You Can't Get There from Here: On Interpreting Learning Experiments

Constantine Lignos
University of Pennsylvania

You Can’t Get There from Here: On Interpreting Learning Experiments

Abstract

Artificial language learning experiments provide a unique opportunity to observe learning under controlled conditions. We cannot, however, observe what learning strategy participants use; we can only carefully design the language and observe the response. This poses an inference problem that I name "the poverty of the experiment." I use computational learning models to address this inference problem, using data from an artificial grammar learning study (Saffran 2001) in which the authors conclude that participants learned hierarchical structure from distributional cues. Simulations show that that learning hierarchical structure is not required to pass the tests administered in those experiments and that a heuristic learner is the best fit for the observed human performance. Artificial language learning experiments cannot in themselves provide evidence for a particular learning strategy; they must be paired with appropriate modeling work to confirm that an implementation of a proposed learning strategy actually produces the expected results.

You Can't Get There from Here: On Interpreting Learning Experiments

Constantine Lignos*

1 Introduction

In investigating the mechanisms of language of acquisition, we are faced with a difficult induction problem. While we may observe the input to the learner and its performance over time, the method of learning occurs is always unobserved. We must infer the structure of learning mechanisms from incomplete, and sometimes even contradictory, data. While experimental work can provide more control over the input than natural settings, the same problem persists. We may carefully design stimuli and observe participants' responses, but we must infer *how* and *what* participants learned. I name this inference problem the *poverty of the experiment* as it may pose as equally great a challenge for language acquisition researchers as its namesake.

In this paper I discuss the issue that this inference problem poses for interpreting learning experiments. I use an artificial language learning experiment as a case study to demonstrate the difficulty of inferring learning methods from participant performance. I propose that computational learning models can aid in addressing this inference problem and use simple computational learning models to evaluate the claim that participants in the experiments of Saffran 2001 learned hierarchical structures from an artificial language. I find that a simple baseline learner that makes no attempt to discover hierarchical information predicts the observed human performance on the given task better than learning the structure intended by the experimenters. While I do not suggest that this simple model is a useful model of language acquisition, its success demonstrates that we must exercise caution in drawing conclusions about what participants in an experiment learn based on their ability to distinguish grammatical and ungrammatical strings generated from an small artificial language.

More broadly, I argue that experiments in which participants show discrimination of stimuli at greater than chance or control levels *cannot* in themselves provide evidence for a particular learning strategy; they must be paired with appropriate modeling work to confirm that an implementation of a proposed learning strategy actually provides the expected results. More succinctly, one cannot simply leap from a finding of discrimination to a specific learning model: *you can't get there from here*.

2 Background

The issue of the poverty of the experiment is most apparent in studies of artificial language learning. In the artificial language learning paradigm (e.g., Reber 1967), an extension of perceptual learning studies (e.g., Gibson and Gibson 1955) into the language domain, a tightly-controlled artificial language is designed for the purpose of testing what participants are able to learn using limited cues. The control afforded by designing an artificial language allows experimenters to define what should be learned and limit the possible learning strategies that participants may use.

In recent years, interest in artificial language learning studies has been reinvigorated by evidence that young infants can use statistical information to learn properties of artificial language input (Saffran et al. 1996a et seq.). The artificial language learning paradigm has been most famously applied to word segmentation (e.g., Aslin et al. 1998, Johnson and Jusczyk 2001, Lew-Williams et al. 2011, Thiessen and Saffran 2003), but artificial languages have been created to explore learning in many domains, including syntactic structure (e.g., Gomez and Gerken 1999, Morgan and Newport 1981, Reber 1969, Saffran 2001). The ability to learn artificial languages is not restricted to speech input (Saffran 2002) or even to humans (Saffran et al. 2008).

The most useful application of the artificial language learning paradigm may be the ability to manipulate the set of cues available for learning a single language. Focusing on the difference in

*Many thanks to Charles Yang, Robert Frank, and the audience at PLC 36 for their enlightening comments regarding this work.

performance when two different sets of cues are available provides a much simpler hypothesis to test: whether additional cues lead to improvement in participant performance, as opposed to whether the grammar has been learned. In word segmentation tasks, the relative importance of statistical and prosodic cues to segmentation has been heavily studied, evaluating word-level stress (e.g., Johnson and Jusczyk 2001, Thiessen and Saffran 2003), prosodic boundaries (e.g., Shukla et al. 2011), and phonotactic cues (e.g., Mattys et al. 1999). The confluence of cues has been studied in grammar learning tasks as well (e.g., Morgan and Newport 1981).

It is difficult to determine the relevance of artificial language learning studies to acquisition as the structure of artificial languages has little in common with that of natural languages. While carefully constructed languages may prove useful for examining specific cues in isolation, attempts to make artificial language learning experiments more naturalistic have encountered difficulties. For example, in a word segmentation task, if Italian syllables are used instead of artificial language syllables, subjects can successfully discriminate between words and non-words (Pelucchi et al. 2009). However, when the length of words in the stimuli varies (Lew-Williams and Saffran 2012) or words in isolation are added to the input (Lew-Williams et al. 2011) learning can fail.

While valid questions of ecological validity regarding artificial language learning can be raised, for the purpose of this paper I put those objections aside. The question at hand here is how to interpret the performance of participants as evidence of their knowledge of the artificial language they are exposed to. Traditionally, discrimination between items in the language and items not in the language has been used as the criterion for whether subjects successfully "learned" the artificial language presented to them. For example, in the task of syntax learning, participants must discriminate between grammatical and ungrammatical items in the language; in word segmentation, participants must discriminate between words and plausible non-words.

While necessary, discrimination between grammatical and non-grammatical items, is not, however, sufficient to show that participants have successfully learned the target language. For syntactic learning, showing that the grammar of the artificial language and the grammar learned by participants are at least extensionally equivalent would be sufficient demonstration that the grammar was correctly learned. Such a rigorous standard would likely prove impractical in experimental settings. A discrimination metric can verify that a participant has learned *something* about the artificial language, but what the participant has learned may not match what the experimenters intended. For example, in a word segmentation task, it is assumed that the reason that participants are able to discriminate between words and non-words is that they have successfully segmented the speech stream into words. But as Endress and Mehler (2009) demonstrate, it is possible to pass a discrimination task in these experiments by not segmenting utterances at all, instead only learning the surface pattern of transitional probabilities manipulated by the experimenters. Reber (1969) explicitly discusses the potential gap between what the experimenter desires the participant to learn and what the participant actually learns. As discussed further in Section 4, this gap has not been adequately addressed in modern work on artificial grammar learning.

In addition to learning a different representation than intended, it is possible that when faced with a discrimination task participants may learn through altogether different means than the experimenters sought to elicit. Careful experimental design can constrain the possible learning strategies a participant can use; for example, Aslin et al. (1998) and subsequent word segmentation studies match words and non-words in frequency to demonstrate that simply recognizing frequent chunks is not sufficient to complete the task. It is generally impossible, however, to design a practical experiment where only a single learning strategy can succeed. For example, while Gentner et al. (2006) claim that starlings are capable of learning recursive structures, further investigation shows that they do not actually adopt such a generalization, relying on heuristics that fail in more difficult tasks (Van Heijningen et al. 2009). Similarly, while statistical word segmentation studies have suggested the use of transitional probabilities (e.g., Saffran et al. 1996b) as a mechanism for word segmentation based on the design of the stimuli of those studies, Perruchet and Vinter (1998) demonstrate that the same effects can be obtained by chunking without the use of transitional probabilities. Modeling can be used as a part of data analysis to verify that that the proposed learning mechanisms are actually consistent with experimental results. Little work matching experiments with explicitly simulated models has been performed. Frank et al. (2010) explicitly compare the performance of multiple

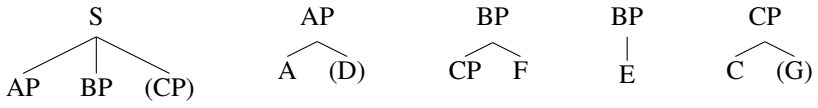

Figure 1: Rules of the context-free grammar used in Saffran 2001. Optional productions are given in parentheses.

models against participant performance in an experimental setting, but the comparison is between simple off-the-shelf models and a highly-articulated model that was customized for the experiment.

In summary, while artificial language learning studies have increased in sophistication over time, the core problem presented by the poverty of the experiment has only sporadically been addressed. Few studies explore alternative strategies participants could use to perform the task.

3 Evaluating Artificial Grammar Learning

In this section, I use the study of Saffran 2001, hereafter S2001, as a case study to explore simple computational baselines for learning tasks. I propose two baseline learning models, both of which can produce the discrimination required to pass the forced-choice grammaticality judgment task used in the study *without* learning the intended structure of the language.

3.1 An Artificial Grammar

The grammar used in the S2001 study, given in Figure 1, was adapted from an earlier artificial language learning study (Morgan and Newport 1981), in which the authors reported several correlated cues were required to learn the grammar successfully. This grammar is capable of generating 18 unique "sentences," strings of symbols that represent syntactic categories.

In the S2001 study, this language is represented using a context-free grammar, but the formalism used to represent small languages of this type has varied arbitrarily as different authors have performed artificial grammar learning experiments. The earliest work in this domain (e.g., Reber 1967) uses a finite state representation, and some researchers continue to do so (e.g., Gomez and Gerken 1999), others (e.g., Morgan and Newport 1981) recognize that many context-free grammars can also be expressed as finite state grammars and represent the grammar using both formalisms. In the simplest cases, hierarchical and linear structure may not be distinguishable. As demonstrated by Takahashi (2009), the grammar of the S2001 study is also expressible as a finite state grammar; to rule out a finite state representation and ensure that learners must discover some type of phrase structure, more sophisticated artificial language learning experiments designed to replicate syntactic phenomena such as movement or recursion are required. The simple languages used in S2001 and all studies listed above are all finite automaton recognizable, Type 3 in the Chomsky Hierarchy.

Thus for learners to learn these languages, learning a finite state automaton would be sufficient. While the limited duration and participant attention inherent in experimental settings inhibit testing of grossly more complex grammars, Takahashi (2009) discusses these limitations and proposes more complex artificial grammars to allow for deeper examination of learning mechanisms for phrase structure rules while maintaining small language size.

In this study, I focus on a different question than work which has evaluated the formal complexity of the grammars in these experiments, instead asking: how do simple learners that do not attempt to learn the formal structure of the grammar behave? If participants are actually learning the intended language, regardless of whether they represent it as a context-free grammar, finite state system, or another formal representation of sufficient power, they will be able to reliably discriminate between grammatical and ungrammatical items in the language. A question thus far unexplored is what behavior simple but less formal learning techniques may yield.

3.2 Baseline Learning Mechanisms for Artificial Grammar Studies

The typical standard of evidence in evaluating whether participants have successfully learned a language is comparing of participant performance against some set of baseline levels. The primary reason for selecting the S2001 study as a case study in this paper is the significant effort undertaken by Saffran in that study to demonstrate that participants performed better than chance and a control group, and that their performance was best explained by grammaticality as opposed to other surface characteristics of the strings. While I question its conclusions below, the S2001 study represents one of the most thorough analyses of participant performance, crucially reporting test-by-test performance and exploring possible confounds through ANCOVA modeling.

In the S2001 study, Saffran examines whether surface characteristics of the strings used in testing predict participant performance, testing the legality of the first word, chunk and anchor strength, similarity to grammatical items, presence of unique pairs in test strings, and string length. It is important to assess the predictive power of these attributes to verify that discrimination between grammatical and ungrammatical items cannot come from low-level perceptual sensitivities. The reasoning suggested by the tests performed in the S2001 study is that if no low-level mechanism can account for participants' successful discrimination, we can infer that they learned the artificial language. As predictive dependencies are a strong cue to passing the administered language test, it is inferred that subject used predictive dependencies to perform the discrimination task.

To further investigate participants' performance, I evaluate several "mid-level" mechanisms for learning an artificial language of the type used in the S2001 study. By "mid-level," I refer to learning mechanisms that do not refer to gross properties of the surface strings participants were exposed to (e.g., length, chunk strength) but lack the formal (high-level) power needed to completely learn the language. These mid-level models are able to successfully discriminate between the grammatical and ungrammatical sequences of syntactic categories without "learning" the language.

The models are trained on sequences of syntactic categories, as opposed to the words participants in the study were exposed to. The ANCOVA results given in Table 6 of the S2001 study demonstrate that participants are can successfully at look beyond the actual CVC words used in exposure to the language (e.g., *biff, klor, cav*) and infer that the relevant structure must be at the levels of syntactic categories. The careful design of the training and test items in that study prevents surface generalizations between words seen during training from being useful during testing. As a result, in this paper the learners are trained on strings of the symbols of the grammar, *A*, *C*, *D*, *E*, *F*, and *G*. For the purpose of modeling, we assume that participants are capable of abstracting from words to syntactic categories and model how they might identify relationships between syntactic categories.

The two models I propose represent simple hypotheses about how a learner may attempt to learn relationships between syntactic categories without using a formally sufficient method such as a context-free grammar or finite state automaton. The two models learn different kinds of relationships between syntactic categories: transitional probabilities (bigram model), as identified as useful in word segmentation studies, and predictive dependency relationships, as suggested by Saffran.

As I explain the models, I will use the following simple context-free grammar to demonstrate what each model will learn:

This grammar is capable of producing two strings: *A B* and *A B C*.

In both of these baseline models, the learner makes the assumption that combinations of symbols not observed during training can be treated as impossible. This assumption relies on *indirect negative evidence* in that the failure to observe something is taken as evidence that it cannot be generated by the grammar. We set aside the question of whether computational or human learners assume unobserved events are impossible or merely improbable, as both lead to the required *discrimination* discussed in this study.

A C F	A D C F	A D E C G
A C F C	A D C F C	A E
A C F C G	A D C G F	A E C
A C G F	A D E	A E C G
A C G F C	A D E C	

Table 1: Syntactic category sequences given as input to the learner.

I explain the learning mechanism used in each model below and compare what the two models learn on a simple grammar.

3.2.1 Bigram Model

A bigram model estimates the probability that one symbol follows another in the language, including special symbols for the beginning and ends of strings, BEGIN and END. Probabilities are estimated by maximum likelihood; for example, the probability of $A \rightarrow B$ is computed by dividing the number of times B follows A by the number of times A occurs. In this paper I follow the tradition of artificial language learning experiments of referring to the conditional probability of B following A ($p(B|A)$) in sequence as "transitional probability" and represent it as $p(A \rightarrow B)$. While this notation is non-standard for representing statistical models, it is easier to understand in the context of previous artificial language studies. As each transition is only conditioned on the previous symbol, this is a first-order Markov model, also called a bigram model.

The probability of a string is the product of the probability of all transitions involved. Thus the probability of the string $A\ B\ C$ would be computed as:

$$p('ABC') = p(\text{BEGIN} \rightarrow A)p(A \rightarrow B)p(B \rightarrow C)p(C \rightarrow \text{END})$$

When trained on the strings $A\ B$ and $A\ B\ C$, the transitions BEGIN $\rightarrow A$, $A \rightarrow B$, and $C \rightarrow$ END have a probability of 1.0. The transitions $B \rightarrow C$ and $B \rightarrow$ END have a probability of 0.5 assuming both the strings A B and A B C appear with equal frequency.

The assumption of equal frequency raises the question of whether probabilities are computed over *types* in the grammar definition as opposed to *tokens* resulting from the frequency of applying rules in the grammar to form strings. The exact probability is, however, irrelevant in this study; all that will be required for simulation of the S2001 study is determining whether a sequence's probability is non-zero. In the case of the example grammar, as long as both strings in the grammar are observed, the probabilities of the transitions $B \rightarrow C$ and $B \rightarrow$ END are non-zero.

3.2.2 Predictive Dependency Model

The predictive dependency model, based on the types of cues Saffran suggests are used based in the S2001 study, forms two generalizations based on the co-occurrence of syntactic categories in a string. It learns that A *requires* b if $p(B|A) = 1$, that is every string containing A also contains B. It learns that A *excludes* B if $p(B|A) = 0$, that is every string containing A does not contain B. Co-occurrence within the same syntactic category is defined such that if category A never occurs more than once in a string, A excludes A.

When trained on the strings $A\ B$ and $A\ B\ C$, the learner detects that A requires B, B requires A, and C requires A and B. It also learns that A, B, and C exclude themselves; that is, each cannot appear more than once.

4 Experiment

To simulate the S2001 study, computational models were trained using the 14 unique syntactic category strings generated by the grammar that the participants in the study of Saffran (2001) were

	Learned Representation	
Category	Requires	Excludes
A		A
C	A	
D	A	D
E	A	E, F
F	A, C	E, F
G	A, C	

Table 2: Predictive dependency rules learned from the artificial language data.

Transition	Probability	Interpretation
BEGIN $\rightarrow A$	1.0	All sentences begin with A.
$A \rightarrow D$	0.428	D may follow A.
$A \rightarrow C$	0.357	C may follow A.
$A \rightarrow E$	0.214	E may follow A.
$C \rightarrow F$	0.313	F may follow C.
$C \rightarrow G$	0.375	G may follow C.
$C \rightarrow$ END	0.313	C may end a sentence.
$D \rightarrow C$	0.5	C may follow D.
$D \rightarrow E$	0.5	E may follow D.
$E \rightarrow C$	0.667	C may follow E.
$E \rightarrow$ END	0.333	E may end a sentence.
$F \rightarrow C$	0.5	C may follow F.
$F \rightarrow$ END	0.5	F may end a sentence.
$G \rightarrow F$	0.5	F may follow G.
$G \rightarrow$ END	0.5	G may end a sentence.

Table 3: Non-zero transitional probabilities learned from the artificial language data. As discussed in Section 3.2.1, probabilities are computed by assuming all strings produced by the grammar are equally frequent.

exposed to examples of (Table 1).[1] The patterns learned from exposure to the artificial language are given in Tables 2 and 3.

Participants in the S2001 study were tested using a two-way forced choice between a grammatical and ungrammatical item of the language.[2] Participants that had learned the structure of the language would be expected to perform above chance and better than that control group on all tests. These tests, show in Table 4, attempt to assess the degree to which participants preferred the grammatical items to ungrammatical ones.

[1] While the grammar generates 18 unique strings, the S2001 study excluded strings with more than five symbols, leaving 14 strings.

[2] An additional test given in that study, not simulated here for reasons of length, examines discrimination between grammatical and ungrammatical chunks sentence chunks.

Test	Grammatical Item	Ungrammatical Item	Pred. Dependency Model Response
Test 1: Every sentence must contain an A word.	A C F	*C F	**Pass:** C and F each require A
Test 2: No sentence may contain more than one A word.	A D E C	*A A D E C	**Pass:** A excludes another A
Test 3: A BP expands to contain an E or a C but not both at once.	A D E	*A D C E	**Fail:** Learner labels *A D C E* as grammatical.
Test 4: If there is a D word, then there must be an A word.	A D C F C	*D C F C	**Pass:** D requires A
Test 5: If there is an F word, then there must be a C word.	A C F	*A F	**Pass:** F requires C
Test 6: If there is a G word, then there must be a C word.	A E C G	*A E G	**Pass:** G requires C

Table 4: Forced-choice grammar tests administered in Saffran 2001. The performance of the predictive dependency learner is marked as "pass" if the simulation correctly rejected ungrammatical items.

The bigram model responds perfectly to the tests. In each test it assigns the ungrammatical item zero probability and the grammatical item a non-zero probability, demonstrating perfect discrimination between the grammatical and ungrammatical items tested. As all grammatical sequences of syntactic categories are observed during training, when seen again in testing they will be assigned a non-zero probability. The tested ungrammatical items are rejected for the following reasons. In Test 1, **C F* is rejected because BEGIN $\rightarrow$ C has zero probability. In Test 2, **A A D E C* is rejected because $A \rightarrow A$ has zero probability. In Test 3, **A D C E* is rejected because $C \rightarrow E$ has zero probability. In Test 4, **D C F C* is rejected because BEGIN $\rightarrow$ D has zero probability. In Test 5, **A F* is rejected because $A \rightarrow F$ has zero probability. In Test 6, **A E G* is rejected because $E \rightarrow G$ has zero probability.

While the bigram model is capable of passing all tests used in this study, this does not imply that it has "learned" the grammar. As the model only has a memory of a single symbol, sequences for which all two adjacent symbols form an acceptable transition will be accepted by this model. For example, as *A D C F C* is produced by the grammar, the transitions $C \rightarrow F$, $F \rightarrow C$, and $C \rightarrow$ END have non-zero probability. Thus, any grammatical sequence ending with *C* can have *F C* appended to it and still be accepted by the bigram model; sequences of the form **A D C F C F C*, **A D C F C F C F C*, **A D C F C F C F C F C* can be used to create ungrammatical examples *ad infinitum* that will be incorrectly accepted by the bigram model.

While the bigram model correctly responds to all of the tests, the dependency learner fails Test 3 while passing all others (Table 4). As the tests were designed by Saffran with predictive dependencies in mind, it is unsurprising that this simple learner does so well in these tests. The reason for passing each test is given in Table 4. The failure of the predictive dependency learner in Test 3 merits further discussion. Test 3, which compares *A D E* and **A D C E* tests whether the learner recognizes that there are two ways to expand *BP* that are in complementary distribution; if *BP* expands to *E*, it cannot also expand to a *CP* which will then expand to a *C*. This type of mutual exclusivity cannot be detected by the predictive dependency learner unless it holds true on the entire string, i.e., if *C* globally excludes *E*. As the grammatical strings *A D E C* and *A D E C G* show, this global exclusion does not hold, and thus the predictive dependency learner cannot reject the ungrammatical item in Test 3.

Comparison to the human participant data in the S2001 study suggests that the failure of this test may be a characteristic pattern of human performance as well. Adults and children in the exper-

imental group did not perform significantly better than the control group in either administration of the forced choice grammaticality judgment task.

5 Discussion

Evaluating human subject performance on the discrimination task simulated in this paper, Saffran reports:

> The results suggest that learners can detect phrasal units in the absence of relevant cues other than predictive dependencies. (Saffran 2001:503)

The results presented above demonstrate that this analysis is partially correct. The fit between the predictive dependency learner and the human participant pattern of performing well with the exception of Test 3 suggests that it is likely that participants used predictive dependencies to discriminate grammatical and ungrammatical items. The fact that participants succeeded in this task with the very limited of cue of predictive dependencies *pace* Morgan and Newport 1981 suggests that predictive dependencies are the crucial cue to success in this task.

However, the conclusion that learners detected phrasal units is questionable on several grounds. First, it assumes that the language used in S2001 has phrasal units at all. While Saffran describes the language using a context-free grammar, suggesting it has phrasal units, as previously discussed it is a finite-state language. A formally correct analysis of it need not contain any phrasal units (i.e., *AP*, *BP*, *CP*) at all. Second, the tests administered are not enough to determine that participants learned any phrase structure; the bigram learner which learns no form of phrase structure can pass all tests perfectly. Finally, participants consistently fail Test 3, the test perhaps that is the closest to an acceptable diagnostic of a phrase-structure-like analysis as it could test whether learners identify that *C F*, *C G F*, and *E* each form a constituent of the same type.

Returning to the poverty of the experiment, recall that this inference problem may manifest itself in at least two fashions in artificial language learning experiments:

1. Participants may demonstrate discrimination without performing the intended learning task (cf. Endress and Mehler 2009).

2. Participants may use learning cues different than those intended by the experimenters (cf. Perruchet and Vinter 1998).

The modeling performed in this paper suggests that the best explanation of participant performance in the S2001 study is that they did not perform the intended learning task. While they do appear to attend to the intended learning cue, predictive dependencies, they use it to solve this task directly without making any attempt to infer phrase structure. This is analogous to the pattern observed by Endress and Mehler (2009) where participants appear to extract transitional probabilities without actually performing word segmentation.

While this modeling study has provided a useful starting point for exploring the issue of the poverty of the experiment, further work is required to thoroughly analyze participants' behavior in the S2001 study. Future work should seek to model participant performance on individual trials, allowing a more quantitative approach to as opposed to the more qualitative comparison between simulation and aggregate participant performance reported here.

A larger set of case studies across domains is beyond the scope of this paper. However, there are a number of domains in which further investigation is necessary, most notably word segmentation. The data collected by Frank et al. (2010) regarding subject performance during a word segmentation-like task can provide an excellent testing ground for both complex and simple models beyond those explored in their study.

This paper demonstrates that rather than providing a postmortem for existing experiments, modeling simple baseline techniques that participants may use should be adopted into the experimental process:

1. When an experiment is designed, a set of baseline approaches that may discriminate grammatical items from ungrammatical ones must be developed in parallel. This allows the performance of these models to be evaluated as a part of hypothesis testing as opposed to post hoc analysis.

2. The stimuli for the experiment and the tests that will be administered to participants should be tested against baseline models to verify that the baselines cannot succeed in the testing phase. If, as in the case of the S2001 study, baseline techniques are capable of performing well at the intended task, it may be impossible to determine that subjects are using a method different than the (typically undesirable) baseline approach.

3. Participant performance must not only be compared against chance and control groups but against simple baseline computational models. To support the conclusion that participants learned the intended representation, participant performance must be best explained by learning the intended representation and not by "shortcuts" used by baseline learners.

This study highlights the reality that careful stimulus design and observing participants' discrimination between grammatical and ungrammatical items are not enough to suggest how learners accomplish a learning task. Additional steps must be taken to be sure that the strategy expected by the experimenters is the one that participants used. If we simply assume that participants' success in learning is attributable to the intended cue(s), this is a form of the *post hoc ergo propter hoc* fallacy; the design of the stimulus is assumed to have a causal relationship with the learning strategy without verification that such a relationship must necessarily exist. Repeated experiment after experiment, such inferential leaps risk creating a canon out of repeated confirmation bias.

6 Conclusion

The modeling and analysis of the Saffran 2001 study presented in this paper demonstrates that the conclusions of that study regarding what participants learned reach further than what can be supported by the data; put more simply, it has been shown that *you can't get there from here*. We find that the proposed dependency learning model provides a better fit to human performance than assuming participants learned a context-free, finite state, or bigram representation of the structure governing syntactic categories of the artificial language presented.

This finding highlights the caution that must be used in the inference problem I have named *the poverty of the experiment*. Strong claims about the mechanisms of language learning must be accompanied by equally strong verification of those mechanisms and the experiments that suggest them. Using of simple computational models as a part of experimental design and analysis requires little effort and should become a necessary part of this verification process.

References

Aslin, R.N., J.R. Saffran, and E.L. Newport. 1998. Computation of conditional probability statistics by 8-month-old infants. *Psychological Science* 9:321–324.

Endress, A.D., and J. Mehler. 2009. The surprising power of statistical learning: When fragment knowledge leads to false memories of unheard words. *Journal of Memory and Language* 60:351–367.

Frank, M.C., S. Goldwater, T.L. Griffiths, and J.B. Tenenbaum. 2010. Modeling human performance in statistical word segmentation. *Cognition* 117:107–125.

Gentner, T.Q., K.M. Fenn, D. Margoliash, and H.C. Nusbaum. 2006. Recursive syntactic pattern learning by songbirds. *Nature* 440:1204–1207.

Gibson, J.J., and E.J. Gibson. 1955. Perceptual learning: Differentiation or enrichment? *Psychological Review; Psychological Review* 62:32.

Gomez, R.L., and L.A. Gerken. 1999. Artificial grammar learning by 1-year-olds leads to specific and abstract knowledge. *Cognition* 70:109–135.

Johnson, E.K., and P.W. Jusczyk. 2001. Word segmentation by 8-month-olds: When speech cues count more than statistics. *Journal of Memory and Language* 44:548–567.

Lew-Williams, C., B. Pelucchi, and J.R. Saffran. 2011. Isolated words enhance statistical language learning in infancy. *Developmental Science* 14:1323–1329.

Lew-Williams, C., and J.R. Saffran. 2012. All words are not created equal: Expectations about word length guide infant statistical learning. *Cognition* 122:241–246.

Mattys, S.L., P.W. Jusczyk, P.A. Luce, and J.L. Morgan. 1999. Phonotactic and prosodic effects on word segmentation in infants. *Cognitive Psychology* 38:465–494.

Morgan, J.L., and E.L. Newport. 1981. The role of constituent structure in the induction of an artificial language. *Journal of verbal learning and verbal behavior* 20:67–85.

Pelucchi, B., J.F. Hay, and J.R. Saffran. 2009. Statistical learning in a natural language by 8-month-old infants. *Child development* 80:674–685.

Perruchet, P., and A. Vinter. 1998. Parser: A model for word segmentation. *Journal of Memory and Language* 39:246–263.

Reber, A.S. 1967. Implicit learning of artificial grammars. *Journal of verbal learning and verbal behavior* 6:855–863.

Reber, A.S. 1969. Transfer of syntactic structure in synthetic languages. *Journal of Experimental Psychology* 81:115–119.

Saffran, J., M. Hauser, R. Seibel, J. Kapfhamer, F. Tsao, and F. Cushman. 2008. Grammatical pattern learning by human infants and cotton-top tamarin monkeys. *Cognition* 107:479–500.

Saffran, J.R. 2001. The use of predictive dependencies in language learning. *Journal of Memory and Language* 44:493–515.

Saffran, J.R. 2002. Constraints on statistical language learning. *Journal of Memory and Language* 47:172–196.

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

Saffran, J.R., E.L. Newport, and R.N. Aslin. 1996b. Word segmentation: The role of distributional cues. *Journal of Memory and Language* 35:606–621.

Shukla, M., K.S. White, and R.N. Aslin. 2011. Prosody guides the rapid mapping of auditory word forms onto visual objects in 6-mo-old infants. *Proceedings of the National Academy of Sciences* 108:6038–6043.

Takahashi, E. 2009. Beyond statistical learning in the acquisition of phrase structure. Doctoral dissertation, University of Maryland College Park.

Thiessen, E.D., and J.R. Saffran. 2003. When cues collide: use of stress and statistical cues to word boundaries by 7-to 9-month-old infants. *Developmental psychology* 39:706–716.

Van Heijningen, C.A.A., J. De Visser, W. Zuidema, and C. Ten Cate. 2009. Simple rules can explain discrimination of putative recursive syntactic structures by a songbird species. *Proceedings of the National Academy of Sciences* 106:20538–20543.

Department of Computer and Information Science
Institute for Research in Cognitive Science
University of Pennsylvania
Philadelphia, PA 19104
lignos@cis.upenn.edu