Available online at www.sciencedirect.com

SCIENCE DIRECT®

COGNITION

Cognition 93 (2004) 147–155

www.elsevier.com/locate/COGNIT

Discussion

Learning the unlearnable: the role of missing evidence

Terry Regier[a,*], Susanne Gahl[b]

[a]Department of Psychology, University of Chicago, 5848 South University Avenue, Chicago, IL 60637, USA
[b]University of Illinois at Urbana-Champaign, Urbana, IL, USA

Received 24 October 2003; accepted 16 December 2003

Abstract

Syntactic knowledge is widely held to be partially innate, rather than learned. In a classic example, it is sometimes argued that children know the proper use of anaphoric *one*, although that knowledge could not have been learned from experience. Lidz et al. [Lidz, J., Waxman, S., & Freedman, J. (2003). What infants know about syntax but couldn't have learned: Experimental evidence for syntactic structure at 18 months. *Cognition, 89*, B65–B73.] pursue this argument, and present corpus and experimental evidence that appears to support it; they conclude that specific aspects of this knowledge must be innate. We demonstrate, *contra* Lidz et al., that this knowledge may in fact be acquired from the input, through a simple Bayesian learning procedure. The learning procedure succeeds because it is sensitive to the *absence* of particular input patterns—an aspect of learning that is apparently overlooked by Lidz et al. More generally, we suggest that a prominent form of the "argument from poverty of the stimulus" suffers from the same oversight, and is as a result logically unsound.

Keywords: Language acquisition; Syntax; Innateness; Poverty of the stimulus; Emergence; Bayesian learning; Indirect learning

One of the core questions of cognitive science is whether human language relies on innate syntactic knowledge. On one influential view, at least some aspects of syntax must be innate, since the child possesses syntactic knowledge that could not have been learned from his or her impoverished linguistic input (Chomsky, 1981; Pinker, 1989). While this "argument from poverty of the stimulus" has generally met with wide acceptance, it has also recently been challenged. A growing number of researchers have suggested that

[*] Corresponding author.
E-mail address: regier@uchicago.edu (T. Regier).

0022-2860/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.cognition.2003.12.003

the child's linguistic input may suffice to allow general-purpose learning mechanisms to acquire syntactic regularities, without the benefit of specifically syntactic innate knowledge (Christiansen & Chater, 1999; Elman, 1993; Pullum & Scholz, 2002; Rohde & Plaut, 1999; Seidenberg, 1997; Tomasello, 2000).

Lidz, Waxman, and Freedman (2003) respond to these challenges, with an empirical investigation of young children's syntactic knowledge and linguistic input. They conclude that specific aspects of children's knowledge are not learnable from the input—and therefore must be innate.

We suggest that Lidz et al.'s innatist conclusion does not follow from their data. We support this claim by demonstrating that a simple Bayesian learning model can account for their findings, without the innate knowledge they propose. We suggest that the flaw in the argument is not theirs, however; rather, it was inherited from the "poverty of the stimulus" tradition. Like some (but not all) earlier work in this tradition, Lidz et al. overlook the fact that much may be learned by noting which patterns are *absent* from the input. Ultimately, we suggest that the argument from poverty of the stimulus is unsound if the role of missing evidence is ignored.

1. The syntax of anaphoric *one*

Lidz et al. (2003) approach the broad question of innate syntactic knowledge by examining a specific phenomenon: the anaphoric use of *one*, as in sentence (1).

(1) I'll play with this red ball and you can play with that one.

Here, the word *one* refers anaphorically to *red ball*. Following Hornstein and Lightfoot (1981), the authors suggest that such sentences implicitly pose a learning problem concerning the structure of the antecedent noun phrase—here, *this red ball*. Such a noun phrase could in principle be analyzed in at least two ways, as illustrated in Fig. 1.

Lidz et al. argue that the nested structure must be the correct one: it is commonly assumed that an anaphoric element may substitute only for a constituent; here, *one* refers

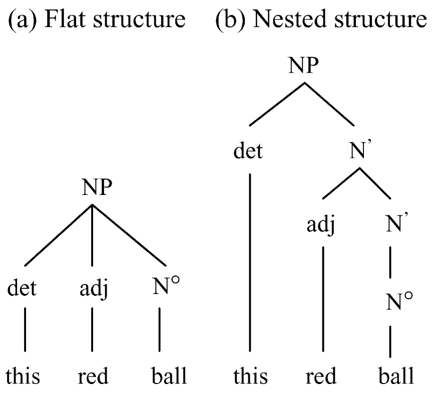
(a) Flat structure (b) Nested structure

Fig. 1. Two possible structures for noun phrases of the form determiner–adjective–noun, such as "this red ball." (Adapted from Lidz et al., 2003)

anaphorically to *red ball*, which appears as a constituent (i.e. a node containing that string and nothing else) only in the nested structure, as the upper N′. Given this assessment of adult grammar, Lidz et al. then ask how children could arrive at this knowledge. Their core learning problem is reproduced here, and marked (†) for future reference:

> (†) Suppose that a learner is exposed to small discourses like [(1)] in which *one* is anaphoric to some previously mentioned discourse entity and that the learner has recognized that *one* is anaphoric. In order to understand this use of *one*, the learner must know that it is anaphoric to the phrasal category N′, which is possible only under the nested structure hypothesis. However, the data to support this hypothesis are not available to the learner for the following reason. Every situation that makes *one* = [$_{N'}$ *red ball*] true also makes *one* = [$_{N^\circ}$ *ball*] true [since any actual red ball, to which *one* might refer, is also a ball—TR & SG]. Thus, if the learner had come to the flat structure hypothesis or to the hypothesis that *one* is anaphoric to N° and not N′, evidence that this is wrong would be extremely difficult to come by. (p. B67)[1]

Lidz et al. (2003) proceed to empirically demonstrate two points. The first is that children's input almost completely lacks a particular form of evidence that, if present, would lead children to the correct hypothesis (i.e. [$_{N'}$ *red ball*]). Children must instead learn on the basis of evidence such as (1), which is consistent with both correct and incorrect hypotheses. The second point is that children near the beginning of language learning nonetheless behave as if they know the correct hypothesis. Thus, the authors argue, children know the correct use of anaphoric *one* without ever having encountered evidence that would allow them to acquire that knowledge. They conclude that aspects of that knowledge are not acquired, but rather innate: children never entertain the incorrect hypothesis that *one* = [$_{N^\circ}$ *ball*] (p. B67).[2]

The general logic is that of the argument from poverty of the stimulus (APS). However, the APS is not a single argument, but rather a family of related arguments (Pullum & Scholz, 2002). The specific version of the APS deployed by Lidz et al. makes a critical assumption: that evidence consistent with multiple hypotheses cannot discriminate among those hypotheses; for instance, that sentences such as (1) cannot discriminate between [$_{N'}$ *red ball*] and [$_{N^\circ}$ *ball*]. This assumption may also be found in some earlier presentations of the APS (e.g. Baker, 1978:416; Hornstein & Lightfoot, 1981:18–20; Pinker, 1989:6).

This assumption is incorrect. Given evidence that is consistent with several hypotheses, a learner can come to discriminate among them, for principled, domain-general reasons. In particular, if one of the hypotheses predicts not only the input that is seen, but also input of

[1] On this argument, the hypothesis *one* = [$_{N'}$ *ball*] (the lower N′ in the nested structure) should also be difficult to disconfirm, given only data such as (1). Our unverified intuition is that *one* = [$_{N'}$ *ball*] is false when (1) is spoken with neutral prosody, but true when it is spoken with emphasis on the word *red*, implying a contrast with the other (non-red) ball. Anaphoric *one* also refers to the lower N′ in a variety of other sentences. In this article, we restrict attention to sentences in which the correct hypothesis is the upper N′ rather than the lower N′—as do Lidz et al. A fuller treatment of the acquisition of anaphoric *one* would need to also cover those cases in which *one* refers to the lower N′.

[2] We use the notations [$_{N^\circ}$ *ball*] and [$_{N'}$ *red ball*] without loss of generality, to also refer to analogous structures in sentences with different adjectives and nouns.

another sort that is never seen, that *absence* can serve as evidence against the hypothesis. Some presentations of the APS explicitly consider this possibility (e.g. Chomsky, 1981:9).

How does this idea apply to anaphoric *one*? The hypothesis *one* = [$_{\mathrm{N'}}$ *red ball*] predicts that the referent of *one* will be red—this prediction is always confirmed. In contrast, the hypothesis *one* = [$_{\mathrm{N^\circ}}$ *ball*] predicts that the referent of *one* will be a ball of any color—and thus we should expect that at least some of its uses will refer to balls that are not red. This will never happen since the hypothesis is incorrect. Thus, our expectations for this hypothesis are not fully met: we do not see the full range of expected referents. This absence can serve as implicit negative evidence against *one* = [$_{\mathrm{N^\circ}}$ *ball*].[3]

The same general intuition underlies Laplace's (1825) law of succession, which may be used to estimate the probability that the sun will rise tomorrow, given evidence that it has risen every morning for some number of days. These successive sunrises are consistent with the correct hypothesis, namely that the sun rises without fail every day and therefore will tomorrow as well—but they are also consistent with false hypotheses such as that the sun has only a 50% chance of rising on any given day: each time we saw it rise, it may have just happened to rise. But on this latter hypothesis, we would also expect the sun to *not* rise on at least some of the days we have observed. This expected evidence never appears, and its absence undercuts the false "50%" hypothesis, despite the fact that the observed data are compatible with that hypothesis.

This mathematical parallel suggests that formal models may be helpful in addressing the role of absent evidence in the argument from poverty of the stimulus. It also suggests that there may be nothing particularly linguistic about the learning processes involved—the learning may fall out of domain-general considerations. We now turn to explore this possibility.

2. Models of indirect learning

Recently, a number of computational learning models have pursued probabilistic approaches to language learning (e.g. Brent & Cartwright, 1996; Eisner, 2002; Niyogi & Berwick, 1996). In addition, several learning models have demonstrated that indirect evidence may shape learning (e.g. Landauer & Dumais, 1997; Merriman, 1999; Regier, 1996; see Regier, 2003 for a review). The Bayesian learning model of Tenenbaum and Griffiths (2001) is particularly relevant for present purposes. This model and variants thereof have accounted for Shepard's (1987) roughly exponential generalization gradient (Tenenbaum & Griffiths, 2001), and several aspects of word-learning: learning from a small number of examples (Tenenbaum & Xu, 2000), the interaction of syntactic and semantic knowledge (Niyogi, 2002), and lexical contrast (Regier, 2003). Critically, this model formalizes the idea of learning from the absence of expected data.

Tenenbaum and Griffith's model assumes that learning is rational, governed by the normative standard of Bayes' rule:

$$(2)\quad p(H|e) \propto p(e|H)p(H)$$

[3] This idea is distinct from the "subset principle" (Berwick, 1986; Pinker, 1995:172–175), which holds that children do not consider broad hypotheses until the input requires them to do so. We suggest, in contrast, that children initially consider all hypotheses, and *learn* to discard overly broad ones.

Here H is a hypothesis in a hypothesis space, and e is the observed evidence. Bayes' rule determines the probability of each hypothesis in the hypothesis space given the observed evidence, as a function of the likelihood $p(e|H)$ (that is, the probability of observing that evidence, given that the hypothesis is true), and the prior probability $p(H)$ (that is, the a priori probability of that hypothesis being true).

At the heart of Tenenbaum and Griffiths' model is their "size principle". This principle holds that the likelihood of seeing a particular sort of evidence e, given that hypothesis H is true, can be determined by considering the full range of different sorts of evidence that H could give rise to, and assuming that e was randomly selected from this set. Formally, if H can give rise to $|H|$ different sorts of evidence, and *"$e \in H$"* means that e is one of those sorts, then:

$$(3)\ \ p(e|H) = \begin{cases} \dfrac{1}{|H|} & \text{if } e \in H \\ 0 & \text{otherwise} \end{cases}$$

This principle causes the likelihood $p(e|H)$ to be largest for those hypotheses that support the smallest range of possible evidence. For example, if we observe a dog barking, this principle would support the hypothesis "only dogs bark" more strongly than it would support the also-consistent hypothesis "animals of all sorts bark." This follows since the hypothesis "animals of all sorts bark" can give rise to a broader range of possible evidence, making it less likely that a random selection from that range would yield a barking dog. In the absence of additional evidence such as a barking cat, this hypothesis will lose support. It is in this sense that the model formalizes the notion of learning from the absence of expected evidence.

Tenenbaum and Griffiths obtain the likelihood $p(e^n|H)$ of observing n occurrences of evidence e by assuming the observations are independent, and multiplying the likelihood for a single observation n times:

$$(4)\ \ p(e^n|H) = \begin{cases} \dfrac{1}{|H|^n} & \text{if } e \in H \\ 0 & \text{otherwise} \end{cases}$$

We may now reconsider the argument from poverty of the stimulus as applied to anaphoric *one* (†), through the lens of this model. We consider four hypotheses, each of which represents a possible node to which anaphoric *one* might refer. The initial word of each hypothesis' name ("nested" or "flat") indicates the structure from Fig. 1 in which the node resides. The hypotheses are:

i. nested:$[_{\mathrm{N'}}$ *red ball*$]$
ii. nested:$[_{\mathrm{N'}}$ *ball*$]$
iii. nested:$[_{\mathrm{N^\circ}}$ *ball*$]$
iv. flat:$[_{\mathrm{N^\circ}}$ *ball*$]$

The first hypothesis is the correct one for sentences such as (1). Critically, however, the hypothesis space also contains two hypotheses (nested:[$_{\mathrm{N}^\circ}$ *ball*] and flat:[$_{\mathrm{N}^\circ}$ *ball*]) that Lidz et al. claim must be innately excluded from consideration for successful learning. On their view, it should be impossible to learn the correct hypothesis, given this hypothesis space and realistic input.

Let the evidence e^n consist of n observations of *one* referring to a red ball while (1) is uttered—this is the same as the scenario envisaged in (†). Assume that the world contains balls of c different colors, including red. Under these circumstances, the likelihood $p(e^n|H)$ is shown below for each of the above four hypotheses.

$$p(e^n|\text{nested} : [_{\mathrm{N}'}\, \mathit{ball}]) = \frac{1}{1^n} = 1$$

$$p(e^n|\text{nested} : [_{\mathrm{N}'}\, \mathit{ball}]) = \frac{1}{c^n}$$

$$p(e^n|\text{nested} : [_{\mathrm{N}^\circ}\, \mathit{ball}]) = \frac{1}{c^n}$$

$$p(e^n|\text{flat} : [_{\mathrm{N}^\circ}\, \mathit{ball}]) = \frac{1}{c^n}$$

These values are derived from (4), which is based on the size principle. For each hypothesis, $|H|$ is the number of different colors that balls may appear in, when referred to by *one*. While the evidence is consistent with all four hypotheses, it supports the correct one more strongly than the others.

Fig. 2 shows the results of applying this model to the learning problem described in (†), under the assumption that $c = 2$ (e.g. there are only red balls and blue balls). Higher values of c yield qualitatively similar results, although with faster learning. A uniform prior was assumed, such that the four hypotheses were equally probable before observing evidence: $p(H) = \frac{1}{4}$ for each. We combined the prior $p(H)$ and likelihood $p(e^n|H)$ to obtain the posterior probability $p(H|e^n)$ for each hypothesis, for $n = 0$ through $n = 5$ exposures to the evidence. The [$_{\mathrm{N}'}$ *red ball*] hypothesis is found to be quite probable, and the three [*ball*] hypotheses quite improbable, demonstrating that this allegedly unlearnable knowledge is in fact learnable, from evidence of the sort described in (†).

Does the child receive enough data to support such learning? Lidz et al. found 31 adult utterances of the form shown in (1) (i.e. anaphoric *one* with an antecedent noun phrase of the form determiner–adjective–noun) in the pooled Adam (Brown, 1973) and Nina (Suppes, 1974) corpora in the Childes database (J. Lidz, personal communication). The Adam corpus spans 116 h of caretaker–child interaction, while the Nina corpus spans 48 h, for a total of 164 h. If we take these numbers to be representative, we may assume that the child receives a sentence of the form shown in (1) every 164/31 = 5.3 h. The simulation shown here required only five exposures for nearly complete learning; these exposures could be supplied in $26\frac{1}{2}$ h of interaction—that is, a few days. Thus, even very young children may receive enough input to allow them to learn the knowledge in question.

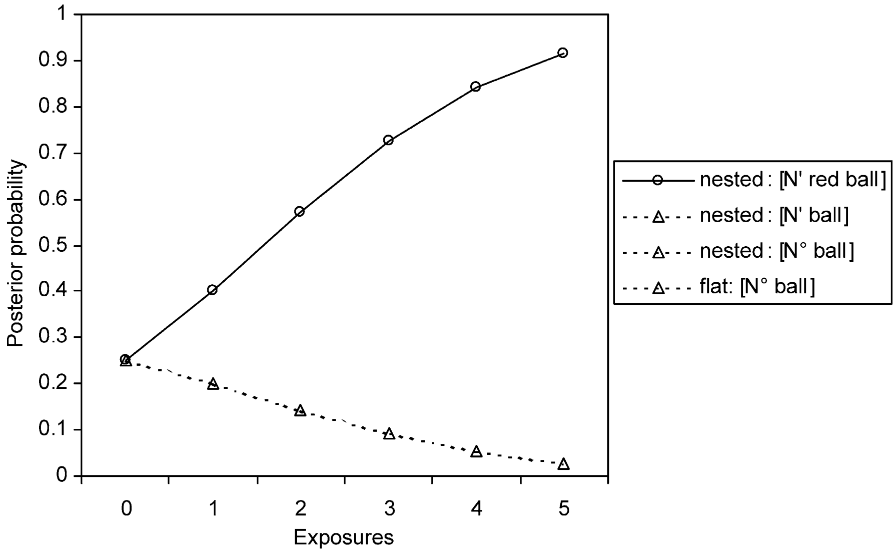

Fig. 2. Learning the unlearnable. The knowledge that *one* = $[_{N'}$ *red ball*] can be learned in only a few exposures. The probabilities of the three [*ball*] hypotheses are identical throughout, and thus appear as a single dotted line.

3. Discussion

We have shown that a simple Bayesian learning model can learn syntactic knowledge that was claimed to be unlearnable. We anticipate two potential objections to such a demonstration.

The first potential objection is that the model is unrealistic. It assumes perfect memory, a capacity that both children and adults lack. The model also fails to capture the full range of uses of anaphoric *one*: it falsely assumes that *one* will always refer to the same node in the antecedent noun phrase. In actuality, *one* will refer to either the upper or the lower N' in the nested structure, depending on the sentence. We followed Lidz et al. in focusing on the restricted case in which *one* refers to the upper N'—and we suggest that the broader lessons of this demonstration are not undermined by that restriction, nor by the idealization of perfect memory. The broader point that the model illustrates, despite its shortcomings, is that evidence that is consistent with multiple learning hypotheses can sometimes discriminate among those hypotheses in a principled manner. Since some prominent versions of the argument from poverty of the stimulus assume that such learning is not possible, the model highlights a flaw in these arguments.

The second potential objection is that the model's hypothesis space is very small. How, one may ask, can such a constrained model challenge the proposition that syntax learning must be constrained? We would respond that the model does not challenge that rather general proposition. Instead, it challenges the more specific proposal that particular sorts of syntactic knowledge must be innate (e.g. that *one* $\neq [_{N^\circ}$ *ball*]). The model is also constrained in preferring narrow hypotheses over broad ones—but this is not a language-specific constraint, nor an arbitrary one. Rather, as we have seen, it emerges from the general process of determining how likely a particular observation is, given a hypothesis.

More generally, learning of any sort is impossible without constraints of some kind—stemming from the structure of the hypothesis space, the nature of the learning process, or both. The central question then should not be whether syntax learning is constrained; it must be, like any other form of learning. Instead, in our view, the central question should be whether the constraints that govern syntax learning are themselves syntactic in nature—part of a specifically syntactic predisposition for language—or whether they emerge from more general aspects of cognition. A fuller answer to this question will require continued investigation, to determine how much of syntax learning can be explained in domain-general terms.

Acknowledgements

This work was supported by NIH grant DC03384. We thank Larry Hedges, Jeff Lidz, and Josh Tenenbaum for helpful discussions.

References

Baker, C. (1978). *Introduction to generative-transformational syntax*. Englewood Cliffs, NJ: Prentice-Hall.

Berwick, R. C. (1986). Learning from positive-only examples: The subset principle and three case studies. In R. S. Michalski, J. C. Carbonell, & T. M. Mitchell (Eds.), *Machine learning: An artificial intelligence approach* (*Vol. 2*). Los Altos, CA: Morgan Kaufmann.

Brent, M., & Cartwright, T. (1996). Distributional regularities are useful for segmentation. *Cognition*, *61*, 93–105.

Brown, R. (1973). *A first language*. Cambridge, MA: Harvard University Press.

Chomsky, N. (1981). *Lectures on government and binding: The Pisa lectures*. Berlin: Mouton de Gruyter.

Christiansen, M., & Chater, N. (1999). Toward a connectionist model of recursion in human linguistic performance. *Cognitive Science*, *23*, 157–205.

Eisner, J. (2002). Discovering syntactic deep structure via Bayesian statistics. *Cognitive Science*, *26*, 255–268.

Elman, J. (1993). Learning and development in neural networks: The importance of starting small. *Cognition*, *48*, 71–99.

Hornstein, N., & Lightfoot, D. (1981). Introduction. In N. Hornstein, & D. Lightfoot (Eds.), *Explanation in linguistics: The logical problem of language acquisition*. London: Longman.

Landauer, T., & Dumais, S. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*, 211–240.

Laplace, P.-S (1825). *Philosophical essay on probabilities*. Translated by A. Dale (1995) from the fifth French edition. New York: Springer.

Lidz, J., Waxman, S., & Freedman, J. (2003). What infants know about syntax but couldn't have learned: Experimental evidence for syntactic structure at 18 months. *Cognition*, *89*, B65–B73.

Merriman, W. (1999). Competition, attention, and young children's lexical processing. In B. MacWhinney (Ed.), *The emergence of language* (pp. 331–358). Mahwah, NJ: Lawrence Erlbaum.

Niyogi, S. (2002). Bayesian learning at the syntax–semantics interface. In W. Gray, & C. Schunn (Eds.), *Proceedings of the 24th Annual Conference of the Cognitive Science Society* (pp. 697–702). Mahwah, NJ: Lawrence Erlbaum.

Niyogi, P., & Berwick, R. (1996). A language learning model for finite parameter spaces. *Cognition*, *61*, 161–193.

Pinker, S. (1989). *Learnability and cognition: The acquisition of argument structure*. Cambridge, MA: MIT Press.

Pinker, S. (1995). Language acquisition. In L. Gleitman, & M. Liberman (Eds.), *Language: An invitation to cognitive science* (2nd ed., *Vol. 1*, pp. 135–182). Cambridge, MA: MIT Press.

Pullum, G., & Scholz, B. (2002). Empirical assessment of stimulus poverty arguments. *The Linguistic Review*, *19*, 9–50.

Regier, T. (1996). *The human semantic potential: Spatial language and constrained connectionism*. Cambridge, MA: MIT Press.

Regier, T. (2003). Emergent constraints on word-learning: A computational review. *Trends in Cognitive Sciences*, *7*, 263–268.

Rohde, D., & Plaut, D. (1999). Language acquisition in the absence of explicit negative evidence: How important is starting small? *Cognition*, *72*, 67–109.

Seidenberg, M. (1997). Language acquisition and use: Learning and applying probabilistic constraints. *Science*, *275*, 1599–1603.

Shepard, R. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*, 1317–1323.

Suppes, P. (1974). The semantics of children's language. *American Psychologist*, *29*, 103–114.

Tenenbaum, J., & Griffiths, T. (2001). Generalization, similarity and Bayesian inference. *Behavioral and Brain Sciences*, *24*, 629–640.

Tenenbaum, J., & Xu, F. (2000). Word learning as Bayesian inference. In L. Gleitman, & A. Joshi (Eds.), *Proceedings of the 22nd Annual Conference of the Cognitive Science Society* (pp. 517–522). Mahwah, NJ: Lawrence Erlbaum.

Tomasello, M. (2000). Do children have adult syntactic competence? *Cognition*, *74*, 209–253.