COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 35 (2011) 1207–1242

ISSN: 0364-0213 print / 1551-6709 online
DOI: 10.1111/j.1551-6709.2011.01189.x

Poverty of the Stimulus Revisited

Robert C. Berwick,[a] Paul Pietroski,[b] Beracah Yankama,[c] Noam Chomsky[d]

[a]*Department of EECS and Brain and Cognitive Sciences, MIT*
[b]*Departments of Philosophy and Linguistics, University of Maryland*
[c]*Department of EECS, MIT*
[d]*Department of Linguistics and Philosophy, MIT*

Received 10 March 2009; received in revised form 5 February 2011; accepted 7 April 2011

Abstract

A central goal of modern generative grammar has been to discover invariant properties of human languages that reflect ''the innate schematism of mind that is applied to the data of experience'' and that ''might reasonably be attributed to the organism itself as its contribution to the task of the acquisition of knowledge'' (Chomsky, 1971). Candidates for such invariances include the *structure dependence* of grammatical rules, and in particular, certain constraints on question formation. Various ''poverty of stimulus'' (POS) arguments suggest that these invariances reflect an innate human endowment, as opposed to common experience: Such experience warrants selection of the grammars acquired only if humans assume, a priori, that selectable grammars respect substantive constraints. Recently, several researchers have tried to rebut these POS arguments. In response, we illustrate why POS arguments remain an important source of support for appeal to a priori structure-dependent constraints on the grammars that humans naturally acquire.

Keywords: Language acquisition; Syntax; Linguistics; Statistics

1. Introduction

Environmental stimuli greatly underdetermine developmental outcomes, including physical growth, in all organisms. For example, insect and vertebrate genomes give rise to different eye lenses, compound versus simple, independent of external stimulus. Focusing on eye organization and ontogenetic maturation, as opposed to environmental variation—taking the target of inquiry to be species-invariant internal aspects of organisms—has led to clearer understanding in studies of vision. Many cases in biology fit this pattern. In our view, human

Correspondence should be sent to Robert C. Berwick, Department of EECS and Brain and Cognitive Sciences, MIT, 32D-728, 77 Massachusetts Ave., Cambridge, MA 02139, USA. E-mail: berwick@csail.mit.edu

language is another example. Before turning to details, though, we illustrate some basic points with an experimentally confirmed case of animal communication.

Foraging bees, in performing the famous ''waggle dance'' for their hive sisters, accurately indicate the location of food sources (Dyer & Dickinson, 1994; discussed in Gallistel, 2007). They do so, in part, by exploiting shared knowledge concerning two sources of navigationally useful information: the time of day, supplied by a circadian clock, and the compass direction of the sun as a function of time (i.e., the local solar ephemeris). Bees are not born with a complete representation of the ephemeris, which depends on latitude; experience plays a role. But bees can acquire the requisite representations given limited observations. Dyer and Dickson let novice bees forage only on a few late afternoons on sunny days. Despite this impoverished experience, the bees successfully navigated to food on their first exposure during heavily overcast mornings. Upon returning to the hive, they communicated correctly, indicating how to fly relative to the unseen sun. More specifically, the novice waggles were directed straight *down*, signaling other bees to fly *away* from the sun. Yet on previous forages, the sun was positioned *above* the terrain in the direction of the food. In effect, the novices ''believed'' that the morning sun was positioned above the terrain in the direction opposite the food, even though their limited late-afternoon experience provided no grounds for any such belief (Gallistel, 2007).

In this case, experience was deliberately restricted and severely impoverished with respect to the acquired knowledge, making this a ''poverty of stimulus'' (POS) example by design. While Dyer and Dickson (and Gallistel) provide some suggestive models of how the bees map their knowledge to the waggle dance, there is no plausible ''learning theory'' of *how* bees acquire the ephemeris function. Rather, the POS argument for an innate endowment motivates the search for a plausible description of what the bee contributes to its mature state of knowledge, independent of experience.

Any such argument depends on a description of the knowledge attained—for example, the compass direction of the sun at all times of the day—or in Marr's (1982) terms, the function computed. Critics can try to show that bees know less than the experiments suggest; and advocates of POS arguments must take care to not overinflate knowledge ascriptions. But critics need to focus on the full range of behavior cited in support of substantive knowledge attributions. For example, if one considered only late-afternoon navigation, one might wrongly conclude that bees do not learn a complete solar ephemeris function. Or one might mistakenly propose an entirely different means of navigation, perhaps based on odor trails.

Discovering *what* an animal comes to know, and how this knowledge is (not) related to experience, often involves seeing how the animal responds to ethologically unusual stimuli. Many POS arguments are based on evidence of this sort. Indeed, in key respects, the bee example parallels the linguistic examples discussed here. Learners can certainly acquire *some* knowledge (e.g., concerning question formation) by generalizing from typical human experience. But in our view, the task is to specify the knowledge actually acquired—in its full richness—and then ask how this knowledge could be acquired given limited experience.

At a suitably high level of abstraction, everyone can agree that humans are linguistically special. Only human infants reflexively acquire languages like English, Japanese, or American Sign Language. Infants somehow select language-related data from the ''blooming

buzzing confusion'' of the external world and develop capacities to use language in ways that go far beyond generalizing from data presented to them. In this respect, human language acquisition is like other examples of species-specific growth and development. At a suitable level of detail, everyone can also agree that experience matters. Whatever the commonalities across human languages, there are also differences. Adequate theories must say how acquiring Japanese differs from acquiring English, *and* how human children differ from other animals in being able to acquire either language (or both) given a suitable course of experience. Such differences in outcome arise from four typically interacting factors:

(1) Innate, domain-specific factors;
(2) Innate, domain-general factors;
(3) External stimuli, such as nutrition, modification of visual input in very early life, exposure to distinct languages such as Japanese-versus-English, or the like; and
(4) Natural law, for example, physical constraints such as those determining that dividing cells form spheres rather than rectangular prisms.

Different conceptions of language acquisition assign different weights to these factors, with emphasis on (2) and (3) often correlated with attention to respects in which human languages differ in learnable ways. But on any view, (1) is crucial, at least in the initial mapping of external data to linguistic experience. Regarding (4), such constraints often play a significant role in the explanation of dramatic biological change; see, for example, Carroll (2005), and we make some suggestions in Section 3 regarding its role in language. In this context, specific POS arguments are offered as part of a larger attempt to identify the role of (1), and thereby isolate the roles of other factors.

Our specific goal is to re-examine one kind of POS argument, based on speakers' knowledge concerning the relation between declarative sentences and the corresponding yes-no questions or ''polar interrogatives'' (Chomsky, 1968, 1971, 1980b). Examples (5a,b), discussed in more detail in Section 2 below, illustrate this relation.

(5a) Can eagles that fly eat?
(5b) Eagles that fly can eat

Such examples were initially used for expository reasons, deliberately simplified so they could be discussed without introducing technical notions. In the last 50 years, many other POS arguments have been offered. But the elementary cases have generated a literature on whether Factor (1) principles are needed to account for knowledge of how (5a) is related to (5b). A common idea is that Factor (2) principles, including domain-general statistical analysis applied to linguistic data, play a predominant role. And it sometimes suggested that such accounts generalize, thereby neutralizing other POS arguments.[1] But we argue that the original POS argument concerning (5) remains unrebutted, and further arguments bolster the case for assigning a significant explanatory role to (1). In Section 4, we consider some different proposals that stress Factors (2) and (3).

These alternatives include a string-substitution inference algorithm (Clark & Eyraud, 2007; Clark, Eyraud, & Habrard, 2008; Clark, 2010); a Bayesian model selection algorithm

that chooses among different types of grammars (Perfors, Tenenbaum, & Regier, 2011); and bigram or trigram statistical methods, along with a neural network model (Reali & Christiansen, 2005; Elman, 2003). In our view, these approaches do not succeed. But we share the desire to reduce any language-specific innate endowment, ideally to a logical minimum. The point of a POS argument is not to replace appeals to ''learning'' with appeals to ''innate principles'' of Universal Grammar (UG). The goal is to identify phenomena that reveal Factor (1) contributions to linguistic knowledge, in a way that helps characterize those contributions. One hopes for subsequent revision and reduction of the initial characterization, so that 50 years later, the posited UG seems better grounded. If successful, this approach suggests a perspective familiar from biology and (in our view) important in cognitive science: focus on the internal system and domain-specific constraints; data analysis of external events has its place within those constraints.

At least since the 1950s, a prime goal for theoretical linguists has been to uncover POS issues and then accommodate the discovered facts while *reducing* the domain-specific component (1). The motivation for such work is clear: The complexity and diversity of descriptive linguistic proposals present deep challenges for any bio-linguistic investigations that address the acquisition, evolution, and neural bases of human language. After arguing that certain knowledge is due to ''the innate schematism of mind,'' as opposed to ''the data of experience,'' responsible nativists try to account for the knowledge attained with the sparest plausible language-specific schematism. But one begins by striving for an adequate conception of the knowledge attained, even if that requires a substantive UG.

In Section 2, we discuss some basic observations regarding questions like (5a). Initially, we try to avoid theoretical description of these facts, which any account must address. Section 3 outlines a potential explanation of central phenomena, from a modern grammatical standpoint, with an eye toward some larger issues. This ''minimalist'' proposal aims to reduce the relevant aspects of Factor (1) to a pair of simple assumptions about grammatical structure. In Section 4, we argue that the alternatives noted above—which stress Factors (2) and (3)—do not handle the original examples, and they do not cover the extended set of examples. We conclude that the best overall strategy, for identifying the relative contributions of (1–4) to human linguistic knowledge, is to formulate POS arguments that reveal a priori assumptions that theorists can reduce to more basic linguistic principles.

2. POS revisited: Empirical foundations

Consider a simple yes-no (polar interrogative) question like (5a), repeated below.

(5a) Can eagles that fly eat?

As illustrated by (5b) and (5c), the auxiliary verb *can* may modify either *eat* or *fly*:

(5b) Eagles that fly can eat
(5c) Eagles that can fly eat

But (5a) is unambiguously the yes-no question corresponding to (5b). The question corresponding to (5c) is perfectly coherent: Is it the case that eagles that can fly do eat? But (5a) cannot be used to ask this question. Rather, (5a) must be understood as asking whether or not eagles that do fly can eat. Competent speakers of English know this, upon reflection, raising the question of how speakers come to know that (5a) is *un*ambiguous in this way.

The original POS questions regarding yes-no questions were often posed by noting, as in Chomsky, 1975:39, ''declarative-question pairs…the man is here–Is the man here?…the man will leave–will the man leave?'' Early accounts of question formation were also framed in terms of transformations, the idea being that a declarative and its polar interrogative were derived from a common ''deep structure'' that served as a basis for interpretation. The POS question concerning (5a) was thus explicitly connected to questions about which *interpretations and* grammatical forms could be competently paired with word strings.

While the original POS facts were described in tandem with development of the transformational framework, many responses to the facts are possible, at least in principle. But before considering the role of any factor (e.g., external experience) in any one case, it is worth stressing that the phenomenon of constrained ambiguity is ubiquitous in the languages that human children naturally acquire. This is why questions concerning (im)possible form/interpretation pairings have been central to modern generative grammar. And note that such questions can be raised in the absence of any deep semantic theory. Facts concerning ambiguity provide much of the initial *data* for theories of what children and adults know about the meanings of linguistic expressions. Moreover, since there are boundlessly many ''negative'' facts to know, their apparent diversity invites a search for unifying principles.

To take another much discussed kind of example, (6) has only the reading paraphrased by (6a), while (7) has only the reading paraphrased with (7b):

(6) Darcy is easy to please
(6a) It is easy for relevant parties to please Darcy
(6b) #It is easy for Darcy to please relevant parties
(7) Darcy is eager to please
(7a) #Darcy is eager for relevant parties to please him
(7b) Darcy is eager that he please relevant parties

Here, ''#'' marks the *absence* of a logically possible reading for the string of words in question. There is nothing wrong with (6b), (7a), or the thoughts expressed with these sentences. On the contrary, they can describe individuals also described with (6a) and (7b). One can know that it is easy for Darcy to please Elizabeth, and that Darcy is eager for Elizabeth to please him, while also knowing that (6) and (7) *cannot* be used to report these facts. Moreover, competent speakers of English know that (8) is as ambiguous as (9).

(8) Bingley is ready to please
(9) The goose is ready to eat

Yet somehow, speakers know that (6) is unambiguous—with *Darcy* understood as the *object* of *please*—while in (7), *Darcy* is taken to be the *subject* of *please*.

Put another way, language acquisition is not merely a matter of acquiring a capacity to associate word strings with interpretations. Much less is it a mere process of acquiring a (weak generative) capacity to produce just the valid word strings of a language.[2] Idealizing, one can say that each child acquires a procedure that generates boundlessly many meaningful *expressions*, and that a single string of words can correspond to more than one expression. A word-sound, like *bank* or *pen*, can be homophonous—that is, shared by several words that may exhibit distinct types (e.g., *noun* or *verb*). Likewise, the sound of a word-string can be shared by structurally distinct expressions, as in (10).

(10) The boy saw the man with binoculars
(10a) The boy [saw [the [man [with binoculars]]]]
(10b) The boy [[saw [the man]] [with binoculars]]

Here, square brackets indicate an assignment of phrase structure to the verb phrase. The structures in (10a) and (10b) correspond, respectively, to readings which imply that *the man had binoculars* or that *binoculars were used*. And note that (13)

(13) The senator called the donor from Texas

is two-but-not-three ways ambiguous. It can mean that the senator called the *donor* who was from Texas, or that the senator made a *call* from Texas to the donor, but not that the *senator* both called the donor and was himself from Texas. So while *called the donor* and *from Texas* can each be used to ascribe properties to an individual, (13) cannot be used to ascribe both properties to the senator. By contrast, (14) can be used in just this way.

(14) The senator from Texas called the donor

More generally, a word-string will have *n* readings for some number *n*, raising the question of how speakers can acquire a grammar that permits that many readings (as opposed to more, or fewer, than *n*). And the ''constrained homophony'' phenomenon must be accounted for, even if the explanation suggests a substantive language-specific innate endowment.

From this perspective, having *zero* readings is a special case that can help make the more general point vivid. For example, speakers know that (15) and (16) are defective on any interpretation; see Ross (1967).

(15) *Who did you hear the rumor that Bob kissed?
(16) *How many movies do you have friends who acted in?

These are not examples of word salad. But neither are they good ways of expressing the thoughts awkwardly expressed with (15a–16a).

(15a) For which person did you hear the rumor that Bob kissed them?
(16a) For which number do you have friends who acted in that many movies?

And while it may be unsurprising that (17) is word salad, a more interesting fact is that (18)

(17) *May before we there been have
(18) *We may been have there before

cannot mean that we may have been there before. This suggests a constraint on how expressions can be generated; see Chomsky (1957).

Such considerations concern *what* knowledge gets acquired, given ordinary experience, not *how* it gets acquired. In terms of Marr's (1982) contrast between ''level one'' and ''level two'' explanations, one can distinguish (a) an abstract mapping from courses of experience to grammars, from (b) various algorithms for computing this mapping. And knowledge acquired, via some combination of ontogeny and learning, can be manifested in many ways.

In particular, a string of words can be comprehensible *and* defective; see Higginbotham (1985). A speaker who would never use (19) can still know that it has the meaning of (20), and not the meaning of (20a).

(19) *The child seems sleeping
(20) The child seems to be sleeping
(20a) The child seems sleepy.

Our capacity to understand (19) is no mere capacity to ''repair'' the string by converting it into an expression of English, since (20a) is well-formed and similar to (19). Descriptions of linguistic competence must accommodate such facts.

Returning to yes-no questions, speakers often agree about how strings cannot be understood, even when unavailable interpretations are *more* expected than mandatory interpretations. Given the list in (21), one might expect the declarative (21a), as opposed to (21b).

(21) hiker, lost, kept, walking, circles
(21a) The hiker who was lost kept walking in circles
(21b) The hiker who lost was kept walking in circles

Yet even if one focuses attention on (21a), (21c) is heard as the yes-no question corresponding to (21b).

(21c) Was the hiker who lost kept walking in circles?

One way or another, the auxiliary verb *was* is associated with the matrix verb *kept*—and not *lost*, which is embedded in a relative clause—as indicated with (21d).

(21d) Was [[the hiker who lost] [__ kept walking in circles]]

One could easily invent a language in which (21c) is homophonous, with (21e) as the *preferred* reading, given what we know about hikers and getting lost.

(21e) Was [[the hiker who __ lost] [kept walking in circles]]

But the relevant constraint trumps considerations of coherence. Note that (22) is a bizarre question, not the polar interrogative corresponding to the sensible (22a).

(22) Was the hiker who fed the granola fed the parking meters?
(22a) The hiker who was fed the granola fed the parking meters

In this context, recall that (5a)—repeated below as (23)—*cannot* be understood as the question indicated with (23a). In noting that (23) must be understood as in (23b), one highlights an interesting fact that illustrates a pervasive phenomenon.

(23) Can eagles that fly eat?
(23a) [Can [[eagles that __fly] eat]]
(23b) [Can [[eagles that fly] [__ eat]]]

One needs a robust account of constrained homophony—an account that handles a wide range of cases, and not just a few examples—to even *begin* discussing how children manage to assign meanings to word strings in human ways. Specifying a procedure that associates (23) with its meaning is of limited value if that procedure also associates (23) with other meanings that (23) does not have.

Moreover, the meaning of (23) clearly depends on its bracketing in ways that go beyond whether or not the auxiliary verb can be understood as somehow rooted in the relative clause. For example, as shown by (23b), *eagles that fly* must be a constituent that serves as the subject of the question. Such examples raise fundamental questions: How do speakers know that words separated in a string can exhibit semantic/phrasal relations that are prototypically exhibited by adjacent words; and how do children come to know under what conditions this is possible?

One can describe aspects of speaker knowledge in terms of displacement, or pairings, without theoretical commitments concerning the responsible principles or mechanisms (syntactic or interpretive). And initially characterizing attained knowledge is the first step toward understanding *how* that knowledge is attained as in Marr (1982). By considering related pairings, theorists can also gain further insight. We offer a few examples to illustrate a much wider range—within and across languages—that call for systematic explanation.

In English, *do* can replace the auxiliary verb *can* (or the main verb *is*), since *do* bears morphological tense (*cf. did*) but is otherwise semantically null. We indicate the actual position of interpretation with **dv**, and the logically coherent but incorrect position by **dv***, now using this notation freely to indicate constraints on ambiguity/homophony.

(24) [do [eagles that **dv*** fly] **dv** eat]

But the notation remains descriptive: The point is that (24) is unambiguous, with *do* understood as related to *eat* and not *fly*. And for these purposes, we take no stand on why *do* is *pronounced* in the fronted position. (Compare *Eagles that fly do eat?*) We simply highlight the fact that in acquiring English, one acquires a grammar according to which *Do eagles that fly eat* fails to have the coherent interpretation indicated with **dv***.

Moreover, in languages that lack a dummy tense marker like *do* (e.g., German), the entire tensed verb can head the interrogative sentence:

(25) [Essen Adler [die **v*** fliegen] **v**]

Moreover, the same form appears in various constructions in languages that exhibit VSO (verb-subject-object) order, even though these need not be questions. McCloskey (2009) offers Irish examples; see also Chung and McCloskey (1987), McCloskey (1991, 1996).

(26a) [gcuirfidh [sí isteach **v** ar an phost]]
put-future she in for the job
''She will apply for the job''

(26b) [An gcuirfidh [sí isteach **v** ar an phost]]
Interrog put-future she in for the job
''Will she apply for the job''

The details concerning *which* verb-like expressions can appear at the front of a sentence depend on the language in question. But many examples testify to a substantive constraint on the position of *interpretation* for a fronted verb-like expression. However, this constraint is described—in terms of a constraint on extraction, or a mandatory principle of construal—one wants explanations for the relevant pairing facts to apply cross-linguistically.

Probing further, examples like (27) demonstrate that a sentence-initial question word need not be associated with the semantic or underlying subject position.

(27) [can [there **v** be [eagles that eat while flying]]]

Here, the position for interpretation follows the *surface* subject *there*, not the underlying (semantic) subject *eagles that eat while flying*. The question is whether eagles that eat while flying can exist. (Compare, *Are there eagles that can eat while flying?*)

To repeat, talk of pairings is simply a way of reporting certain unambiguity facts, not a theory-laden description, much less an explanation at all. But the facts extend to adjectival constructions and *wh*-words (*what*, *who*, *which book*, etc.), as shown below, where positions for correct interpretation are marked with **a** and **w**, respectively, while illicit positions for interpretation are indicated with **a*** and **w***:

(28a) [Happy though [the man who is tall] is **a**], he's in for trouble
(28b) [Though [the man who is tall] is happy], he's in for trouble
(28c) [Tall though [the man who is **a***] is happy], he's in for trouble
(29a) [What did [the man who bought the book] read **w**]
(29b) [What did [the man who bought **w***] read]

The constraints on **v** and **w** pairings partly overlap but are not identical, suggesting that speaker knowledge can be subtle. In (29a,b) the legitimate **w** position is in the main clause, not the embedded clause. But as (30) below shows, while *what* may be paired with the **w** position that lies within an embedded clause, *that eagles like* **w**, in contrast, *will* cannot be paired with the **v*** position in that same embedded clause. Note that (30a), but not (30b), is a possible ''echo question'' counterpart to (30).

(30) [What will John **v** warn [people that we read **w*** to p] [that eagles **v*** like **w**]]
(30a) John will warn people that we read to that eagles like *what*?
(30b) John warns people that we read to that eagles will like *what*?

To be sure, some languages may not exhibit pairings like those in (23)–(30) due to other factors. For example, some languages might not form questions with *wh*-words, and so lack analogs of (30). But where such pairings are possible at all, the general constraints look the same. This uniformity motivates the search for a principled account of why children, with different linguistic experience, should respect the same abstract constraints.

This suggests that explanations for **v**-pairing facts, as opposed to mere descriptions of these explananda, should meet at least four conditions:

I. Yield the correct pairings, for unboundedly many examples of the sort described;
II. Yield the correct structures, for purposes of interpretation, for these examples;
III. Yield the correct language-universal patterning of possible/impossible pairings;
IV. Distinguish v–pairings from w-pairings, in part, while also accounting for their shared constraints.

Criteria I–IV impose a considerable burden on proffered *explanations* for any fact illustrated with an expository example. In particular, they exclude proposals that do not even attempt to account for the pairings and the various options for interpretation. Proposals that do not extend beyond (23)—or even worse, provide descriptions of mechanisms that generate only surface strings of words, rather than the correct bracketed structures—are not responses to the original POS argument illustrated with (23). We return to these points in Section 4.

Chomsky (1968, 1971, 1980b) addressed the questions raised by (23) in terms of a grammatical rule relating (23) to (23b). The idea was that whatever the correct formulation, such a *rule* must make reference to the *structure* (i.e., bracketing) of the sentence. Simply ''counting''—say, until reaching the first occurrence of *can,* and ignoring sentential structure—yields the wrong results. The theoretical issue was framed (Chomsky, 1968:61–62, 1971, 26–27) by imagining a learner faced with the task of accounting for such declarative/question pairs by means of two competing rule hypotheses, H1 and H2: ''H1 simply moves the left-most occurrence of *is* to the front of the sentence; H2 first identifies the subject noun phrase, and then moves the *next* occurrence of *is* to the front of the sentence.'' (Chomsky, 1971:26). Following convention, we call this H2 rule ''V-raising.'' Its generalization to other categories, as described in examples (24)–(30), is called ''raising.''

Crucially, rule H1 refers only to the analysis of the sentence into individual words—or at most part of speech labels—along with the property ''left-most.'' That is, H1 does not depend on the sentence structure and consequently is called a *structure-independent* rule. By contrast, H2 refers to the abstract label ''noun phrase,'' a grouping of words into constituents, and consequently is called *structure dependent.* In this case, the crucial domain-specific Factor (1) principle seems to be the structure dependence of *rules*; see Chomsky (1968, 1971, 1975, 1980a,b), Crain and Nakayama (1987:522).

Theorists can then describe the examples above in terms of two additional principles, which happen to overlap in the case of subject relative clauses. For the V case, the pairing

(or raising) does keep to minimal distance, but ''minimal'' is defined in structural (phrase-based) rather than linear (word-based) terms: The paired/raised element is the one structurally closest to the clause-initial position. More generally, there are no ''counting rules'' in language (for further discussion, see Berwick & Weinberg, 1984). For all cases, the second descriptive principle is that subject relative clauses act as ''islands'' (see Ross, 1967), barring the pairing of an element inside the relative clause with an element outside it (whether an auxiliary verb, a verb, a *do*-auxiliary, an adjective, or a *wh*-word).

For the most part, the status of these further principles remains an open question, though we note below in Section 3 that linear order seems to be a reflex of the sensory-motor system, and so unavailable to the syntax and semantics we describe there. So, while we suspect that such constraints might reflect Factor (1) principles, antecedent domain-specific knowledge, Factor (4) principles, related to computational efficiency, may play a role (Berwick & Weinberg, 1984). The more important point is that a series of related POS observations can together impose serious adequacy conditions, like I–IV, on accounts of what speakers know. And if speakers know more than they learn from experience, then when offering theoretical descriptions of what speakers know, an important goal is to reduce appeals to (1).

3. An optimal general framework

What kind of system covers the examples just discussed, adequately capturing *what* knowledge speakers acquire, while minimizing any posited language-specific innate endowment? In our view, this question is unavoidable given the POS considerations illustrated by the elementary examples. Though as outlined in Section 1, we are not trying to say *how* linguistic knowledge is acquired or implemented. These questions are important, but methodologically secondary: First, theorists need a decent account of what is acquired/implemented. But we contend that learning takes place within constraints of the sort outlined below. This approach, if successful, identifies aspects of linguistic knowledge that need not be learned. In this sense, the aim is to delimit the boundaries of what an acquisition model must accomplish.

Any algorithm that meets conditions I and II, specified at the end of Section 2, generates unboundedly many expressions (e.g., mental representations that can be depicted with phrase markers) that can be paired with strings of lexical (atomic) elements. An algorithm that exhibits constrained homophony—in that a given string may be paired with more than one structure, but not every logically possible structure—presumably generates structures in accord with laws of some kind. But before thinking about sources of constraint, one wants to be clear about the necessary conditions for unbounded generation of any kind.

Modern logic, which has been intertwined with the study of arithmetic notions, has revealed various ways of describing generative procedures. But one way or another, such procedures require a primitive combinatory operation that forms larger elements out of smaller ones. (The requisite operation can be described in terms of a Peano-style axiom system, a Fregean ancestral, a Lambek-style calculus with ''valences,'' or in other ways.) We assume that some aspect of human biology implements a primitive combinatory operation

that can form complex expressions from simpler ones, which themselves may or may not be atomic. Call this basic operation—whatever its source, and without further stipulations about how it should ultimately be characterized—merge.

At a minimum, merge takes as input two available syntactic objects—X and Y, each an ''atom'' for computation (drawn from the lexicon), or else previously constructed by merge from such atoms—and constructs a new extended object. In the simplest case, X and Y are unchanged and unordered by the merge operation, so that merge(X, Y) can be represented simply as the set {X, Y}. We will refer to the condition that X and Y are unchanged as the ''no-tampering condition,'' a general principle of efficient computation, since altering X or Y would require further computational steps. Imposing an order on X and Y also requires a computation that may not belong to the syntactic-semantic component of language.

This is not to deny that an order is imposed, at least for purposes of speech production. But linear ordering of words may reflect *externalization* of linguistic structures by the sensory-motor system (see Berwick & Chomsky, 2011), as opposed to the generative procedures under consideration here: While {Z, {X, Y}} differs from {Y, {X, Z}} in ways suggestive of how *Bob saw Al* differs from *Al saw Bob*, mere word order seems to be irrelevant for semantic composition. In any case, we start with the idea that merge(X,Y) = {X,Y}. To suppose otherwise is to add a stipulation concerning the primitive operation of linguistic combination—and in that sense, to posit a more complex computation.

Further assumptions will be needed to describe various linguistic phenomena. But the basic combinatorial operation, however simple or complex, is surely manifested in many interaction effects involving independent cognitive systems. This makes it hard to discover the nature of merge and the expressions generated via this operation. But if merge is an aspect of the endowment that makes human language possible, then an obvious research strategy is to start with the simplest conception that meets the ''no-tampering condition,'' and see where (if anywhere) this conception of merge proves inadequate.

If X is a lexical item and SO is any syntactic object, their merger is the set {X, SO} with SO traditionally called the *complement* of X. For example, *see the man* is a verb phrase consisting of the verb *see* and its noun phrase complement *the man*, in which the noun *man* is combined with a determiner, *the*. For expository convenience, the verb phrase can be represented as the set {**v**, {**det**, **n**}}, using part of speech labels.

Since merge can apply to its own output, without limit, it generates endlessly many discrete, structured expressions; where ''generates'' is used in its mathematical sense, as part of an idealization that abstracts away from certain performance limitations of actual biological systems. Each syntactic object formed via merge has properties that enter into further computation, including semantic/phonetic interpretation: A verb phrase functions differently from a noun phrase. In the best case—that is, if phrasal properties are determined in an especially simple way, requiring few if any additional factors—this information will be carried by a designated element of the syntactic object. In {**v**, {**det n**}}, **v** can serve as this ''category marker'' or *label* for the phrase. Ideally, the label will also be locatable by an efficient search procedure, perhaps as a result of some Factor (4) principle; this in fact turns out to be the case in computer-implemented versions of the approach we sketch (Fong, 2011), as further mentioned briefly below.

If speakers employ a more sophisticated procedure to classify (say) verb phrases as such, then the role for Factor (1) principles may be even larger. But again, we want to start with the sparest assumptions about how properties like *being a verb phrase* and *being a noun phrase* might be realized and distinguished. The idea is emphatically not that complex operations are biologically computed in surprisingly efficient ways. The hypothesis is rather that the core linguistic operations are simple enough to be computed by whatever biology underpins the generative processes that are exhibited by natural language grammars.

In this context, let us also set aside the interesting question of which labeling algorithms are optimal *in general*, and confine attention to the simple but frequent case of a lexical item (or ''head'') H merging with a phrasal complement XP to form {H, XP}. At least here, it is clear that an optimal/minimal search algorithm should locate H as the label of {H, XP}: the simplest algorithm would look at H and XP, detect H as a lexical item, and so not search into XP. In our example, **v** would be identified as the head of {**v**, {**det**, **n**}}—making this expression a verb phrase, regardless of how {**det**, **n**} is labeled. Since this simple model seems to be empirically adequate, we adopt it, absent contrary evidence.

Say that Y is a *term of* X if Y is a subset of X or a subset of a term of X. If we think of Y merged to X, there are two possibilities: Instances of *external* merge, where Y *is not* a term of X; and instances of *internal* merge, where Y *is* a term of X. In either case, the result is {X, Y}. But if Y is not a term of X, then (external) merger with X yields {X, Y} in which there is *just one* instance of Y. By contrast, if Y is a term of X, then (internal) merger with X yields {X, Y} in which there are *at least two* instances of Y, one of which was externally merged to some term of X.

This distinction apparently has interpretive/articulatory correlates. External merge typically underlies *argument structure*, as in {admires, her}; where *admires* merges with and is not a term of *her*, which is thereby assigned some thematic role. Ignoring irrelevant details, the verb phrase *admires her* might then in turn merge with *John* to form {John, {admires her}}, with *her* receiving its normal intonation when this sentential structure is ultimately articulated. Internal merge typically underlies *non-argument relations*, encompassing discourse effects and scope related interpretation. For example, {John, {admires her}} might merge *internally* with *her* to form {her, {John, {admires her}}}, whose articulation as, *her John admires her* is characterized by (a) an intonation peak on the ''new'' information indicated with the fronted and pronounced occurrence of *her*; and (b) a second, unpronounced occurrence of *her*. But note that *her* retains its thematic role from this latter externally merged occurrence.

Appeal to internal merge might be called a ''copy theory of movement.'' Given the no-tampering condition, internally merging Y with X will yield an expression {X, Y} such that Y is an *unaltered* copy of some constituent of X: It would take extra computational work to delete or modify either copy of Y in {X, Y}. So the possibility of expressions containing multiple copies of the same expression—and in that sense, the possibility of copies ''displaced'' from positions created by external merge—''comes for free,'' because it requires no additional stipulations, following directly from positing an operation that must be invoked on any view. A *ban* on copies would require special explanation. Correlatively, one can describe the pairing facts described in Section 2 in terms of displacement, without

appeal to any special operation of ''forming'' or ''remerging'' copies; see Hornstein (2000). From this perspective, the possibility of ''movement'' (i.e., copies) is a *consequence* of making the *fewest* assumptions about merge.

In this context, consider the simplified structure (31a),

(31a) $[_{\text{Comp}}$ [you wrote *what*]]

formed by external merge and headed by a complementizer (Comp) that is typically not pronounced in modern English. (Compare the more Chaucerian *the which that you wrote.*) Taking all of (31a) to be X, and letting Y be *what*, Y is a term of X. Merging X with the covert complementizer, and then applying internal merge to Y yields (31b),

(31b) $[_{\text{Spec}}$ *what* $[_{\text{Comp}}$ [you wrote *what*]]]

with the new instance of *what* in the so-called *Specifier* position of Comp. Again, we assume that pronunciation of the initial (lower) instance is suppressed.

It should be apparent that internal merge—the special case where X or Y is a term of the other—yields ''raising'' constructions of the sort discussed in Section 2. The structurally lower occurrence of *what* in (31b) is in the proper position for semantic interpretation as an *argument* of *wrote*. The structurally higher occurrence is in the proper position to be pronounced and interpreted as an *operator* ranging over the construction, with the resulting meaning being roughly ''whatever Φ such that you wrote Φ.'' This description of the facts meets requirement II noted at the end of Section 3: Given the basic assumptions outlined here, concerning linguistic ''atoms'' and phrases, internal merge generates structured objects that provide the requisite positions for interpretation.

To repeat, supposing that merge operates freely—allowing for both external and internal merge as possibilities, exhibiting slightly different properties as a matter of course—is the simplest theoretical option. Given a biology that somehow implements merge, why expect it to allow for *only* external merge? There is no sense in which external merge is a simpler operation than merge itself, and no sense in which internal merge is a more complex operation than external merge. Theorists can posit an additional constraint/stipulation that rules out internal merge: This is to hypothesize that the operation permitting construction of *you wrote that* and *you wrote what* can be used to construct *whether you wrote that* but not *what you wrote what*. But this would be to posit more than the minimum required in order to explain how a system could generate unboundedly many phrases. Moreover, given an additional constraint/stipulation that limits merge to external merge, one would need to posit an additional operation of combination to account for the facts (describable in terms of pairings, as in Section 2) which suggest that human grammars generate expressions like (31b). The net result would be to require a *richer* substantive UG—that is, a *greater* reliance on the domain-specific Factor (1) principles—compared with the assumption that merge operates freely. So rejecting this minimal assumption, in favor of some alternative, requires additional empirical evidence. More specifically, evidence would be needed for the double stipulation of *barring* internal merge (or external merge) and then *adding* new descriptive technology to replace what internal merge and

external merge do without stipulation. Absent such evidence, we keep to the simplest merge-based system.

As with (31a), (5a) may now be rewritten as (32), with two copies of *can*, with the structurally *lower* copy indicating the proper position for the interpretation associated with *eat*, and the structurally *higher* one indicating the proper position for pronunciation:

(32) [can [eagles that fly] can eat]]

The relation between the declarative and interrogative is thus established via the internal merge operation and the externally merged copy. Note that the **v** notation, used earlier to describe the pairing facts, can *now* be seen as more than a mere expository convenience: (32) captures the relevant pairing, with the fewest possible stipulations, by exhibiting a syntactic structure transferrable to language components responsible for articulation and interpretation. We assume that at the former (morpho)-phonological stage—involving a ''performance'' component that makes its own ''pass'' through the syntactic structure, linearizing it for speech output—pronunciation of the second occurrence of *can* is suppressed. This is no different in spirit than the assumption operative since Chomsky (1957), according to which some component maps *s*+*run* into *runs*, ''hopping'' the *s* to the verb stem. The details are of interest for purposes of developing a full account of human language that includes input and output systems (i.e., parsing and production); see, for example, Fox and Pesetsky (2004). But we abstract from the (morpho)-phonological details, since we take our claims here to be compatible with any independently plausible specific proposal.

Running through examples (24)–(36), one can easily check that in each case, the copying account fills in the legitimate locations for **v**, **dv**, **a**, or **w** interpretation, meeting our requirements (I) and (II), and most of (III). For example, in (28a), repeated below as (33), *happy* is interpreted properly in its position after the predicate *is*.

(33) [Happy though [the man who is tall] is happy], he's in for trouble
Compare: though the man who is tall is happy, he's in for trouble.

Finally, to capture the constraints on pairings, we need to add the two language-dependent principles mentioned earlier: First, for **v**-pairing, the ''raised'' **v** is the one structurally closest to the clause-initial position; second, in all cases, subject relative clauses act as ''islands.'' Given these additional assumptions, flagged at the end of Section 2, criteria (I)–(IV) are satisfied for these relatively simple examples.

As we have emphasized, our main goal is to account for the representations and generative machinery required to describe *what* competent speakers know. So for the most part, we have abstracted from computational considerations. But since we have sometimes adverted to computational considerations, as with the ability to ''check'' features of a head/label, this raises a legitimate concern about whether our framework is computationally realizable. So it is worth noting that the copy conception of movement, along with the locally oriented ''search and labeling'' procedure described above, can be implemented computationally as an efficient parser; see Fong, 2011, for details.

The same minimal system can also account for more complex cases in which quantificational structure cannot simply be read off surface word order. This is especially desirable, since such cases present a serious POS argument that might otherwise lead to a puzzle about where the requisite innate endowment could come from. For instance, the copying account renders (34a) as (34b):

(34a) [[which of his pictures] did they persuade the museum that [[every painter] likes best?]]]
(34b) [[which of his pictures] did they persuade the museum that [[every painter] likes [which of his pictures] best?]]]

This seems correct, since in (34a), *which of his pictures* is understood to be the object of *likes*, analogous to *one of his pictures* in (35).

(35) [they persuaded the museum that [[every painter] likes [one of his pictures] best]]]

In (34b), the lower copy of *which of his pictures* occupies the correct position for interpretation. Further, the quantifier-variable relationship between *every* and *his* in (34a) is understood to be the same as that in (35), since the answer to (34a) can be *his first one*—different for every painter, exactly as it is for one of the interpretations of (35). By contrast, no such answer is possible for the structurally similar (36), in which *one of his pictures* does *not* fall within the scope of *every painter*:

(36) [[which of his pictures] persuaded the museum that [[every painter] likes flowers?]]]

Here too, then, the correct (and different) structure is supplied by (34b). In which case, a learner need not acquire any new knowledge to properly interpret such examples, even if they initially seem to be especially complicated.

4. Other accounts

In response to the facts outlined in Section 2, other proposals have been advanced, often with a prominent role assigned to learning from contingent experience. In this section, we review three recent and illustrative examples, in light of the desiderata listed at the end of Section 2. We endorse the general aim of reducing the linguistic domain-specific Factor (1). But as we show, these proposals do not address the original POS problem concerning the structure dependence of linguistic rules. Nor do they address the cross-linguistic examples; and so far as we can tell, the alternative proposals cannot be extended to generate the attested broader patterns of correct and incorrect pairings. Moreover, even with regard to the basic examples, we think the alternatives suffer from either or both of two serious defects: They do not aim to specify the correct *structures* for interpretation (along with correct pairings, as emphasized in Section 2), and so they fail to capture what speakers know about the basic examples; or they do aim to capture the right pairings, at least implicitly, but fail to do so. One might be able to specify a learnable procedure that generates the correct

polar interrogative *word strings* of English. But absent any reasons for thinking that such a procedure will meet other empirically motivated desiderata, this project seems orthogonal to that of understanding how Factors (1)–(4) conspire to yield human language. And as discussed in Section 2, the POS considerations at issue here arose in the context of attempts to say how word strings are related to interpretations *via* grammatical forms.

4.1. String-substitution for acquisition

We begin by considering a string-based approach to learning motivated by some of Zellig Harris' proposed ''discovery procedures'' for grammars (e.g., Harris, 1951) and pursued in a series of papers, Clark and Eyraud (2007); Clark et al., 2008; Clark, 2010); hereafter, CE. CE advance an inference algorithm, that, given examples like (37a) and (37b), generalizes to a much larger derivable set that includes examples like (37c), while correctly excluding examples like (37d).

(37a) men are happy.
(37b) are men happy?
(37c) are men who are tall happy?
(37d) *are men who tall are happy?

Briefly, the method works by weakening the standard definition of syntactic congruence, positing that if two items *u* and *v* can be substituted for each other in a *single* sentence context, then they can be substituted for each other in *all* sentence contexts. For example, given the string *the man died* and the string *the man who is hungry died*, by CE's notion a learner can infer that *the man* and *the man who is hungry* are intersubstitutable in these (and so all) sentences. Similarly, given the new sentence *the man is hungry*, intersubstitutability of *the man* and *the man who is hungry* licenses *the man who is hungry is hungry*.

CE call this notion *weak substitutability* to distinguish it from the more conventional and stronger definition of substitutability, which does not extend existential substitutability to universal substitutability. This extension does the work in CE's system of generalizing to examples that have never been encountered by a learner. Weak substitutability imposes a set of (syntactic) congruence classes, a notion of constituency, on the set of strings in a language. For example, *the man* and *the man who is hungry* are said to be in the same congruence class. This yields an account of sentence structure, in the limited sense of how words are grouped into phrases. But CE's proposal fails for two basic reasons.

First, it fails for English, Dutch, and in all likelihood other natural languages, even when restricted to only the *strings* that a language generates (i.e., a language's weak generative capacity). Second, it does not address the original POS question, which—as Section 2 illustrate—depends on which *structures* a language generates (i.e., a language's strong generative capacity).

Regarding the first point, CE themselves remark that ''it is certainly true that the grammars produced from these toy examples overgenerate radically. On more realistic language samples [weak substitutability] would eventually start to generate even the incorrect forms

of polar questions'' (2007:1742). But the issues are deeper, and they arise quickly. English is simply not a substitutable language in the relevant formal sense. For CE, the sentences *eagles eat apples* and *eagles eat* show that *eat* and *eat apples* are in the same class. However, this licenses the ill-formed *eagles eat apples apples*. While *eat* and *eat apples* are both verb phrases, the latter cannot be substituted for the former in *eagles eat apples*. Indeed, virtually no two phrases will be substitutable for each other in *all* texts. In particular, given *can eagles fly* and *eagles fly*, the substitution method implies that *can eagles* and *eagles* are congruent—and likewise, given *eagles can fly*, that *eagles can* and *eagles* are congruent—thus licensing *can can eagles fly*, *eagles can can fly*, *can eagles can fly*, and so forth.

Unsurprisingly, the points hold for other languages. Consider the West Flemish sentences in (39), from (Huybregts, 2008).

(39a) da is niets. (''that is nothing'')
(39b) wa is niets. (''what is nothing'')
(39c) da betekent niets. (''that means nothing'')
(39d) da betekent da zwart schoon is. (''that means that black beautiful is'')

Given (39a,b), *da* and *wa* are substitutable; and given (39c,d), *niets* and *da zwart schoon is* are substitutable. And given 40(a), the ensuing sequence of substitutions ends with the ungrammatical aux-inverted form 40(c).

(40a) is da niets.
(40b) is wa niets.
(40c) *is wa da zwart schoon is.

Turning to the second point, insofar as CE focus only on weak generative capacity, they cannot even represent the POS problems regarding structure dependence. As we stressed above, the original questions concerned the structures required for interpretation and knowledge of the relevant pairing facts. CE recognize this, noting that pairs of sentences can be related semantically, but that ''this does not imply that there need be any particular syntactic relation (2007:1742).'' However, they offer no alternative account of the facts reviewed in Section 2. And to recapitulate our main theme, the original POS question concerns the knowledge that certain interrogatives have certain declarative counterparts. So while CE may offer answers to questions other than the ones posed here, they do not offer an approach that handles even simple illustrations of the original POS problem.

4.2. Bayesian model selection of grammars

Recently, Perfors et al. (2011), hereafter PTR, have also considered the key question of domain-specific versus domain-general knowledge in language acquisition. Their perspective, which involves explicit focus on grammars and the structures that grammars impose on strings, differs from that of CE. Further, they apply their method to actual child-directed input from the CHILDES corpus. In these respects, we find PTR's approach congenial. But as discussed below, their result does not address what we take to be the central issues

concerning the structure dependence of grammatical *rules*. So *pace* their suggestions to the contrary, we see no reason for thinking that their result throws doubt on the initial argument: Facts of the sort discussed in Section 2 illustrate specific knowledge that is rooted in domain-specific principles, as opposed to domain-general principles of learning. If *other* facts fail to make the case for a more substantive linguistic nativism, this highlights the need to distinguish such facts from those reviewed in Section 2.

Before turning to PTR's model however, we need to flag an exegetical complication that corresponds to a fundamental and much discussed linguistic distinction. Taking grammars to be *procedures that generate* expressions (I-languages, in Chomsky's, 1986 sense, directly mirroring Church, 1941 terminology for his interpretation of the lambda-calculus), one can distinguish grammars that generate structured expressions only via structure-dependent principles—of combination or transformation—from grammars that can generate structured expressions in more permissive ways, allowing for rules like ''front the *first* auxiliary verb.'' And note that the expressions (trans)formed by a structure-independent rule can themselves be structured. The distinction here concerns the generative procedures, not the generated objects. This matters, since many POS arguments purport to reveal constraints on the expression-generating procedures that children can naturally acquire. If skeptics demand evidence that children acquire grammars that generate structured expressions, as opposed to mere word strings that can be classified in certain ways, one can respond with facts of the sort reviewed in Section 2. But these facts were not offered merely, or even primarily, as rejoinders to those skeptical of appeals to any mentalistic/structural notions. The deeper point is that with regard to human language, the core generative procedures seem to have a character not explained by domain-general learning.

The exegetical complication is that PTR begin their discussion by suggesting that their proposal will differ from that of Chomsky (1965, 1971, 1980a,b) with regard to how children acquire generalizations concerning ''the hierarchical phrase structure of *language*,'' why children appear to favor ''hierarchical *rules that operate on grammatical constructs* such as phrases and clauses over linear rules that operate only on the sequence of words,'' and whether children ''must innately know that syntactic *rules* are defined over hierarchical phrase structures rather than linear sequences of words'' (Perfors et al., 2011:306; including the abstract, our italics). But after saying that their goal is to ''reevaluate'' POS arguments for such conclusions (2011:307), PTR note that their way of framing the issue differs from Chomsky's. And later, they characterize their goal as follows (2011:310, our italics): ''to show that *a disposition to represent syntax* in terms of hierarchical structure rather than linear structures need not be innately specified as part of the language faculty, but instead could be inferred using domain-general learning and representational capacities.''[3]

Setting aside questions about PTR's precise goals, they are surely right to say that learners make ''a significant inductive leap'' when inferring that expressions are structured. And it is well worth knowing how this leap is made. But even if a Bayesian learner can acquire grammars that generate structured expressions, given child-directed speech but no additional language-specific knowledge, this does not yet make it plausible that such a learner can acquire grammars that exhibit constrained ambiguity of the sort illustrated in Section 2. Such grammars generate expressions *in certain ways*. In particular, children acquire

grammars that generate expressions in accord with specific structure-dependent rules that govern interpretations; see Crain and Pietroski (forthcoming). So especially if the generative principles can be plausibly reduced, as suggested in Section 3, the question is whether learners can induce that expressions are generated in these (human) ways.

So in our view, the main issue is not whether a Bayesian learner can acquire grammars that generate structured expressions. For us, the issue is whether PTR's model of such acquisition makes it plausible that a similar model—with similar stress on general learning as opposed to domain-specific constraints—can capture the constrained ambiguity facts. The theoretical task is not to show how children might acquire grammars that generate structured expressions in ways that would make these facts puzzling. Rather, the task is to describe grammars so that these facts follow from more basic generative principles, which may reflect a language-specific innate endowment. But so far as we can tell, PTR's model does not include or suggest hypotheses about how structured expressions are generated in accord with the apparently language-specific constraints discussed above.

What is PTR's learning model? In brief, PTR propose a notion of a ''Bayes learnable'' grammar, employing a method advanced by Solomonoff (1964) and located within a coherent Bayesian probabilistic framework by Horning (1969). Horning begins with the traditional notion of an evaluation metric for grammars, a proxy for grammar size that estimates a grammar's a priori probability as directly proportional to the number of symbols in that grammar—making smaller, more compact grammars more probable, and so more highly valued. A grammar's prior probability is then adjusted by how well it ''fits'' the data it receives: The a priori grammar probability is multiplied by that grammar's empirical *likelihood* (the probability that the data is generated by the grammar being evaluated). The resulting computation yields the Bayesian *posterior* probability of a particular grammar *given* a particular corpus. Given a set of grammars and a training corpus, the most highly valued and selected grammar is the one with the highest posterior probability.

Horning's original formulation of an ideal Bayesian learner thus goes beyond a naïve empiricist model. For it admits both antecedently given information, as encapsulated by a grammar's prior probability, and a ''goodness of fit'' measure that reflects the probability of the observed data given the grammar. But despite establishing some general learnability results with respect to Bayesian learning of probabilistic context-free grammars, Horning did not implement his method. Importantly, PTR have implemented Horning's original formalism, applying it to a particular CHILDES corpus of child-directed utterances, in order to draw conclusions about what types of grammars ''best fit'' this data.

From their analysis, PTR reach two main conclusions that are relevant here. First, they address the ''inductive leap'' mentioned above: ''given typical child-directed speech and certain innate domain-general capacities, an ideal learner could recognize the hierarchical phrase structure of language without having this knowledge innately specified as part of the language faculty'' (Perfors et al., 2011:306). Second, they note that their best-ranked ''context-free grammars can parse aux-fronted interrogatives containing subject NPs that have relative clauses with auxiliaries—Chomsky's critical forms—despite never having seen an example of these forms in the input'' (2011:325).

Crucially, however, it does not follow that such learners will acquire grammars in which *rules* are *structure dependent.* On the contrary, as we show below, the acquired grammars may still operate structure-independently. In short, inferring that language is hierarchical (in PTR's sense) leaves the original POS question untouched, and their Bayesian model does not explain the constrained ambiguity facts. Let us examine why.

PTR consider three candidate grammar *types* for covering their corpus: memorized finite lists of part of speech sequences; stochastic context-free grammars generating such sequences, where each rule has a particular production probability; and regular right-linear grammars, either found by automatic search or derived from covering context-free grammars. The finite list, which calls for memorizing each particular part of speech sequence as a special case, can be immediately excluded as a viable option for any realistic acquisition model. Not only is each sentence totally unrelated to every other, but storage quickly grows far beyond any reasonable memory capacity. And of course, there is no way to address the basic conditions (I) or (II) of Section 2.

That leaves the context-free and right-branching regular grammar types. So it is quite important, for this discussion, that the ones *not* selected by a Bayesian learner be grammars that *don't* assign hierarchical structures of the relevant sort to strings. But if a grammar is unbounded, in any sense of interest for the study of human languages, then one or more expression-combining operations must be applicable to their own outputs (via recursion or some logical equivalent, like a Fregean ancestral). And given such operation(s), any *derivational sequence* for a generable *string* will determine a hierarchical structure—thus providing one notion of structure generation (strong generation) distinct from the (weakly generated) string. For instance, let f be a successor operation that applies to an element a so that $f(a) = a$; $f(f(a)) = aa$; etc. Any nested object in this series can be identified with some string a^n. But any such string can still be described as an instance of the derivational structure associated with $f(\ldots(f(f(a)))\ldots)$, even if it is also (weakly) generated by other recursive procedures. In this sense, finitely statable grammars that generate boundlessly many expressions will at least associate expressions with structures, even if the expressions are strings.

As a concrete example, consider PTR's most highly valued (highest posterior probability) *regular* grammar (found by local search) labeled REG-B, level 6 (the last line of table 3 of Perfors et al., 2011:321). According to this grammar, *eagles can fly*—or more precisely, the part of speech string ''n aux vi''—has the parse displayed in Fig. 1(A).

Even in this simple case, the regular grammar assigns hierarchical structure to the input. Longer parsed sequences display even more nested structure, as shown in Fig. 1b. Obviously, these structures do not provide plausible analyses of English sentences (though they are drawn from PTR's acquired grammars). The point is simply that in one perfectly fine sense, PTR's regular grammars assign hierarchical structures to strings. In this sense, their learner did not have a chance to acquire an ''unbounded'' grammar that does *not* assign such structures. So it crucial to bear in mind that PTR have *not* offered an argument that their Bayesian learner prefers unbounded grammars that assign such structures to unbounded grammars that do not.

Their learner *can*, however, distinguish among *types* of grammars—and in that sense, choose among different kinds of hierarchical structure: uniformly right- or left-branching

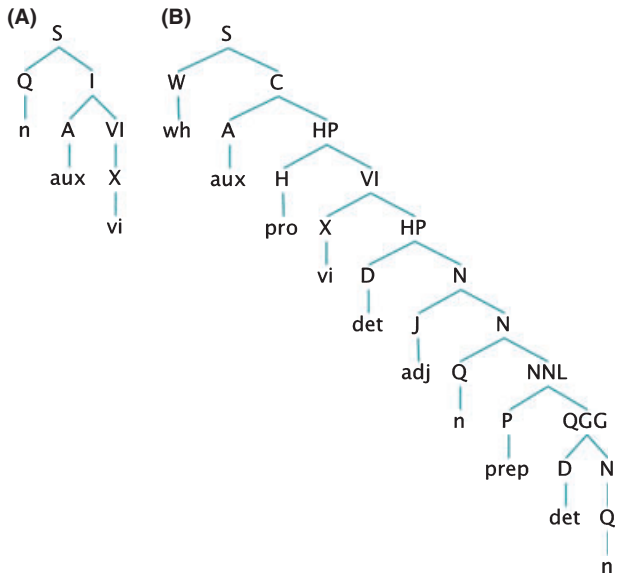

Fig. 1. (A) Regular grammar parse of ''n aux vi.'' (B) Regular grammar parse of the more complex sentence corresponding to the part of speech sequence, ''wh aux pro vi det adj n prep det n.'' Both parses use the REG-B grammar from Perfors et al. (2011) applied to CHILDES part of sequences from Perfors et al.'s (2011) training set.

structure, of the sort generated by regular grammars, as opposed to the more complex tree structures produced by general context-free grammars. And in fact, as PTR themselves say in their note 1, they take hierarchical grammars to be those in which ''constituents (e.g., noun phrases) are coherent chunks that can appear in different locations in a generically tree-structured syntactic parse'' (2011:308). So their argument is that at least for the corpus data in question, an ideal Bayesian learner will prefer grammars that yield ''generically tree-structured'' parses (generated by general context-free grammars) to grammars that do not (right- or left-regular grammars). (For PTR, an ''ideal'' learner is one that can search the space of hypothesizable grammars in ways that are relatively unconstrained by considerations of computational feasibility; see Perfors et al., 2011:315.)

This is welcome confirmation, applied to a dataset derived from actual child-directed input, for the independently plausible claim that regular grammars are descriptively inadequate for natural language: They do not provide the resources needed to capture *what* competent speakers know, much less how they know it. But what of the learner's choice between structure-dependent and structure-independent rules? At this point, PTR seem to assume that if a grammar generates expressions that exhibit hierarchy, then the rules defined over these expressions/structures must be structure dependent: ''if the hierarchical nature of phrase structure can be inferred, then any reasonable approach to inducing rules defined over constituent structure should result in appropriate structure-dependent rules'' (Perfors et al., 2011:313). The antecedent of this conditional is automatically satisfied if it is known that the target grammar is unbounded. But in any case, and as noted above, the inference is fallacious: Structured expressions can be (trans)formed by a structure-independent rule—for

example, fronting the first auxiliary. One can of course proceed to say that such a grammar is ''unreasonable,'' in a technical sense, if it permits structure-independent generation of expressions. But this is to assume a logically contingent *and presumably language-specific* constraint on the space of ''reasonable'' grammars. From a domain-general perspective, there is nothing wrong with generating unboundedly expressions via operations that allow for (say) structure-independent ways of forming interrogatives. Indeed, this was one moral of examples like (5a).

(5a) Can eagles that fly eat?

More generally, structure dependence of rules is an abstract property of certain grammars, and it is orthogonal to the property of generating expressions that exhibit hierarchical constituency. But since PTR speak in terms of ''what rules *defined over* constituent structure *should* result in'' (2011:313, our italics), it may be worth noting that a flat—and in that sense *non*-hierarchical—sequence of phrases like [[$_{NP}$ det n] [$_{VP}$ aux v]] could be modified by structure-*dependent* rule, like ''invert the NP and VP,'' yielding a different sequence like [[$_{VP}$ aux v][$_{NP}$ det n]]. So again, one needs to ask what PTR's learner tells us about the abstract properties of grammars revealed by the examples from Section 2.[4]

To test whether or not Bayesian learners prefer grammars that exhibit structure dependence of rules, one needs to examine a hypothesis space with a greater variety of linguistic models, or model *types* in PTR's sense. To frame the structure-dependent/independent question, one could add a set of rules defined over *both* the regular and context-free grammars, with one set of rules being structure dependent, and the other not—yielding *four* possible model choices. Then one could see whether Bayesian analysis favors models with structure-dependent rules. In more detail, one might draw on any linguistic theory that systematically accounts for the relevant pairing facts. For example, given Generalized Phrase Structure Grammar or Head Driven Phrase Structure Grammar (HPSG), the pairing of (5a) with (5b)

(5b) Eagles that fly can eat

might be described in terms of a meta-rule such as: Given the form [$_{VP}$ X V X], there is the form [$_{Q}$ V NP X], with V an auxiliary verb and Q denoting a question phrase (Gazdar, 1981). Or in terms of HPSG, one could adopt a lexical rule to capture the auxiliary-inversion property. The same augmentation could be applied to PTR's regular grammars. But it is now less clear what the results of comparative Bayesian analysis would be, especially for a more realistic corpus with further computational constraints on how the space of hypothesizable grammars is actually searched.

On the one hand, the addition of meta-rules (or lexical mapping rules) can often reduce grammar size, replacing two sets of rules with one; it is well-known from formal language theory that the consequent compression gain for a context-free grammar may be exponential or even greater (Meyer & Fischer, 1969; Berwick, 1982) when compared to a regular grammar that (weakly) generates the same strings. On the other hand, adding such rules threatens to enrich the role of Factor (1) principles. Moreover, adding meta-rules threatens massive over-generation, as outlined in Uszkoreit and Peters (1986) or Ristad (1986). So while

grammar compression would likely boost a priori probability—the first factor in Bayesian analysis—over-generation would tend to reduce the *likelihood*, or corpus ''fit,'' as PTR note. The outcome remains an open question. In any case, PTR do not carry out the requisite comparisons to distinguish (unbounded) grammars that only generate expressions via structure-dependent rules and ''more permissive'' grammars. So on this score, no conclusions can be drawn from their results.

Of course, critics of POS arguments might settle for a Scotch Verdict on issues concerning the structure dependence of rules. And PTR report that they, ''along with many cognitive scientists, are less sure about whether explicit movement rules provide the right framework for representing people's knowledge of language, and specifically whether they are the best way to account for how a child comes to understand and produce utterances like (4c) but not like (4b)'' (2011: 310, their examples below):

(41) (= PTR's 4b) * Is the boy who smiling is happy?
(42) (=PTR's 4c) Is the boy who is smiling happy?

However, the issue is not just about how speakers understand (4c), or how (4b) is recognized as degenerate. As stressed in Section 2, one also needs to explain why *Is the boy who lost left* fails to have the following interpretation: (*is it the case that*) *the boy who is lost left?* Why does this string only mean: *the boy who lost is left?* Likewise, while *Is the boy who smiling left* is as degenerate as PTR's example (4b), one wants to know why this string fails to have even *one* reading: *the boy who is smiling left?* PTR might say that a string fails to have any meaningful reading if it cannot be assigned a (legitimate) syntactic structure. But recall examples like (19), which are defective but meaningful:

(19) *The child seems sleeping

And in any case, the *un*ambiguity of *Is the boy who lost left* still calls for explanation. Repeating an earlier point, having zero readings seems to be a special case of having *n* but not *n + 1* readings. So other things equal, one wants a common explanation for why all such strings fail to have the readings they would fail to have given the apparent constraints on movement discussed in Sections 2 and 3.

Put another way, marking (4b) with an asterisk notes a fact that calls for explanation. The account we favor involves a system ideally reduced to the notion of external merge, as discussed in Section 3. We are unsure what alternative explanation PTR and many others have in mind. But we do think that alternatives should be responsive to the full range of facts—concerning constrained ambiguity, and not *just* contrasts like (4b/c)–which motivated the initial proposals that posited constrained displacement operations.

Insofar as PTR's model is not intended to cover such cases, or explain such facts, this is no complaint. PTR are free to focus on (what they take to be) ''the more basic question of how and whether a learner could infer that representations with hierarchical phrase structure provide the best way to characterize the set of syntactic forms found in a language'' (2011:312). But they add, ''We see this question as the simplest way to get at the essence of the core inductive problem of language acquisition posed in the generative tradition''

(2011:312). From our perspective, this tendentious theoretical claim—about what the ''core'' inductive problem is—is not empirically motivated. We see no reason for thinking that PTR's question is more ''basic'' than those concerning how grammars pair sounds with interpretations *via syntactic forms*. From this perspective, the core problem may be that of acquiring a generative procedure that gives rise to certain kinds of constrained ambiguity. And these questions animated many of the original proposals in generative linguistics. Again, this is not to say that PTR's model fails to address their own questions. The point is simply that in assessing POS arguments, one needs to be clear about the premises and conclusions.

Like PTR, we are unimpressed by arguments that ''consider some isolated linguistic phenomenon that children appear to master and conclude that because there is not enough evidence for that phenomenon in isolation, it must be innate (2011: 330).'' But we have highlighted some better arguments that focus on available interpretations for strings, and how syntactic structures are generated by constrained basic operations. In this context, we have located examples like *Can eagles that fly eat* within a broader range of cases that (so far as we can tell) PTR's model does not accommodate.

4.3. Learning from bigrams, trigrams, and neural networks

Reali and Christiansen (2005), hereafter RC, constructed three sets of models to explore the acquisition of yes-no questions: (1) a bigram statistical model; (2) a trigram statistical model; and (3) a simple recurrent neural network (SRN) model, to conclude that ''there is indirect statistical information useful for correct auxiliary fronting in polar interrogatives and…such information is sufficient for distinguishing between grammatical and ungrammatical generalizations, even in the absence of direct evidence'' (Reali & Christiansen, 2005:1007).

Like PTR, RC used a corpus of child-directed speech as training data. For their first two models, RC computed the frequency word bigrams or trigrams and then an overall sentence likelihood for any word sequence, even for previously unseen word sequences. This sentence likelihood was then used to select between opposing test sentence pairs similar to *Are men who are tall happy—Are men who tall are happy*, the idea being that sentences with the correct auxiliary fronting would have a greater likelihood than those with incorrect auxiliary fronting. We note at once that engagement with the original POS arguments would presumably call for discussion of *Are men who married like women* and why it fails to have the meaning: *men who are married like women?* But in what follows, we review RC's experiments with bigram models—relying on Kam et al.'s thorough review of this model type (2008)—with an eye toward further assessment of RC's trigram model, as pursued by Kam (2007). We then turn to RC's neural network model.

Reali and Christiansen (2005) Experiment 1 demonstrated that on 100 grammatical-ungrammatical test sentence pairs the bigram likelihood calculation successfully chose the correct (grammatical) form 96% of the time. Kam et al.'s replication, using a slightly different training corpus and a different test pairs, got 87% correct. But as Kam, Stoyneshka, Tornyova, Fodor, and Sakas (2008) note, this success seems attributable more to accidental

facts about English and the specific test sentences, rather than the bigram method itself. The bigram model apparently exploits the fact that *who* and *that* are homographs, textually ambiguous as to whether they are pronouns or complementizers/relativizers. For example, the bigram probability of *Is the little boy who is crying hurt* is greater than that of *Is the little boy who crying is hurt*. Kam et al. note that this results entirely from the high bigram frequency of *who-is* in the first sentence, as opposed to *who-crying* in the second. But the high bigram score for *who-is* reflects its high-frequency occurrence in main sentences, with *who* as a pronoun, rather its appearance in relative clauses. Kam et al. concluded that the training corpus examples spuriously boost the bigram frequency for *who-is*. And when they scored examples without such ''winning bigrams'' then the bigram performance declined precipitously. Performance also dropped when *who* and *that* were removed by replacing their occurrences in relative clauses with the nonce labels *who-rel* and *that-rel*. To be sure, this might have the ''knock-on'' effect of removing otherwise useful cues for a word or construction-based account of acquisition. As Kam et al. observe, ''an alternative conclusion might be that homography (homophony) is a good thing, which permits learners to bootstrap from a word form (superficially defined) in one context to its occurrence in another context'' (2008:785, note 11). But as they go on to remark, absent concrete evidence on this matter, one can only speculate about such possibilities.

4.3.1. RC's trigram model

RC extended their bigram approach to a *trigram* model, calculating sentence likelihoods according to three-word frequencies. Kam (2007, 2009) analyzed this model in detail, confirming that trigram performance essentially tracked that of bigrams, succeeding (and failing) for the same reason that bigrams did. In what follows below, we extended Kam's findings regarding trigrams, by adapting Kam et al.'s homographic removal methodology to RC's original corpus and test sentences (see the last row of Table 1, Experiment 6). With this information removed, trigram performance degraded significantly. We conclude that the homography issue that Kam et al. uncovered as the apparent source for the discriminative power of bigrams also arises for trigrams, so that neither bigrams nor trigrams serve as a convincing model for auxiliary-inversion learning. Details follow below.

To test RC's trigram approach, we used child-directed utterances from two versions of the Bernstein-Ratner (1984) corpus, one used by Kam et al., with 9,634 sentences, and the second supplied to us by RC, with 10,705 sentences. For Experiments 1 and 2, we replicated Reali and Christiansen's (2005) Experiment 1 and Kam et al.'s (2008) Experiment 1, using first the 100 test sentence pairs constructed by RC and second the 100 different grammatical and ungrammatical test sentence pairs constructed by Kam et al. Each test sentence was in the template form, ''*is* NP {*who*|*that*} *is* A B''/''is NP {*who*|*that*} A *is* B''. We computed the bigram and trigram frequencies for each pair according to the smoothing formulas specified in RC.

The first two lines of Table 1 display our results. The % correct, incorrect, and undecided perfectly match both earlier reported results. The small differences between the original RC and Kam et al. results arises from using slightly different training corpora and completely distinct test sentences. The RC test sentences are nearly uniform in their ''B''

Table 1
Percentage of test sentences classified correctly versus incorrectly as grammatical or undecided, using RC's bigram or trigram analysis (Expts. 1 & 5); Kam et al.'s methodology (Expts. 2, 3, & 4); and adapting Kam et al.'s homograph removal methodology (Expt. 6)

Experiments	Sentences Tested	% Correct	% Incorrect	% Undecided
1. Replication of Reali and Christiansen (2005), bigram test	100	96	4	0
2. Replication of Kam et al. (2008), Expt. #1, bigram test	100	87	13	0
3. Replication of Kam et al. (2008), Expt. #2, Disambiguated *rel-* pronouns	100	20	38	42
4. Kam et al. (2008) Expt. #2, Disambiguated *rel-* pronouns, using Reali and Christiansen (2005) test sentences	100	69	16	15
5. Replication of Reali and Christiansen (2005), trigram test	100	95	5	0
6. Disambiguated *rel*-pronouns, trigram test, using Reali and Christiansen (2005) test sentences	100	74	11	15

portions, often with a single adjective, for example, *happy,* and not as varied as the ''B'' portion of the Kam et al. test sentences, which included full non-adjectival phrases, for example, *for Paul's birthday*. This change makes the *Kam* et al. test sentences more difficult to discriminate via bigrams alone because the displacement of *is* in the grammatical version of each sentence pair sometimes yields a collocation with a bigram frequency of 0. This can be enough to throw the decision in favor of the ungrammatical form ''A *is* B,'' despite the presence of a relatively high-frequency bigram such as *who is* in the grammatical version.

Given this replication, we proceeded in Experiments 3 and 4 to reproduce Kam et al.'s Experiment 2 using both Kam et al.'s and RC's test sentences, replacing occurrences of *who* and *that* where they introduced relative clauses with the forms *who-rel* and *that-rel*. The corresponding results are given in the third and fourth rows of Table 1. As Kam et al. observed, without *who-is* or *that-is* cues, bigram performance dropped from 96% correct to 20% correct. Similarly, using RC's original test sentence pairs, performance degrades, from 96% to 69%. While this is not as steep a performance decline, it is still a significant drop,

with the smaller decline attributable to the relative simplicity of the RC's test sentences as compared to Kam et al.'s.

Following Kam (2007, 2009), we then turned to trigrams. Experiment 5 (Table 1, row 5) displays our replication of RC's trigram analysis. RC's trigram analysis gave exactly the same correct/incorrect results as their bigram analysis, while our trigram replication made one additional mistake, classifying a single ungrammatical test sentence, *is the box that there is open,* as more likely than the corresponding grammatical form, *is the box that is there open.* The reason for the single difference is straightforward. Aside from this one sentence, all trigram frequencies in both the grammatical and ungrammatical RC test sentences are 0, so the trigram values are actually estimates based on smoothed bigram and unigram values. The exceptional test sentence pair noted above is different: The ungrammatical form contains a trigram with frequency 1, namely, *is-open-end-of-sentence-marker*, while the corresponding grammatical form does not contain any non-zero trigrams. This suffices to render the ungrammatical form more likely under trigram analysis than its grammatical counterpart.

For Experiment 6, we adapted Kam et al.'s homograph replacement methodology to the trigram case, testing the resulting trigrams using RC's sentence pairs. The results are displayed in row 6 of Table 1, with 74% correct, 11% incorrect, and 15% undecided, again a substantial decline from the near-perfect trigram performance when *who-is* and *that-is* cues were not removed. We conclude that the high accuracy for *both* bigrams and trigrams in discriminating between grammatical and ungrammatical sentences seems to be due to exactly to the effect Kam et al. found: the accidental homography between pronouns and complementizers in English, rather than anything particular to the yes-no question construction itself.

4.3.2. Learning from simple recurrent networks (SRNs)

In their final experiment, RC adopted an SRN as a learning model. SRNs are presumed to be more powerful than either bigrams or trigrams, because they contain a ''hidden'' context layer. RC trained 10 different networks on the Bernstein corpus, and then tested whether these could discriminate between grammatical versus the ungrammatical example minimal pairs as in the bigram case, such as (their 9a, b) pair, *Is the boy who is hungry nearby/Is the boy who hungry is nearby*. Note that they had to recode actual words into 1 of 14 possible part of speech categories, for example, DET (*the*), N (*boy*), PRON (*who*), V (*is*), ADJ (*hungry*), PREP (*nearby*), and the like, as described further below.

Since SRNs output a *distribution* over possible predicted outputs after processing each word, one must verify their performance in a more nuanced way. For instance, to see why the networks preferred ''grammatical'' strings to the others, RC provided the part of speech prefix of the test sentences up to the point at which grammatical versus ungrammatical divergence would occur—for example, for the pair, *Is the boy who is hungry* versus *Is the boy who hungry is*, this would be the point immediately following prefix sequence V DET N PRON. RC then examined the predicted output of the trained networks over all word categories at this point to see whether more network activation weight was assigned to the grammatical continuation V (corresponding to *is*) as opposed to the ungrammatical continuation ADJ (corresponding to *hungry*). RC confirmed that the V activation level was nearly an

order of magnitude larger than that for ADJ, in this sense confirming that the network had acquired the knowledge that the proper continuation was V (see Figure 5 in Reali & Christiansen, 2005:1021).

But how secure is this result? As Kam et al. remark, ''the only results to date, even for these stronger learners, are for the original *is-is* [construction], which we have seen is trivially acquirable by the simplest *n*-gram learner and is not a bellwether for human language acquisition at large'' (2008:783). In other words, one might wonder whether the success of the SRNs in selecting, for example, V as opposed to ADJ as above might *also* be attributable simply to ''brute statistical facts.'' Kam et al.'s and our findings above suggest that bigram information alone could account for most of the statistical regularity that the networks extract. Indeed, RC themselves come close to the same conclusion in their own discussion when they state that ''The networks are capable of distinguishing chunks of lexical categories that are more frequent in the training input from less frequent ones (that is, the lexical categories corresponding to PRON V ADJ [*who is hungry*] versus PRON ADJ V [*who hungry is*])'' (2005:1020). There are several suggestive indications that this is so, and that the bulk of even a recurrent network's power results from the sheer fact that pronouns are most often followed by verbs rather than adjectives.

While it would be easy to overstate such a conclusion, it is straightforward to see that a network might largely rely on bigram statistics even from the crudest analysis of the Bernstein corpus when recast using RC's word categories. In particular, in our version of the Bernstein corpus, the bigram PRON-V occurs 2,504 times (the highest bigram count over all bigram pairs), and PRON-ADJ 250 times, an order of magnitude difference of V over ADJ, precisely reflecting the RC's network prediction. Furthermore, the network's next most highly activated word category given the prefix V DET N PRON prefix is N, with the (visually estimated) value of 0.13. Once again, a simple bigram fits this result fairly well, with the PRON-N bigram count of 893 being the next highest-frequency bigram count after PRON-V. (Note that for RC the category PRON included possessive pronouns like *her*, so it is unsurprising that PRON would often be followed by the category N, as in *her toy*.)

We examined this idea more carefully by replicating RC's SRN experiment as closely as possible given their published data, ensuring that our training corpus, test sentences, and procedures matched theirs. We first tagged the Bernstein corpus words with one of 14 part of speech categories as in RC, and then, following RC, trained simple recurrent networks consisting of RC's Elman-style architecture (Lewis and Elman, 2003). As with RC, we initially used 10 different SRNs (with different random weight initializations), training them on several passes through the corpus. However, for completeness, since network weights can depend on random seeding, we also trained SRNs consisting of 50, 100, 250, 1,000, and 2,500 individual networks, as well as carrying out a partially exhaustive search of the recurrent network parameter space to ensure that the results were relatively stable and insensitive to initial conditions.

The first two rows of Table 2 display the comparison between RC's results and ours. There are some differences. In particular, the activation level obtained for ADJ is somewhat higher in our networks than RC's (0.040 vs. 0.03); and our activation level for V somewhat lower (0.184 vs. 0.23). But the order of magnitude difference between the activation levels

Table 2
Simple recurrent neural network activation levels for V and ADJ categories given 10 simple recurrent networks following a given prefix consisting of part of speech categories, as in Reali and Christiansen (2005). Reported values display the mean activation level averaged over 10 trained networks, with values in parentheses indicating standard deviations

Experiment	Activation Level of V Category	Activation Level of ADJ Category
Prefix: V Det N PRON (Reali & Christiansen, 2005)	0.23	0.03
7. Replication of Reali and Christiansen (2005) V Det N PRON	0.184 (0.0381)	0.055 (0.0025)
8. Det N PRON	0.150 (0.0140)	0.054 (0.0018)
9. PRON	0.173 (0.0117)	0.055 (0.0015)

for V versus ADJ agrees with RC's result. The correspondence between the two network simulations is displayed more graphically in Fig. 2: Here, open triangles correspond to RC's mean network activation levels while our networks' mean activation levels are displayed as the gray histogram bars for each part of speech category, with standard deviation brackets as shown. The notable outlier is the ADV category, where the RC network predicted an average activation level of approximately 0.13 as compared to 0.047 for our networks. While we cannot pinpoint the exact reason for this discrepancy, we suspect that it may be due to the random fluctuations inherent in the seeding and convergence of such networks. Given that we ran our simulations over many hundreds of thousands of initial weight sets, all with the same resulting mean activation levels for ADV, we were reasonably confident that our SRN prediction was stable. Putting this difference to one side, we were otherwise satisfied that we had replicated RC's trained SRNs, and we proceeded to further explore the bigram hypothesis by testing our networks on different word category prefixes, with the results shown in the last two rows of Table 2, rows 8 and 9.

In each case, we supplied a different prefix, and then recorded the comparative SRN activation levels for the possible predicted categories of V versus ADJ. The first two test prefixes were Det N and PRON. On the assumption that the SRN was in fact simply acquiring knowledge about bigram ''lexical chunks'' we would again expect that the category V should be more highly activated than ADJ. The results confirm this prediction: V is preferred over ADJ by approximately the same amount as with the prefix V DET N PRON. The same preference for V over ADJ holds for similar prefixes generally, for example, V PRON or DET ADJ N PRON (data not shown). This again seems unsurprising, since the bigram PRON ADJ is rare relative to PRON-V.

Given these suggestive results, in a final experiment we directly tested whether the SRNs were to a large extent extracting a statistical regularity corresponding to bigram frequencies. We calculated the bigram predictions for the alternative V versus ADJ continuations given a preceding PRON, comparing this bigram prediction to the SRN predictions for V versus ADJ when given the prefix V DET N PRON. Fig. 2 displays the results. Bigram predictions

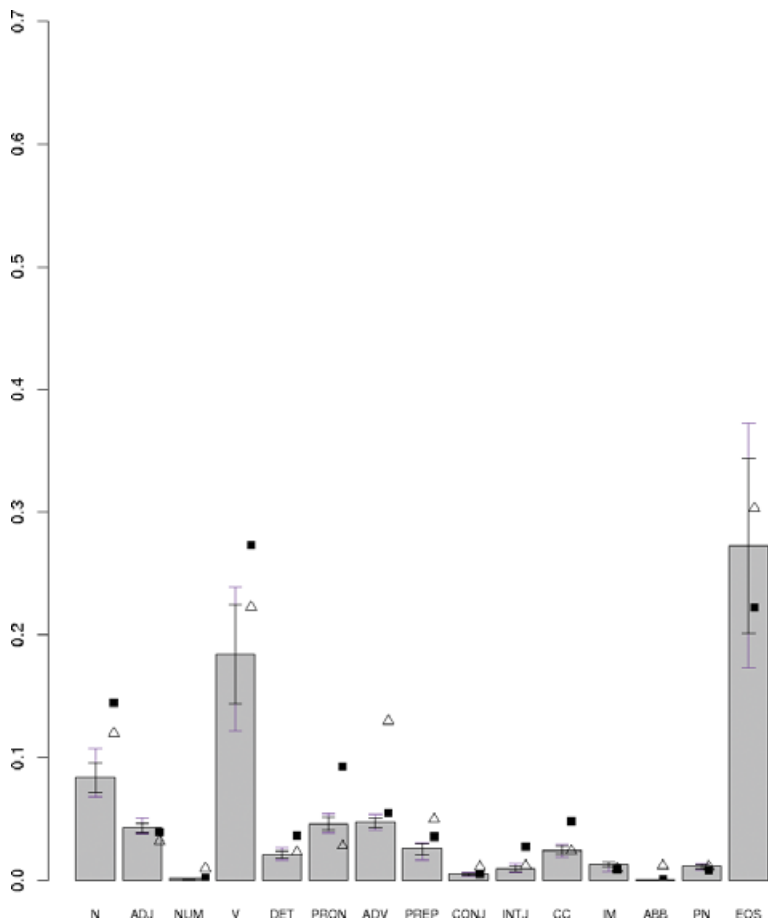

Fig. 2. Mean SRN activation levels (gray bars) compared against Reali and Christiansen's (2005) predictions (open triangles) and bigram predictions (black squares), given the prefix V, Det, N, PRON, for 10 trained SRNs. Brackets on each histogram bin indicate standard deviations. EOS is the beginning/end of sentence marker (not reported in Reali & Christiansen, 2005, but estimated from their plotted figure). SRN, simple recurrent neural network.

for the part of speech category following PRON are shown as black squares, with our and RC's network predictions as previously noted, histogram bars and open triangles, respectively. The bigram model correctly selects V over ADJ, by roughly the same order-of-magnitude proportion as expected from the raw bigram counts, though it is clear that the PRON-V bigram value is even larger than the SRN activation level. Fig. 2 also suggests that the SRN and bigram levels are often in general correspondence, rising and falling together, with high bigram frequencies corresponding to high activation levels, though there are some outstanding differences. One is the outlier noted previously, RC's network category ADV, which we consequently set aside. A second discrepancy is between the absolute bigram frequencies and network activation levels for N and V, where the bigram values exceed cor-

responding SRN activation levels. Despite these differences, the *relative* ordering of bigram and SRN activation levels are in general agreement. In short, one would not go too far wrong in simply using bigrams as proxies for SRN activation levels, and this indicates the extent to which the SRNs have essentially drawn on the bigram regularities.

Since SRNs are explicitly designed to extract sequentially organized statistical patterns, and given that the *is-is* question types can be so easily modeled by sequential two-word patterns, this is not at all surprising. Indeed, it is difficult to see how SRNs could possibly fail in such a statistically circumscribed domain. Further, it is known that recurrent networks bear a close relationship to so-called auto-regressive moving average (ARMA) statistical models, of which bigrams are a particularly simple and degenerate case (Sarle, 1994; Resop, 2006). From this perspective, it is also unsurprising that SRNs would *not* exploit the bigram evidence already explicitly shown by RC and Kam et al. to be present in the corpus data.

In any case, it remains to be seen how an SRN would deal with more complex interrogatives of the sort examined by Crain and Nakayama (1987), for example, *Is the boy who was holding his plate crying,* where the matrix auxiliary *is* differs from the relative clause auxiliary *was.* Note that under RC's encoding, both auxiliaries are mapped to the single category V, along with main verbs. So as it stands the question cannot even be posed. Similarly, the homograph replacement methodology cannot be applied, because both possessive, personal, and complementizer forms of *who* and *that* are all mapped to the same category, PRON. Until these extensions are carried out, we find that we must agree with Kam et al. that the SRN results, along with the bigram and trigram results remain far from compelling, being ''trivially acquirable by the simplest *n*-gram learner and … not a bellwether for human language acquisition at large.''

5. Conclusion: The POS remains

In acquiring a grammar for a language, a child acquires a procedure that pairs sounds with interpretations in particular, constrained ways. While ambiguity is ubiquitous, endlessly many examples like (5a) are strikingly *un*ambiguous.

(5a) Can eagles that fly eat?

Taking each grammar to be a biologically implementable procedure that generates complex expressions from a finite stock of primitive expressions, the phenomenon of constrained ambiguity raises a pair of related questions: Which expression-generating procedures do children acquire; and how do children acquire these procedures? Answers must do justice to the facts concerning what examples like (5a) do and do not mean, in English and in other languages. Proposals can fail in many ways. Here, we stressed the fact that a procedure might pair sounds with *more* interpretations than competent speakers permit.

In Section 2, we reviewed some familiar reasons for thinking that human grammars generate complex expressions via *structure-dependent* operations that respect substantive constraints. With regard to (5a), the idea is that a prohibition against extracting an auxiliary

verb from a relative clause precludes generation of any expression that pairs the sound of (5a) with the following interrogative meaning: (is it true that) eagles that can fly eat? If this is correct, one wants to know how children acquire the knowledge that respects such prohibitions. In particular, one wants to know the role of innate, language-specific factors. For on the one hand, it seems that these factors must impose substantial limits on the space of possible human grammars that can be acquired, given domain-general learning and contingent experience. But on the other hand, one wants to reduce such factors—ideally, to the minimum required in order to explain unbounded generation of expressions that can (via interfacing cognitive systems) pair sounds with interpretations.

Section 3 sketched an attempt at such reduction, in order to illustrate how POS considerations can motivate a substantive linguistic nativism *without* requiring an embarrassingly large language-specific innate endowment. One should not resist the conclusions of POS arguments—and insist that grammars respect only those constraints that reflect domain-general learning applied to ordinary linguistic experience—on the grounds that such arguments lead to non-explanatory (and/or biologically implausible) proposals about our distinctively human innate endowment. On the contrary, by taking POS arguments seriously, theorists are led to a question that turns out to be fruitful: To what degree would the simplest imaginable procedures for generating unboundedly many meaningful expressions be procedures that give rise to the observed phenomena of constrained ambiguity?

This leaves room for skeptics to reply that we have either over described ordinary linguistic knowledge, or underestimated the ways in which domain-general learning can lead to the acquisition of highly constrained expression-generating procedures. (Recall, in this regard, our introductory analogy with bee communication.) And as conceptions of learning evolve, along with technology, skeptics have ever more tools at their disposal. But in Section 4, we argued that three recent replies to some old POS arguments fail to engage with the real force of such arguments: These replies underdescribe the relevant linguistic knowledge, overestimate the power of domain-general learning, or both. This leads us to believe that some 50-plus years after examples like (5a) were initially offered, the basic points still hold—unchallenged by considerable efforts, in the interim and earlier, to show that real progress lies with the study of how contingent experience shapes cognition. In our view, the way forward begins with the recognition that environmental stimuli radically underdetermine developmental outcomes, and that grammar acquisition is a case in point. Then one can try to describe the gap between experience and linguistic knowledge attained, reduce that gap to basic principles that reflect the least language-specific innate endowment that does justice to the attained knowledge, and thereby help characterize the true role of experience in a manner that illuminates cognition.

Notes

1. See, for example, Ross (1967), Hornstein and Lightfoot (1981), Travis (1984), Chomsky (1986), Jackendoff (1993), Crain and Thornton (1998), Baker (2001), Laurence and Margolis (2001), and Crain and Pietroski (forthcoming). On

alternatives to the classical POS argument and its extensions, see Lewis and Elman (2001) and Pullum and Scholz (2002).

2. If each word-string could express any thought built from concepts indicated by the words—allowing for concepts corresponding to unpronounced subjects/objects—then even with constraints on *order*, so that *Darcy pleased Elizabeth* cannot mean that Elizabeth pleased Darcy, (6–9) would all be ambiguous.
3. In the literature—including experimental studies by Crain and Nakayama (1987), who PTR offer as a contrast to their view—auxiliary inversion is regularly used to illustrate the structure dependence of *rules*, and not (mere) hierarchical structure. PTR cite a rare exception: the transcript of an oral question-and-answer session, regarding a written conference paper, in Piattelli-Palmarini (1980). But in that paper—Chomsky (1980b), which appears in the cited volume—the relevant major heading is ''The Structure Dependence of Rules.''
4. It may be that PTR had a different inference or ''meta-rule'' in mind: If a learner chooses a context-free grammar, then *all* grammatical rules must be stated in the same context-free format. But this is even more clearly an additional language-specific assumption about the class of acquirable grammars. Further, why should learners be precluded from ''cobbling together'' generative strategies?

Acknowledgments

We would like to thank Michael Coen, Janet Fodor, Riny Huybregts, Amy Perfors, Luigi Rizzi, William Sakas, Aditi Shrikumar, Joshua Tenenbaum, and all the anonymous reviewers, for helpful discussions and comments that have greatly improved this article. We would especially like to thank Amy Perfors, Joshua Tenenbaum, and Terry Regier, along with Morten Christiansen and Florencia Reali for so graciously and generously making available their experimental materials, including grammars and training corpora, without which it would have been impossible to carry out the comparative analyses described here.

References

Baker, M. (2001). *The atoms of language*. New York: Basic Books.

Bernstein-Ratner, N. (1984). Patterns of vowel modification in motherese. *Journal of Child Language*, *11*, 557–578.

Berwick, R. (1982). *Locality principles and the acquisition of syntactic knowledge*. Ph.D. dissertation, Cambridge, MA: MIT Press.

Berwick, R., & Chomsky, N. (2011). Biolinguistics: The current state of its evolution and development. In A. M. Di Sciullo & C. Boecx (Eds.), *Biolinguistic investigations* (pp. 19–41). Oxford, England: Oxford University Press.

Berwick, R., & Weinberg, A. (1984). *The grammatical basis of linguistic performance*. Cambridge, MA: MIT Press.

Carroll, S. (2005). *Endless forms most beautiful*. New York: W.W. Norton.

Chomsky, N. (1957). *Syntactic structures*. The Hague, The Netherlands: Mouton.
Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.
Chomsky, N. (1968). *Language and mind*. New York: Harcourt, Brace, Jovanovich.
Chomsky, N. (1971). *Problems of knowledge and freedom*. London: Fontana.
Chomsky, N. (1975). *Reflections on language*. New York: Pantheon.
Chomsky, N. (1980a). *Rules and representations*. New York: Columbia University Press.
Chomsky, N. (1980b). On cognitive structures and their development. In M. Piattelli-Palmarini (Ed.), *Language and learning: The debate between Jean Piaget and Noam Chomsky* (pp. 35–54). London: Routledge and Kegan Paul.
Chomsky, N. (1986). *Knowledge of language*. New York: Praeger.
Chung, S., & McCloskey, J. (1987). Government, barriers and small clauses in modern Irish. *Linguistic Inquiry*, *18*, 173–237.
Church, A. (1941). *The calculi of lambda conversion*. Princeton: Princeton University Press.
Clark, A. (2010). Efficient, correct, unsupervised learning of context-sensitive languages. In M. Lapata & A. Sakar (Eds.), *Proceedings of the Fourteenth Meeting on Natural Language Learning* (pp. 28–37). Uppsala, Sweden: Association for Computational Linguistics.
Clark, A., & Eyraud, R. (2007). Polynomial time identification in the limit of substitutable context-free languages. *Journal of Machine Learning Research*, *8*, 1725–1745.
Clark, A., Eyraud, R., & Habrard, A. (2008). A polynomial algorithm for the inference of context free languages. In A. Clark, A. F. Coste & L. Miclet (Eds.), *Grammatical inference: Algorithms and applications, lecture notes in computer science 5728* (pp. 29–42). New York: Springer.
Crain, S., & Nakayama, S. (1987). Structure dependence in grammar formation. *Language*, *63*, 522–543.
Crain, S., & Pietroski, P. (forthcoming). The language faculty. In E. Margolis, S. Laurence, & S. Stich (Eds.), *The handbook for philosophy of cognitive science*. New York: Oxford University Press.
Crain, S., & Thornton, R. (1998). *Investigations in universal grammar: A guide to experiments in the acquisition of syntax and semantics*. Cambridge, MA: The MIT Press.
Dyer, F., & Dickinson, J. (1994). Development of sun compensation by honeybees: How partially experienced bees estimate the sun's course. *Proceedings of the National Academy of Sciences, USA*, *91*, 4471–4474.
Elman, J. (2003). Learning and development in neural networks: The importance of starting small. *Cognition*, *48*, 71–99.
Fong, S. (2011). *Minimalist parsing: Simplicity and feature unification*. Workshop on Language and Recursion. Mons, Belgium: University of Mons. March.
Fox, D., & Pesetsky, D. (2004). Cyclic linearization of syntactic structure. *Theoretical Linguistics*, *31*, 1–46.
Gallistel, C. R. (2007) Learning organs. English original of L'apprentissage de matières distinctes exige des organes distincts. In J. Bricmont & J. Franck (Eds.), *Cahier n° 88: Noam Chomsky* (pp. 181–187). Paris: L'Herne.
Gazdar, G. (1981). Unbounded dependencies and coordinate structure. *Linguistic Inquiry*, *12*, 155–184.
Harris, Z. (1951). From morpheme to utterance. *Language*, *22*, 161–183.
Higginbotham, J. (1985). On semantics. *Linguistic Inquiry*, *16*, 547–593.
Horning, J. (1969). *A study of grammatical inference* (Tech. rep. #139). Stanford, CA: Stanford University.
Hornstein, N. (2000). *Move! a minimalist theory of Construal* (2000). Oxford, England: Blackwell.
Hornstein, N., & Lightfoot, D. (1981). Introduction. In N. Hornstein & D. Lightfoot (Eds.), *Explanation in linguistics* (pp. 9–31). London: Longman.
Huybregts, R. (2008). *Linguistic argumentation and poverty of the stimulus arguments*. Utrecht, The Netherlands: University of Utretcht.
Jackendoff, R. (1993). *Patterns in the mind: Language and human nature*. New York: Harvester Wheatsheaf.
Kam, X. N. C. (2007). Statistical induction in the acquisition of auxiliary-inversion. In H. Caunt-Nulton, S. Kulatilake & I. Woo, (Eds.), *Proceedings of the 31st Boston University Conference on Language Development* (pp. 345–357). Somerville, MA: Cascadilla Press.

Kam, X. N. C. (2009). *Contributions of statistical induction to models of syntax acquisition*. PhD dissertation, New York: The Graduate Center of the City University of New York.

Kam, X.-N. C., Stoyneshka, I., Tornyova, L., Fodor, J. D., & Sakas, W. G. (2008). Bigrams and the richness of the stimulus. *Cognitive Science*, *32*, 771–787.

Laurence, S., & Margolis, E. (2001). The poverty of the stimulus argument. *The British Journal for the Philosophy of Science*, *52*, 217–276.

Lewis, J. D., & Elman, J. (2001). Learnability and poverty of stimulus arguments revisited. In B. Skarabela, S. Fish, & A. H. Do (Eds.), *Proceedings of the twenty-sixth annual Boston University conference on language development* (pp. 359–370). Somerville, MA: Cascadilla Press.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. New York: W.H. Freeman.

McCloskey, J. (1991). Clause structure, ellipsis and proper government in Irish. *Lingua*, *85*, 259–302.

McCloskey, J. (1996). On the scope of verb raising in Irish. *Natural Language and Linguistic Theory*, *14*, 47–104.

McCloskey, J. (2009). *Irish as a configurational language*. Berkeley, CA: Berkeley Syntax Circle.

Meyer, A., & Fischer, M. (1969). Economy of description by automata, grammars, and formal systems. *Mathematical Systems Theory*, *3*, 110–118.

Perfors, A., Tenenbaum, J., & Regier, T. (2011). Poverty of the stimulus: A rational approach. *Cognition*, *118*, 306–338.

Piattelli-Palmarini, M. (1980). *Language and learning: The debate between Jean Piaget and Noam Chomsky*. London: Routledge and Kegan Paul.

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

Reali, F., & Christiansen, M. H. (2005). Uncovering the richness of the stimulus: Structure dependence and indirect statistical evidence. *Cognitive Science*, *29*, 1007–1028.

Resop, J. (2006). *A comparison of artificial neural networks and statistical regression with biological resource applications*. MS thesis, College Park, MD: University of Maryland at College Park.

Ristad, E. (1986). *Computational complexity of current GPSG theory*. AI Lab Memo 894, Cambridge, MA: MIT Press.

Ross, J. (1967). Constraints on variables in syntax. Doctoral dissertation, Massachusetts Institute of Technology, *Published as, Ross, J. R. (1986). Infinite syntax!* Norwood, NJ: Ablex.

Sarle, W. (1994). *Neural networks and statistical models*. *Proceedings of the 19th annual SAS users group meeting*, Cary, NC: SAS Institute, 1538–1550.

Solomonoff, R. (1964). A formal theory of inductive inference. *Information and Control*, *7*, 1–22.

Travis, L. (1984). Parameters and effects of word order variation. Doctoral dissertation, Cambridge, MA: MIT.

Uszkoreit, H., & Peters, S. (1986). On some formal properties of metarules. *Linguistics and Philosophy*, *9*, 477–494.