Context-freeness Revisited

Willem Zuidema (zuidema@uva.nl)
Institute for Logic, Language and Computation, University of Amsterdam
PO Box 94242, 1090GE Amsterdam, the Netherlands

Abstract

A series of papers have appeared investigating the ability of various species to learn context-free languages. From a computational point of view, the experiments in this tradition suffer from a number of problems concerning the stimuli used in the training phase of the experiments, the controls presented in the test phase of the experiments, and the motivation for and the conclusions drawn from the experiments. This paper discusses in some detail the problems with the existing work in this domain before presenting a new design for this type of experiments that avoids the problems identified in existing studies. Finally, the paper presents results from a small study demonstrating the benefits of the new design.

Keywords: Context-free languages, formal language theory, artificial language learning, animal experiments

Introduction

Since the publication of (Fitch & Hauser, 2004), a small but highly visible literature has emerged investigating the ability of various species to learn and process a context-free language (e.g., Friederici, 2004; Perruchet & Rey, 2004; Gentner, Fenn, Margoliash, & Nusbaum, 2006; Hochmann, Azadpour, & Mehler, 2008; van Heijningen, de Visser, Zuidema, & ten Cate, 2009; Abe & Watanabe, 2011; ten Cate & Okanoya, 2012). It is not difficult to see why the questions addressed in this literature appeal to a wide audience: the grammars generating context-free languages are context-free by virtue of their ability to generate hierarchical structures and to implement center-embedding. Hierarchy and centerembedding are, since (Chomsky, 1957), widely recognized to be hallmark features of human language. Hence, experimentally establishing whether non-human animals can handle a context-free language[1] seems to address a prime candidate in the search for uniquely human, and perhaps uniquely linguistic, cognitive skills.

However, on a closer look, there are many problems with this literature, and almost a decade of investigation and debates have not brought the clarity about this issue that we might have hoped for. In this paper I will first discuss in some detail the problems with the existing work in this domain before presenting a new design for this type of experiments. I will present results of a small experiment with this design, that show it is workable. For lack of space, I will not review elementary formal language theory here; see (O'Donnell, Hauser, & Fitch, 2005) for an introductory and (Jäger & Rogers, 2012) for a more advanced discussion of the formal background of the experiments discussed here.

[1]Throughout this paper I will use the phrase "a context-free language" as denoting a member of the subset of the context-free languages that is not also in the set of regular languages.

Problems with the experimental record

From a computational point of view, the experimental record suffers from a number of problems concerning the stimuli used in the training phase of the experiments, the controls presented in the test phase of the experiments, and the motivation for and the conclusions drawn from the experiments.

The first major problem is a lack of clarity about which ability is really investigated: the ability to *implement* a context-free language, the ability to *learn* a context-free language, or a *preference* for selecting a context-free strategy from the set of strategies adequate for solving the task. Much of the rhetoric seems to be about the ability to implement, but all existing experiments that I am aware of really at best address the weaker hypothesis that non-human animals lack the human preference.

This problem is exacerbated as all existing studies allow a great deal of ambiguity in the training phase about which strategies are adequate. Some ambiguity is unavoidable: all real-world experiments can only present a subset of the infinite stringsets that make up context-free languages, leaving the learner fundamentally uncertain about whether or not sub- or supersets of the intended context-free language are the target (see figure 4). Moreover, studies using familiarization/habituation paradigms can only present positive stimuli in the training phase. However, in a reinforcement paradigm some ambiguity is avoidable, but existing studies using such a paradigm fail to provide learners with the information that some plausible alternatives are not intended. For instance, (Gentner et al., 2006) presented their starlings with in the order of 300000 stimuli with positive and negative feedback to learn to distinguish A^nB^n from $(AB)^n$, but not with a single stimulus that would help the birds exclude A^nB^m. If the question we want to ask is whether these birds can learn the context-free language at all, it would be better to avoid unnecessary ambiguity about the task (**desideratum 1**).

A second unclarity in existing work comes from unnecessary variation in the syllables of (song) elements used to compose the stimuli. Thus, when testing whether subjects can learn A^nB^n, all studies I am aware off use multiple instances of A's and B's. This means the subjects are really confronted with two tasks at the same time: the task to *categorize* a_1, a_2, ... as instances of class A, and the task to learn sequencing rules. While the interaction between categorization and sequence learning is certainly interesting, this interaction has in fact not been explicitly addressed in this paradigm. In most studies the categorization task is made rather trivial because A's and B's are carefully selected to be acoustically very similar within one category and very dissimilar between categories. In these studies the variation in stimuli probably has

little impact on the results and just makes describing the experiments unnecessarily complicated; in other cases, it introduces confounds. It would be better, therefore, to avoid these complications and start with experiments using an alphabet with just 2 items: a and b (**desideratum 2**).

A third major problem with the existing literature concerns inadequate controls (see also Beckers, Bolhuis, Okanoya, & Berwick, 2012). (Fitch & Hauser, 2004) present no data on controls for alternative strategies (although the supplementary material states – without presenting details – that various alternative explanations have been controlled for). Unlike many other experiments, (Gentner et al., 2006) did test a number of these alternative strategies, and presented results that seemed to exclude all except for the most "heavily contrived" finite-state grammar hypotheses. It turns out that even their quite elaborate efforts to control for various alternative strategies are insufficient, as I will discuss below. It would seem necessary, therefore, to work out better ways to evaluate the plausibility of alternative explanations for the results (**desideratum 3**).

A case-study: Gentner et al. 2006

To make these problems concrete, I will here discuss them in the context of (Gentner et al., 2006). This is not because this paper has more methodological problems than others; on the contrary, in fact, this paper probably represents one of the most serious efforts to control for alternative explanations among the experimental papers in this domain. As I will show below, however, the results from this study have nevertheless little to say about the ability or inability of song birds to learn a context-free language.

Training Gentner et al. studied whether starlings (*Sturnus vulgaris*) are able to learn a context-free language. As in many other studies, the stimuli in this experiment were strings of elements that fall into two easily distinguishable categories, A and B, each with a small number of members, i.e. $A = \{a_1, a_2, a_3, \ldots, a_8\}$ and $B = \{b_1, b_2, b_3, \ldots, b_8\}$. Gentner et al. extracted these stimuli from the starling's own song, where the A's were "rattle" motifs and the B's were "warble" motifs. Stimuli in the training phase consisted of strings of length 4 from two patterns[2]: (i) $(AB)^n$ and (ii) A^nB^n. I will refer to string sets defined by these patterns as the FINITE-STATE-0 and the CONTEXT-FREE language (the 0 indicating that this is just the first of many finite-state languages that I will consider).

The birds were trained in a go-nogo operant conditioning procedure to respond selectively to stimuli from one or the other pattern. In the experiment, birds did indeed learn to distinguish the stimuli sets, at levels far exceeding chance, also when new A- and B-category elements were used. This in itself is not enough to prove context-freeness, as Gentner et al note. For instance, the two groups of birds could have internalized +FINITE-STATE-0 and -FINITE-STATE-0 instead[3]. I.e., they could do with a model for the finite-state stimuli set, and only accept/reject stimuli that do not/do conform to it. Or, because the string length is set to 4, A^nB^n is indistinguishable from A^2B^2 (which, again, doesn't need context-free power to be recognized). Worse even: there are many other alternative strategies to distinguish the training stimuli-sets, the simplest of which are based on detecting specific element-to-element transitions, or memorizing the beginning or end of strings. For instance, the BA transition, and the AB beginning, are diagnostic, because they both only occur in the +FINITE-STATE-0 set.

If one could show, in the test phase, that the birds have learned a context-free language, this ambiguity in the training phase is not a problem. However, if the birds turn out to choose one of the simpler strategies that also suffice to distinguish the two classes, we are left almost empty-handed. We cannot make plausible, then, that birds cannot learn context-free language, because we haven't tried very hard to force them to.

Testing In the test phase, Gentner et al did consider a relevant set of alternative strategies.

The first test is whether subjects generalize from A^2B^2 to a larger subset of A^nB^n. Of course, in formal language theory the language A^nB^n contains an infinite number of strings, where n can be any integer. Gentner et al. argue, quite reasonably, that in an experimental setting we should be concerned about whether subjects generalize to unseen n. (This is completely analogous to the use of formal language theory in the study of natural language: if we can demonstrate the right generalization mechanisms on necessarily finite data, we can reason about an infinite competence under a hypothetical lifting of performance constraints.) Gentner et al. report, for birds trained with A^2B^2, a strong preference for A^3B^3 and A^4B^4 strings over $(AB)^3$ and $(AB)^4$ respectively (and an inverse preference for birds trained on $(AB)^2$). This rules out the -FINITE-STATE-0 (or +FINITE-STATE-0) strategy.

There remain, however, still many alternative hypotheses that predict successful discrimination of A^nB^n and $(AB)^n$ strings. It is useful to define the following simple, but effective strategies for positively responding to the +CONTEXT-FREE stimuli[4]:

+ANBN: A^nB^n, with $n \geq 1$

+ANBM: A^+B^+, the set of strings that consist of 1 or more A's followed by 1 or more B's;

-BIGRAM-BA: $\cdot^* BA \cdot^*$, strings containing transition BA;

[2] I use a conventional shorthand notation for sets of strings of a given pattern, where A's and B's indicate any elements from these classes, X^n indicates n repetitions of X, and brackets are used to disambiguate the scope.

[3] A strategy is defined by a PATTERN, written in smallcaps, and a $+$ or a $-$ in front of it; the $+$ indicates that strings that conform to the pattern are treated as positive stimuli; the $-$ indicates that strings that conform to the pattern are treated a negative stimuli.

[4] I will use more or less standard regular expression notation, where a dot $\cdot$ means any symbol, * means repeated any number (≥ 0) of times, and $^+$ means repeated any number (≥ 1) of times.

	A2B2	A3B3	A4B4	AB2	AB3	AB4	A1B3	A3B1	A2B3	A3B2	A4	B4	ABBA	BAAB
+ANBN	1	1	1	0	0	0	0	0	0	0	0	0	0	0
+ABN	0	0	0	1	1	1	0	0	0	0	0	0	0	0
+ANBM	1	1	1	0	0	0	1	1	1	1	0	0	0	0
+AA-PRIMACY	1	1	1	0	0	0	0	1	1	1	1	0	0	0
+BB-RECENCY	1	1	1	0	0	0	1	0	1	1	0	1	0	0
+AB-RECENCY	0	0	0	1	1	1	0	1	0	0	1	0	0	1
+AA-BIGRAM	1	1	1	0	0	0	0	1	1	1	1	0	0	1
+BA-BIGRAM	0	0	0	1	1	1	0	0	0	0	0	0	1	1
+BB-BIGRAM	1	1	1	0	0	0	1	0	1	1	0	1	1	0
$+\cdot\{1,4\}$	1	0	0	1	0	0	1	1	0	0	1	1	1	1
$+\cdot\{1,6\}$	1	1	0	1	1	0	1	1	1	1	1	1	1	1

Table 1: Predicted response of various hypothesized pure strategies to probe stimuli. See the main text for descriptions of the top 9 strategies; the bottom two strategies check the length of a string, and accept strings up to length 4 and 6 respectively.

+BIGRAM-AA: $\cdot^* AA \cdot^*$, strings containing transition *AA*;
+BIGRAM-BB: $\cdot^* BB \cdot^*$, strings containing transition *BB*;
-PRIMACY-AB: $AB \cdot^*$, the set of strings that start with *AB*;
+PRIMACY-AA: $AA \cdot^*$, the set of strings that start with *AA*;
-RECENCY-AB: $\cdot^* AB$, the set of strings that end with *AB*;
+RECENCY-BB: $\cdot^* BB$, the set of strings that end with *BB*;

Any of these strategies (together with their complements when considering the birds that should NOT respond to $A^n B^n$) suffices to distinguish positive from negative samples in the experimental set-up (and all listed alternative strategies are in the finite-state class). But these nine hypotheses do make different predictions on the behavior of the subjects for previously unseen patterns. For instance, the +PRIMACY-AA strategy classifies all of the +CONTEXT-FREE stimuli as positive, but in addition, for string length 4, also includes AAAA, AAAB, AABA. Table 1 gives for (the + variety of) each of these strategies the predicted response (1 is a GO-response, 0 is a NOGO-response).

There are, in fact, still many other alternative strategies that we could consider, such as memorizing non-adjacent pairs (e.g., $A \cdot B \cdot^*$), or requiring a specific number of a particular transition (e.g., requiring two *AB* transitions, as in $B^* A^+ B^+ A^+ B^+$, or exactly one, as in $A^+ B^+ A^*$). Gentner et al. appeal, quite reasonably again, to considerations of parsimony to ignore such alternatives.

To rule out the 9 remaining alternative explanations, Gentner et al. presented birds with a number of diagnostic strings. For instance:

- AAAB, which is *incorrectly* predicted to give a positive response by -FINITE-STATE-0, -BIGRAM-BA, +BIGRAM-AA, and +PRIMACY-AA;

- BBBB, which is *incorrectly* predicted to give a positive response by -FINITE-STATE-0, -BIGRAM-BA, +BIGRAM-BB, -PRIMACY-AB, -RECENCY-AB and +RECENCY-BB.

In an experimental setup, however, checking these predictions needs to be buffered to unavoidable noise in the data. (It would be unreasonable to reject an hypothesis based on a single unexpected classification by a bird.) Hence, we need to use statistics, but how statistical methods for data analysis are combined with formal language theory is a non-trivial issue that both theoreticians and experimentalists have so far largely ignored. (Even the review by (Jäger & Rogers, 2012), which presents a major effort to bridge formal language theory and artificial language learning experiments, ignores this issue).

Gentner et al. chose to use the *d'*-statistic, which is a measure for discrimination between stimuli classes that corrects for response bias (the tendency to prefer a GO or a NOGO-response regardless of the stimulus). They show that the *d'* between AAAA and ABBA is significantly lower than the *d'* between A2B2 and AB2, and argue this rules out the +AA-PRIMACY strategy. Similarly, they find lower *d'* for BBBB vs BAAB, and for BAAB/ABBA vs. AAAA/BBBB, and argue this rules out +BB-RECENCY and -BA-BIGRAM. Similar analyses can be given for the remaining alternative strategies.

However, it turns out that this approach for ruling out alternatives is only valid if the population is homogeneous – all members follow the same strategy – and if each individual follows a *pure* strategy. If individuals or populations can mix multiple strategies , Gentner et al.'s method leads to invalid conclusions. The data of Gentner et al., reproduced as the blue bars in figure 2, clearly show that the assumption of pure strategies is false: the *d'*-statistic for the $A^n B^n$ vs. AB^n contrast decreases with increasing *n*, and the *d'*'s for the primacy and recency strategies differ significantly. A study on zebra finches, by (van Heijningen et al., 2009), reported major individual differences between birds, further strengthening the case against the pure strategy assumption.

Simulated Data To show how the *d'*-statistic can lead to wrong conclusions, I will now present some artificial data that shows a qualitatively similar pattern of *d'*-scores for both a model that involves an underlying context-free grammar (model I: CFG) and a model that is just a mix of finite-state strategies (model II: MIX).

To generate the CFG data, I assume a population where 70% of the individuals have internalized the +ANBN-strategy, 10% follow a strategy to reject long strings

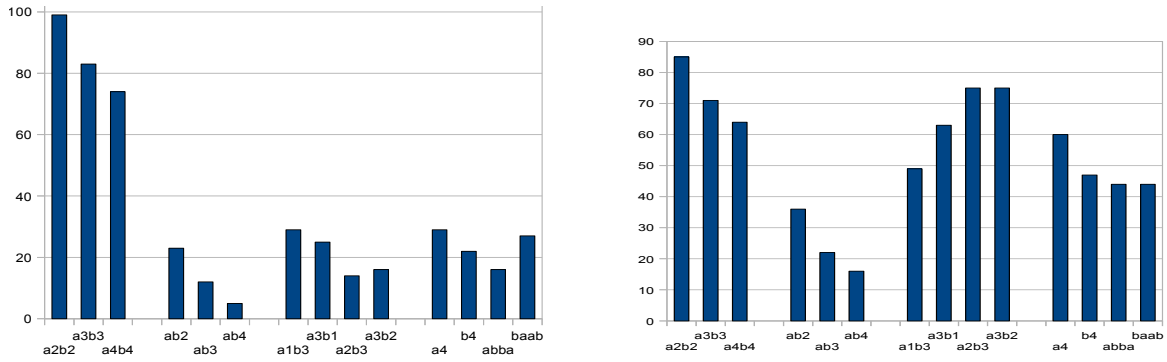

Figure 1: Simulated data, generated from the CFG-model (left) and the MIX-model (right). In both graphs, the first group of 3 bars represent the number of go-responses to A^nB^n-stimuli (out of 100); the second to $(AB)^n$-stimuli; the third group to $A^nB^{m\neq n}$-stimuli; the fourth to the remaining control-stimuli.

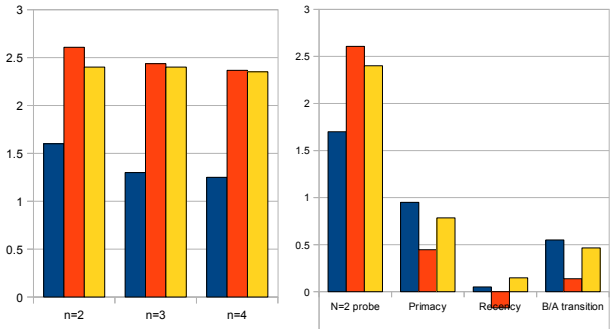

Figure 2: The d'-statistic calculated for the A^nB^n vs. $(AB)^n$ distinction (left) and for various controls (right). *Blue:* Gentner et al, *Red:* CFG, *Yellow:* MIX.

$(+.\{1,6\})$, 10% reject medium and long strings $(+.\{1,4\})$, and 10% randomly choose GO or NOGO. This is equivalent to assuming a mixed strategy with the same proportion, or a combination of intra- and interindividual variation.

To generate the MIX data, I assume a mix of finite-state strategies in the following proportions (roughly based on the findings of (van Heijningen et al., 2009)): 25% +AA-PRIMACY, 15% +BB-RECENCY, and 10% for each of the other strategies in table 1 + RANDOM.

A given mix of strategies defines for each stimulus a specific number f of GO-responses, from a fixed number of 100 presentations. To generate some randomness, I assume each of the f stimuli that should be classified as a GO-response has a fixed probability $p = 0.03$ to receive a NOGO-response, and similarly, that each of the $100-f$ remaining stimuli have a probability p of receiving a GO-response.

Hence, the final number of GO-responses is $g = x + y$, where both x and y are sampled from a binomial distribution:

$$x \sim B(f,p), y \sim B(100-f, 1-p)$$

This generates a dataset as in figure 1 (left) for the CFG model, and (right) for the MIX model (note that the datasets are the result of a single run of the model).

I subsequently calculate the *d'*-statistic in the same way as (Gentner et al., 2006). This statistic is simply the difference between the z-transform of the counts: $d'(x,y) = z(x) - z(y)$ where the z-transform in turn is a way to express the magnitude of the score in terms of how many standard deviations it is away from the mean: $z(x) = (\bar{d} - x)/\sigma_d$ where d is the complete data vector from which x is one value (or the average of several values), and σ_d is the standard deviation over that vector.

Applying these formulas to the contrasts between A^nB^n and $(AB)^n$ for $n \in \{2,3,4\}$ we obtain, in figure 2(left) qualitatively similar results to Gentner et al: significant discrimination for all, but a decrease in discriminability with increasing n. Although unsurprising, this result points to a problem with existing studies that fail to show generalization: this could be due to length effects. It would be better if tests for generalization do not only test on *longer* strings in the test phase than were offered in the training phase (**desideratum 4**).

Applying these formulas to the contrasts between AAAA and ABBA (labeled "primacy" in (Gentner et al., 2006)), BBBB vs BAAB (labeled "recency"), and BAAB/ABBA vs. AAAA/BBBB (labeled "bigram B/A") we observe, in figure 2(right) the exact same pattern of results as Gentner et al reported: overall much lower d'-values than for the baseline, with primacy receiving the second highest score and recency the lowest.

Hence, for both the context-free (CFG) and mix of finite-state strategies (MIX) we see the same pattern of d'-values, showing that when mixed strategies are possible, these values are uninformative about whether or not a context-free language is learned by any individual in the population. Hence, we need better tools to assess which strategies individuals are using and whether there is significant individual variation in a population (**desideratum 5**).

A new design

The problems I discussed with the Gentner et al. study are symptomatic for many studies in this domain. Confusion

about exact goals and methodology are of course typical for the early phase of a new research field. It is now time, however, for an experimental design to emerge that is both methodologically sound and capable of generating useful results. In the following I will present an attempt to give such a design and a first experiment to assess its usefulness. The design follows desiderata 1-4 discussed above:

1. The goal of the design is to test whether or not the subjects can learn the context-free language A^nB^n from the type of data that can be used with animals as well as human infants and adults; i.e., not too long strings, possibly with positive and negative feedback. Some strings are reserved for the test phase only, to assess generalization, but otherwise any training regime is allowed within these constraints. In practice, I choose for a two-stage training phase: a familiarization phase where only positive stimuli are presented, and a feedback phase where positive and negative stimuli are presented with positive and negative feedback. The negative stimuli are not just from $(AB)^n$, but also from other plausible, but incorrect, alternative languages.

2. To make the task as simple and unambiguous as possible, I define the patterns over an alphabet of just two different sounds: *a* and *b*, selected to be short and acoustically clearly distinct.

3. To be able to test for generalization, I reserve 2 values of n for strings from A^nB^n and A^nB^n for the test phase only. I further reserve a number of n,m combinations for strings in $A^nB^{m\neq n}$ for the test phase, to be able to exclude primacy, recency and ANBM-strategies.

4. To make sure the test strings are not much longer than the strings seen at training, I use $n \in \{2,3,5,6\}$ for A^nB^n and $(AB)^n$ strings at training, and $n \in \{3,4,6,7\}$ at test.

The stimuli presented to subjects in the various phases are thus as follows:

Phase	Stimuli
Familiarization	a2b2, a3b3, a5b5, a6b6
Feedback	*Positive:* a2b2, a3b3, a5b5, a6b6
	Negative: ab2, ab3, ab5, ab6, a3b2, a5b4
Test	*Positive:* a3b3, a4b4, a6b6, a7b7
	Negative: ab3, ab4, ab7, a3b2, a4b3, a2b3

Experimental data Ultimately we need a lot of data and new analysis tools to meet desideratum 5 and exclude mixed strategies between and within individuals. However, a first important check on the design is to evaluate whether we can replicate the findings (Hochmann et al., 2008) that humans adults can, *at the population level*, (i) learn to distinguish strings from A^nB^n from several finite-state alternatives, and (ii) generalize to unseen n. We therefore carried out a small experiment with the design above, to test whether its new features 1-4 stand in the way of successful learning.

The experiment was implemented as a simple internet-based applet. Subjects were instructed that they were going to do an experiment that looked a bit like a computer game, where they would have learn an alien language. At the computer screen, subjects were presented with written instructions, and, once they started the game, presented with a space background and UFO's moving over the screen. Subjects were asked to click on disks and listen to the sounds produced. After hearing the sounds they would decide to either shoot the UFO or save the aliens inside. In the familiarization phase (4 exposures to each stimulus), they were told all aliens were 'good aliens' and shooting was disabled. In the feedback phase (1 exposure to each stimulus) feedback was provided in the form of happy or sad face on the screen. In the test phase (2 exposures to each stimulus) no feedback was given.

54 subjects were recruited in the Amsterdam Science Museum (Nemo) in August 2012, and volunteered for the experiment without payment. The experiment last only about 5 minutes per subject. Subject ages ranged from around 10 to around 80; native languages included several major European languages. We thus worked with a very heterogeneous group of subjects and obtained very little data per person. Hence, the experiment was not useful (nor intended) for settling the question of whether human adults can learn a context-free language in such a setup, but to assess whether the experimental design defined above is able to generate useful data.

Figure 3 (left) gives the overall response rates for each of the stimuli presented in the test phase. As can be seen, the response rates for the A^nB^n stimuli on the left are higher than for the A^nB^n stimuli on the right, which in turn are higher than the AB^n stimuli in the middle. All three pairwise between-group difference are highly significant ($p < 0.01$, Kolmogorov-Smirnov test). Crucially, responses to a4b4 are indistiguishable from other positive stimuli, indicating subjects have, at the population level, generalized to unseen n.

These data thus, roughly, replicate earlier results. But can we check for all relevant alternative explanations, including a mix of finite-state strategies, as I argued above would be necessary? Unfortunately, with so little data per subject, we cannot sensibly estimate individual strategies. To get some idea about individual variation I split the data in two based on performance during the feedback phase. 20 subjects were classified as low-performers, with an accuracy in the feedback phase of less than 70% (similar to the criterion used in van Heijningen et al., 2009). The other 34 subjects were classified as high-performers. Figure 3 (right) shows the d' statistic for the three pairwise contrasts. One striking value is a d' of approximately zero for the low performers on the A^nB^n vs. A^nB^m contrast, showing that they did not distinguish between the two classes and clearly had not learned a context-free language.

These data thus suggest that the experimental design presented above is very workable and can be used to obtain useful data to address the question of whether subjects can

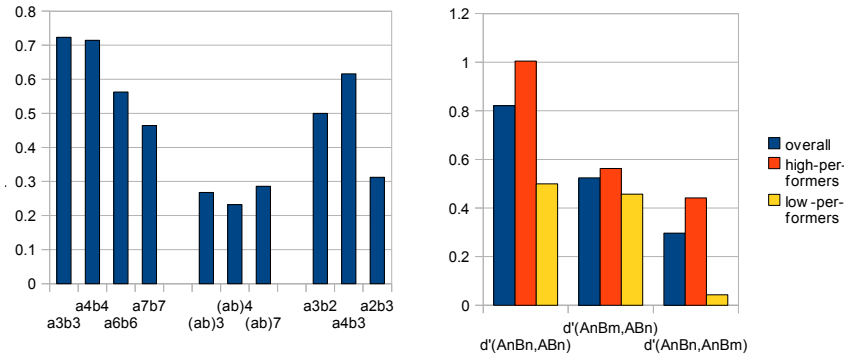

Figure 3: Experimental results.

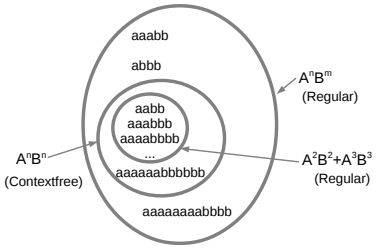

Figure 4: There exist (infinitely) many *regular* subsets and supersets of the *context-free* language A^nB^n (regular languags can be generated by finite-state automata).

learn A^nB^n, provided it is applied in more controlled circumstances and more data per person is gathered. With enough data per subject, we can apply the model selection approach of (van Heijningen et al., 2009) to estimate the most likely strategy for each individual subject.

Conclusions

I have discussed some major problems with existing studies attempting to show that nonhuman animals can or cannot learn a context-free language. In that discussion, I identified four desiderata for a new design of an experiment, and a fifth desideratum for data analysis, to properly address that question. I have shown that these desiderata for the design can be satisfied, and presented some experimental results that suggest there are no major obstacles to apply the new design in animal experiments. I hope the interdisciplinary community that tries to bring formal language theory, artificial language learning and animal cognition experiments together will apply this new design in future experiments, such that the search for uniquely human cognitive skills can be based on a more sound foundation. I have not, in this paper, discussed the difficult question of whether context-freeness is really the most important property to investigate in the search for a biological basis for language (Zuidema, 2013). Even if it is not – and I suspect it isn't – it is essential that the methodological errors in the experimental record on context-freeness get corrected.

Acknowledgments I thank Vanessa Ferdinand, Richard Kunert and Sander Latour for their help in designing the experiment and creating the software for stimulus presentation and data gathering, and the team at Nemo for running the experiment.

References

Abe, K., & Watanabe, D. (2011). Songbirds possess the spontaneous ability to discriminate syntactic rules. *Nature neuroscience*, *14*(8), 1067–1074.

Beckers, G., Bolhuis, J., Okanoya, K., & Berwick, R. (2012). Birdsong neurolinguistics: songbird context-free grammar claim is premature. *Neuroreport*, *23*(3), 139.

Chomsky, N. (1957). *Syntactic structures*. The Hague: Mouton.

Fitch, W. T., & Hauser, M. D. (2004). Computational constraints on syntactic processing in a nonhuman primate. *Science*, *303*(5656), 377-380.

Friederici, A. D. (2004). Processing local transitions versus long-distance syntactic hierarchies. *Trends in Cognitive Sciences*, *8*(6), 245–247.

Gentner, T. Q., Fenn, K. M., Margoliash, D., & Nusbaum, H. C. (2006). Recursive syntactic pattern learning by songbirds. *Nature*, *440*, 1204-1207.

Hochmann, J.-R., Azadpour, M., & Mehler, J. (2008). Do humans really learn anbn artificial grammars from exemplars? *Cognitive science*, *32*(6), 1021–1036.

Jäger, G., & Rogers, J. (2012). Formal language theory: refining the chomsky hierarchy. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1598), 1956–1970.

O'Donnell, T. J., Hauser, M. D., & Fitch, W. T. (2005). Using mathematical models of language experimentally. *TRENDS in Cognitive Sciences*, *9*(6).

Perruchet, P., & Rey, A. (2004). Does the mastery of center-embedded linguistic structures distinguish humans from nonhuman primates? *Psychonomic Bulletin and Review*. (In press)

ten Cate, C., & Okanoya, K. (2012). Revisiting the syntactic abilities of non-human animals: natural vocalizations and artificial grammar learning. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1598), 1984–1994.

van Heijningen, C., de Visser, J., Zuidema, W., & ten Cate, C. (2009). Simple rules can explain discrimination of putative recursive syntactic structures by a songbird species. *Proceedings of the National Academy of Sciences*, *106*(48), 20538–20543.

Zuidema, W. (2013). Language in nature: on the evolutionary roots of a cultural phenomenon. In P. Binder & K. Smith (Eds.), *The language phenomenon* (p. 163-190). Berlin: Springer.