and 2, as conducted by Ravignani et al. [2013].

	Sum of Squares	d.f.	F	p
Grammatical	1.7	1	8.33	0.0045
Test	0.0048	1	0.02	0.88
Grammatical:Test	0.0036	1	0.02	0.89
Residuals	27	136		

Table 5: Results of ANOVA for Tests 1 and 2 combined.

analysis is very sensitive to details. Second, differences in responding between gramatical and non-grammatical sequences were very slight for trials that received more than one response (Fig. 1), thus taking into account multiple responses reduces the differences between grammatical and non-grammatical sequences.

3.3 Inclusion of training sequences in analysis

Sequences with $AB^{1,2,3}A$ structure were included both during habituation training and in Test 1. Ravignani et al. [2013] included responses to these sequences in their analysis of Test 1 results. I have also, so far, included them. It may be argued, however, that it is problematic to assess whether the monkeys learned something about a grammar by using the same stimuli employed to demonstrate the grammar (or very similar ones, see below). This point is crucial in the present case, because the t test for Test 1 becomes non-significant if data from $AB^{1,2,3}A$ sequences are excluded from analysis, even when a dichotomized response is used ($t = 0.96$, d.f. $= 5$, $p = 0.38$). The $AB^{1,2,3}A$ sequences played during Test 1, however, were not (all) identical to the habituation sequences, because each time a sequence was played, the constituent A and B sounds were sampled randomly from two sets of sounds. Thus determining whether we should include or exclude test responses from these sequences requires some analysis.

During habituation, the monkeys heard a total of 720 A sounds and 1680 B sounds. My main point is that, although many test sequences were novel during Test 1, all of them were composed of sounds heard during habituation training, and thus may have been perceptually very similar to habituated sequences. If this is true, we expect significant stimulus generalization [Ghirlanda and Enquist, 2003] that can be mistaken for generalization based on abstract features. To evaluate this possility, I have summarized the charateristics of A and B sounds in Table 8. It appears that many of the A sounds were probably not distinguishable by the monkeys, as the frequency difference between two A sounds could be as small as 0.15 JNDs (just noticeable difference: the difference that can be detected with 50% probability). In fact, the 44 A sounds spanned a range of about six JNDs. B sounds were more easily discriminable, lying about one JND apart. B sounds, however, were played 1680 times during habituation training, so that stimulus generalization appears likely even when considering that these sounds were more distinct. In conclusion, it seems prudent to exclude $AB^{1,2,3}A$ sequences from Test 1 because of possible stimulus generalization of habituation.
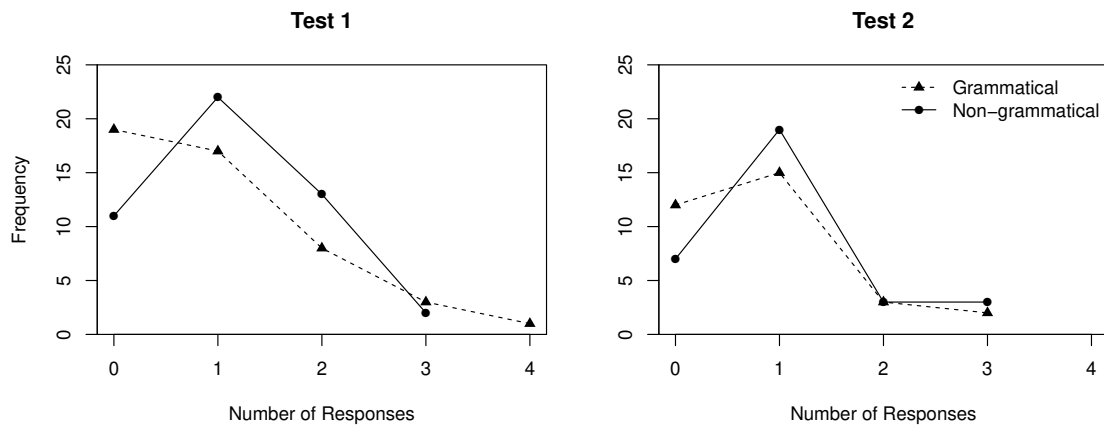
4

Figure 1: Distribution of responses to grammatical and non-grammatical sequences in Tests 1 and 2.

Test	t	d.f.	p
1	1.66	5	0.16
2	1.00	3	0.39

Table 6: Results of t tests for Tests 1 and 2, when responses are not dichotomized. Compare Table 4.

3.4   Omission of data from analysis

Ravignani et al. [2013] excluded from the analysis of Test 2 results the responses elicited by two stimulus sequences, *AB* and *BA*. All analyses reported so far have honored this choice, but there are reasons to revisit it. The rationale offered for excluding responses to *AB* and *BA* was that this pair of sequences, being symmetrical with respect to the exchange of low and high pitch tones introduced in Test 2, had also been presented in Test 1. Thus, the authors argued, the monkeys could have habituated to these sequences during Test 1, confounding the outcome of Test 2. This possibility, however, can be rejected confidently. First, habituation typically proceeds over many trials, while *AB* and *BA* were presented only once to each animal in Test 1. Second, the data do not show any habituation to these sequences between Tests 1 and 2. In both tests, both *AB* and *BA* received an average of 1 head turn per presentation.

This point is relevant becase, without excluding *AB* and *BA*, even the original data analysis yields non-significant results in Test 2, as pointed out by Ravignani et al. [2013] themselves ($t = 1.99$, d.f. $= 3$, $p = 0.14$). The result stays non-significant if responses are not dichotomized ($t = 0.83$, d.f. $= 3$, $p = 0.47$).

4   Discussion

Ravignani et al. [2013] concluded that "Squirrel monkeys consistently recognized and generalized the sequence AB^nA" and that they "are sensitive to abstract dependencies of different lengths and can generalize to new lengths and auditory parameters of the stimuli" ("dependency" here indicates that grammatical sequences were bound to have identical first and last elements). I argued above that these conclusions should be tempered in light of statistical and methodological considerations.

5

	Sum of Squares	d.f.	F	p
Grammatical	1.7	1	2.21	0.14
Test	0.34	1	0.45	0.5
Grammatical:Test	0.008	1	0.01	0.92
Residuals	103	136		

Table 7: Results of the ANOVA in Table 5 when responses are not dichotomized.

Sound class	Number of sounds	Frequency range (Hz)	Interval (Hz)	JND (Hz)
A	44	1800–2200	9	~ 70
B	44	9000–11000	45.5	~ 50

Table 8: Characteristics of sounds used to assemble stimulus sequences. "Interval" refers to the difference between adjacent sounds. "JND" refers to the just-noticeable difference in the considered frequency range, i.e., the difference that is detected with 50% probability [based on Wienicke et al., 2001].

Even setting these aside, however, the data show a very modest generalization ability. Table 9 summarizes results across Test 1 and Test 2 both with dichotomized responses ("Fraction of Trials") and non-dichotomized responses ("Mean Responses"). For the sake of the argument, data from sequences *AB* and *BA* is excluded from Test 2. Let us consider the case in which the effect of grammaticality appears greatest: dichotomized responses in Test 2. Monkeys responded to 83% of non-grammatical sequences and to 62% of grammatical ones. Thus each (dichotomized) response had a probability of $83/(83+62) = 0.57$ of originating from a non-grammatical sequence. If monkeys had responded randomly, this probability would have been very similar, at 0.5. Another way of noting that the monkeys' responses to grammatical and non-grammatical sequences differed only very slightly is to refer to the ANOVA in Table 5, in which sequence grammaticality accounts for only about 6% of the variance in responding.

Ravignani et al. [2013] interpreted their results using the framework of formal language theory [Fitch and Friederici, 2012]. Additionally, we can relate their results to research on animal working memory, which has employed a variety of broadly similar tasks [Roberts and Kraemer, 1981, Honig and Thompson, 1982, Kendrick et al., 1986]. For example, in the "successive matching to sample" paradigm, animals are trained to perform a response if a second stimulus is identical to a first one [Nelson and Wasserman, 1978, Honig and Thompson, 1982]. The current task is similar, in that all habituated sequences could be identified by retaining in memory the first sound, and then checking whether the last was the same or not [Berwick, 2016]. Intervening sounds could be ignored (although we may expect them to make the task harder by introducing interfering memories). From this perspective, the question is not whether squirrel monkeys can conceptualize grammars with long-range dependencies, but whether their working memory span is sufficient to compare the first and last stimuli in a sequence. Estimates of memory span for primates range from about 10 s to several minutes [Roberts and Kraemer, 1981, Lind et al., 2015], while the stimuli used by Ravignani et al. [2013] lasted only 0.75–1.75 s. Comparing the first and last stimuli appears thus within reach of squirrel monkey memory. Indeed, the monkeys' performance was unimpressive compared to what other species have achieved in similar (although not identical) tasks [Honig and Thompson, 1982, van Heijningen et al., 2013, Seki et al., 2013, Lind et al., 2015]. A possible reason is the lower motivational salience of the habituation task, compared to instrumental tasks

6

Test	Grammatical	Fraction of Trials	Mean Responses
1	No	0.77	1.12
1	Yes	0.60	0.96
2	No	0.83	1.08
2	Yes	0.62	0.84

Table 9: Summary of Test 1 and Test 2 results. Test 2 results exclude responses to sequences *AB* and *BA*.

with food reinforcement. In addition, Ravignani et al. [2013] provided "only" 360 training trials, while good performance on working memory tasks often requires thousands of trials [Ghirlanda et al., 2017]. In other words, it is possible that habituation training was not very successful.

There would be at least three potential advantages if grammar learning studies made greater contact with working memory studies. First, adopting more similar methods would make results comparable with a rich literature [Honig and Thompson, 1982, Kendrick et al., 1986]. Second, training techniques could be improved to yield a clear baseline of responding to training stimuli, which in turn would improve the assessment of responding to test stimuli that is so important for grammar learning studies. Lastly, the working memory literature contains many ideas about the content and mechanisms of animal memory [Honig and Thompson, 1982, Kendrick et al., 1986, Ghirlanda et al., 2017]. Some of these may be relevant to grammar learning tasks, as seen above. A more systematic consideration of existing theories of animal working memory would increase the ability of grammar learning studies to inform us about proto-linguistic abilities that go beyond memory abilities that are widespread among species.

5 Acknowledgments

References

Robert C. Berwick. Monkey business. *Theoretical Linguistics*, 42(1–2):91–95, 2016.

Jamie DeCoster, Anne-Marie R. Iselin, and Marcello Gallucci. A conceptual and empirical examination of justifications for dichotomization. *Psychological Methods*, 14(4):349, 2009.

W Tecumseh Fitch and Angela D Friederici. Artificial grammar learning meets formal language theory: an overview. *Phil. Trans. R. Soc. B*, 367(1598):1933–1955, 2012.

S. Ghirlanda and M. Enquist. A century of generalization. *Animal Behaviour*, 66:15–36, 2003.

S. Ghirlanda, J. Lind, and M. Enquist. Memory for stimulus sequences: A divide between humans and other animals? *Royal Society Open Science*, 2017. Revised manuscript submitted.

Werner K. Honig and Roger K. R. Thompson. Retrospective and prospective processing in animal working memory. In G. H. Bower, editor, *Psychology of Learning and Motivation*. Academic Press, New York, 1982.

Donald F Kendrick, Mark E Rilling, and Maurice Ray Denny. *Theories of animal memory*. Psychology Press, 1986.

J. Lind, S. Ghirlanda, and M. Enquist. Animal memory: A review of delayed match-to-sample data from 25 species. *Behavioral Processes*, 117:52–58, 2015.

Robert C. MacCallum, Shaobo Zhang, Kristopher J. Preacher, and Derek D. Rucker. On the practice of dichotomization of quantitative variables. *Psychological Methods*, 7(1):19, 2002.

Keith R. Nelson and Edward A. Wasserman. Temporal factors influencing the pigeon's successive matching-to-sample performance: Sample duration, intertrial interval, and retention interval. *Journal of the Experimental Analysis of Behavior*, 30:153–162, 1978.

R Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2016.

Andrea Ravignani, Ruth-Sophie Sonnweber, Nina Stobbe, and W. Tecumseh Fitch. Action at a distance: dependency sensitivity in a New World primate. *Biology Letters*, 9(6):20130852, 2013.

William A Roberts and Philipp J Kraemer. Recognition memory for lists of visual stimuli in monkeys and humans. *Animal Learning & Behavior*, 9(4):587–594, 1981.

Yoshimasa Seki, Kenta Suzuki, Ayumi M Osawa, and Kazuo Okanoya. Songbirds and humans apply different strategies in a sound sequence discrimination task. *Frontiers in psychology*, 4, 2013.

Caroline A. A. van Heijningen, Jiani Chen, Irene van Laatum, Bonnie van der Hulst, and Carel ten Cate. Rule learning by zebra finches in an artificial grammar learning task: which rule? *Animal Cognition*, 16(2):165–175, 2013.

A. Wienicke, U. Häusler, and U. Jürgens. Auditory frequency discrimination in the squirrel monkey. *Journal of Comparative Physiology A*, 187(3):189–195, 2001.