Part I

Poverty of the Stimulus and Modularity Revised

2

Poverty of the Stimulus Stands: Why Recent Challenges Fail[1]

ROBERT C. BERWICK, NOAM CHOMSKY, AND MASSIMO PIATTELLI-PALMARINI

2.1 Introduction: the Poverty of the Stimulus Revisited

Environmental stimuli greatly underdetermine developmental outcomes in all organisms, including their physical growth. This is in biology at large a familiar truism and is sometimes called, in the domain of language, the 'poverty of the stimulus' (POS). For example, the distinct genomes of insects and vertebrates give rise to quite different eye lenses, compound vs simple, independently of external environmental stimulus. In this case, placing the primary focus on the actual object of study, the internal organization of eyes and their development rather than extraneous external variation, has led to clearer understanding, as in most other biological examples.

Turning to cognition, only human infants are able to reflexively acquire language, selecting language-related data from the 'blooming buzzing confusion'[2] of the external world, then developing capacities to use language that far exceed any data presented to them, much as in other areas of growth and development. The explanation for such differences in outcome arises from four typically interacting factors:[3]

(1) Innate, domain-specific factors (in the case of language, what is called 'universal grammar', obviously crucial at least in the initial mapping of external data to linguistic experience);
(2) Innate, domain-general factors;

[1] For a more detailed critique of several of these challenges see Berwick, Pietroski, Yankama, and Chomsky, 'Poverty of the Stimulus Revisited' (2011).

[2] This expression is famously due to William James (in *the principles of psychology*, 1890/1981), who characterized 'the stream of thought' and the baby's impression of the world 'as one great blooming, buzzing confusion'.

[3] This view accommodates the familiar possibility of so-called 'epigenetic effects', the interaction of external stimuli (factor 3) with innate factors 1 and 2.

(3) External stimuli, such as nutrition, modification of visual input in very early life, exposure to distinct languages such as Japanese vs English, or the like; and
(4) Natural law, e.g., physical constraints such as those determining that most dividing cells form spheres rather than other shapes, and none forms, say, rectangular prisms.

Addressing the same question, Descartes famously observed that an infant presented with a figure with three irregular sides—all that it ever experiences in the natural world—perceives it as a distorted triangle, not as a perfect example of what it actually is. In this case as well, the sample data, 'the stimulus' for selecting the correct concept 'triangle', seems too impoverished without positing antecedently the target concept in question. While Descartes's conclusion may well be too strong—the operative principle might be some kind of a priori Euclidean geometry applied to sensations yielding geometrical figures—the methodological approach stands.

The goal of this chapter is to re-examine familiar examples that were used to motivate one very elementary illustration of a POS question in linguistics, so-called yes-no questions or 'polar interrogatives,' (N. Chomsky 1968, 1971, 1980) in an attempt to determine the proper formulation of factor (1), the domain-dependent linguistic factors required to explain them. We stress at the outset that these examples were selected for expository reasons, deliberately simplified so that they could be presented as illustrations without the need to present more than quite trivial linguistic theory. They are but one example out of a wide array of POS arguments given fifty years ago. Nevertheless, this simplified example, taken in isolation, has given rise to a substantial literature, much of it attempting to show that this knowledge of language can be accounted for by resort to factor (2), for example, statistical data analysis by presumably domain-general methods. Further, it is sometimes suggested that if that effort succeeds, something significant will be revealed about the POS, perhaps even its non-existence in the case of language. As we will show to the contrary, the question of POS in the case of language would scarcely be affected even if such efforts succeeded, since one can resolve this particular POS question with very minimal assumptions about factor (1) principles (that is, UG). However, even this much is academic, since as section 2.4 below demonstrates, these approaches fail completely.

In fact, there is good reason to expect continued failure, for several reasons. First, such approaches misconstrue what is actually at stake, even in this artificially simplified example. Second, they ignore the empirical range of relevant cases from which this example was selected. Perhaps most importantly however, there are long-known, straightforward answers to this particular POS question that have far wider scope. These answers are quickly discovered if we follow standard biological methodology, as in the case of animal eye lenses mentioned earlier. No one would have dreamt of trying to account for the POS problem in the case of animal eye lenses, or innumerably many

others like it, while knowing virtually nothing about eyes. Similarly, incorporating a small part about what is actually known about language happens to yield a very simple solution to the POS problem brought up in the case of yes-no questions, while also addressing the actual issues at stake and covering a much wider empirical range. Pursuing this course also opens new and intriguing questions that have yet to be explored carefully.

Specifically, we will consider some recent attempts to deal with the simple case of polar interrogatives on the basis of domain-general procedures, factor (2) above, eliminating factor (1). These alleged alternatives include a string-substitution inference algorithm (Clark and Eyraud 2007; Clark, Eyrand, and Habrard 2008; Clark 2010), a Bayesian model selection algorithm that chooses among different types of grammars (Perfors, Tenenbaum, and Regier 2006, 2011), and a bigram or trigram statistical method (Reali and Christiansen, 2005).[4] Though these particular approaches do not succeed, we show that it is indeed possible to reduce the domain-specific linguistic component (1) quite radically, perhaps even to what may well be a logical minimum. Our alternative arises from a very different way of looking at the problem than the one adopted by these other approaches, one closer to the biological method: an analysis of the internal system of language and its constraints, rather than data analysis of external events.

More generally we note that the prime concern of serious theoretical work in linguistics since the 1950s has been to uncover potential POS issues, and then attempt to eliminate them, *reducing*, not *increasing*, the linguistic domain-specific component (1). This approach is pursued for obvious reasons: the apparent complexity and diversity of descriptive linguistic proposals raises insuperable burdens for all relevant biolinguistic questions, including the acquisition and evolution of language as well as its neural basis.

The remainder of this paper is organized as follows. In section 2.2 we lay out the basic empirical facts regarding the expository question formation examples, striving to remain neutral as to any particular linguistic formulation insofar as possible, arriving at a basic list of empirical requirements that any explanatory account must address. Section 2.3 turns to explaining the empirical data in section 2.2 from a modern grammatical standpoint—what an answer to the original problems ought to look like. It aims at reducing the linguistic domain-dependent facts (1) to a minimum. We shall see that even when we consider extensions beyond the question formation examples, very few language-specific assumptions are required to provide a simple solution to this particular problem (though as expected, new and non-trivial issues arise). Section 2.4 proceeds to assess the claimed explanatory success of the recent approaches listed above. We shall see that all these approaches collapse, both on the original examples

[4] See also Kam and Fodor, this volume, for a reanalysis of this method, with considerations and conclusions germane to our own.

and on the extended example set. We find that on balance, the elimination of POS problems and the reduction of factor (1) (the domain-dependent linguistic knowledge that must be taken as a priori) remains best advanced by current research in linguistic theory, rather than by the alternative approaches reviewed in section 2.4, a conclusion that we believe generalizes to other cases.

2.2 POS Revisited: Empirical Foundations

We begin our re-examination of the POS with the familiar expository example from N. Chomsky (1968, 1980). Consider a simple yes-no (polar interrogative) question structure as in (1a) below, where square brackets denote an assignment of phrase structure and lower-case v and v* denote possible positions for the interpretation of the word 'can':

(1a) [can [eagles that v* fly] v eat]]

For (1a) to be properly understood, the occurrence of 'can' must be interpreted in the position marked by v, not v*, yielding a question about the predicate 'eat' rather than 'fly'; the question asks whether or not eagles can eat, not whether they can fly. Assigning the proper semantic interpretation to sentences like these has always been the real question of linguistic interest. We note further that the proper interpretation of example (1a) also depends on its bracketing into phrases, that is, the assignment of a structural description to the string of items 'can eagles that fly eat'. This is necessary in order to interpret, e.g., 'eagles that fly' as a single expression that serves as the subject of the question.

How then is the choice made between the alternative positions for interpretation, v and v*? Note that the question (1a) has a clear declarative counterpart with the same semantic properties, differing only in the property of being a declarative rather than an interrogative, where 'can' replaces the correct position for interpretation, v, rather than v*, i.e.,

(1b) [[eagles that fly] can eat]

With no tacit assumptions as to the actual principles involved, we may posit that examples (1a) and (1b) constitute a *pairing*, where the second item of the pair explicitly indicates the correct position for interpretation of 'can'. Such pairings are part of the knowledge of language that children attain, attesting to the relationship between structure and interpretation. It is the relationship between such pairs that is the fundamental question of interest, as clearly posed by the earliest expository examples, e.g., 'the dog that is in the corner is hungry'—'is the dog that is in the corner hungry', with the assumed bracketing and position for interpretation marked by v as: [is [the dog that is in the corner] v happy] (Chomsky, 1968: 61–62, 1980: 39–52). It is this

knowledge, at the very least, that factors (1)–(4) above must account for, as was explicit in the earliest presentations.[5]

Further insight into this knowledge may be gained by considering related pairings beyond this simple expository example. Let us consider some of these here, with the understanding that they by no means exhaust the possibilities, but simply serve to illustrate that there is a much wider range of related pairing examples demanding explanation, both within a single language, and, even more importantly, across all languages, universally. First, in English one may also substitute 'do' for the auxiliary verb 'can' or the main verb 'is' since 'do' bears morphological tense (cf. 'did') but is otherwise semantically a dummy or pleonastic item. We denote its correct position of interpretation by dv, and its incorrect position by dv*:

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

However, in languages that lack a dummy tense marker like 'do', e.g., German, we find that the entire tensed verb may be found heading the sentence:

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

Moreover, the same form appears in various constructions in languages that have basically VSO (verb-subject-object) order, as in Irish, even though these need not be questions (examples from McCloskey 2009):[6]

(4a) [gcuirfidh [sí isteach v ar an phost]]
put-future she in for the job
'She will apply for the job'

(4b) [An gcuirfidh [sí isteach v ar an phost]]
Interrog put-future she in for the job
'Will she apply for the job?'

In other words, the element that may be paired depends on details about the language in question. Crucially, we find that in the rich variety of examples like these,

[5] Such pairings are a part of every linguistic theory that takes the relationship between structure and interpretation seriously, including modern accounts such as HPSG (Head-driven Phrase Structure Grammar), LFG (Lexical Functional Grammar), and TAG (Tree Adjoining Grammar), as also stressed by Kam and Fodor in their chapter in this volume. As it stands, our formulation takes a deliberately neutral stance, abstracting away from details as to how pairings are determined, e.g., whether by derivational rules as in TAG or by relational constraints and lexical redundancy rules, as in LFG or HPSG. For example, HPSG (Bender, Sag, and Wasow, 2003) adopts an 'inversion lexical rule' (a so-called 'post-inflectional' or 'pi-rule') that takes 'can' as input, and then outputs 'can' with the right lexical features so that it may appear sentence-initially and inverted with the subject, with the semantic mode of the sentence altered to be 'question' rather than 'proposition'. At the same time this rule makes the subject noun phrase a 'complement' of the verb, requiring it to appear after 'can'. In this way the HPSG implicational lexical rule defines a pair of the exactly the sort described by (1a,b), though stated declaratively rather than derivationally.

[6] See Chung and McCloskey (1987), McCloskey (1991, 1996) for extensive evidence for a v position in Irish.

the constraint that governs the correct choice for the position of interpretation for v continues to apply. Any explanation of pairings must therefore apply universally, cross-linguistically to cases such as (3) and (4a,b) as well as 1(a,b).

Probing further, the possibility of a construction like (1a) does not necessarily involve the semantic or underlying subject position, as illustrated in (5) below, where the position for interpretation, v, follows the surface subject 'there', not the underlying semantic subject 'eagles that eat while flying':

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

Pairings may also include adjectival constructions, (6a,b), as well as forms with 'wh' words ('what', 'who', which book', etc.), as indicated below. We again mark the position for correct interpretation via a notation for adjectives, a, or wh-words, w. Examples (6c) and (7b) illustrate that here too certain pairings are possible, while other pairings appear to violate some constraint, as marked by the illicit positions for interpretation, a* and w*.[7]

(6a) [Happy though [the man who is tall] is a], he's in for trouble

(6b) [Though [the man who is tall] is happy], he's in for trouble

(6c) [Tall though [the man who is a*] is happy], he's in for trouble

(7a) [What did [the man who bought the book] read w]

(7b) [What did [the man who bought w*] read]

The constraints on v and w pairings partly overlap but are not identical. In both (5) and (7a,b) the legitimate v or w positions are in the main clause, while the forbidden v* or w* positions lie within an embedded clause. However, example (8) below shows that the constraints on v and w pairings must be distinguished. In (8), 'what' *may* be paired with the w position that lies within an embedded clause, 'that eagles like w'; in contrast, 'will' can *never* be paired with the v* position within that same embedded clause:

(8) [what will John v warn [people that we read w* to p] [that eagles v* like w]]
cf. 'John will warn people that we read to that eagles like what'

More generally, although not all languages will necessarily exhibit pairings like those in (1)–(8) due to other, extraneous factors (e.g., some languages might not form questions with wh-words along the lines of (8)), where such pairings are possible at all, the general constraints look the same as they do in English.

Our general conclusion then is that a proposed explanation for v-pairing must meet at least the following conditions:

[7] There are of course many other possible construction pairings and constraints, including some that apparently 'violate' the embedding constraint described in the main text, but they are not relevant to the problem we address in this article. These would be part of a fully articulated theory of language, which we do not present here. We are simply illustrating example pairings that will have to be met by any fully satisfactory account.

I. Yield the correct pairings, for an infinite set of examples, those that exhaust the relevant cases;
II. Yield the correct structures, since interpretation is required by any serious linguistic/cognitive theory, also for an infinite set of examples;
III. Yield the correct language-universal patterning of possible/impossible pairings;
IV. Distinguish v- from w-pairings in part, while also accounting for their shared constraints.

Criteria I–IV impose a considerable empirical burden on possible explanations that go, as they should, beyond the simplified expository example. They exclude proposals that do not even attempt to account for the pairings and the various options for interpretation, or that do not extend beyond (1a), or, even worse, that limit themselves to generating only a correct surface string of words, rather than the correct bracketed structures. As we shall see, these problems arise for all the efforts we consider in section 2.4 below that attempt to frame an account in terms of factor (2), domain-general principles, though in fact they collapse on even simpler grounds.

As is familiar, Chomsky (1968, 1971, 1980) addressed the question of pairings like (1a,b) in terms of a grammatical rule relating the (1a) and (1b) forms, noting that whatever the correct formulation, such a *rule* must make reference to the *structure* (i.e., bracketing) of the sentence, rather than simply 'counting' until reaching the first occurrence of 'can', and ignoring the sentence structure. The question was framed (1968: 61–2, 1971: 26–7, 1980: 39–52) by imagining a learner faced with accounting for such declarative/question pairs by means of two competing rule hypotheses, H1 and H2. H1 'takes the left-most occurrence of 'is' and then moves it to the front of the sentence' (1971: 27, 1980: 39) while H2 'first identifies the subject noun phrase of the sentence' and then moves 'the occurrence of "is" following this noun phrase to the front of the sentence.' (*ibid*: 26). Let's stress here and now (though we will develop this further in section 2.4.2.1) that the mere existence of hierarchical structures and the child's access to them are presupposed. The issue that is so central to **this particular** POS problem is tacit knowledge by the child that grammatical rules *apply* to such structures. Regrettable confusion persists on this distinction in much of the work we are reviewing here. By convention, we call this movement 'V-raising', and its generalization to other categories as described in examples (2)–(8), 'raising'.[8]

[8] From the earliest work in generative grammar in the 1950s, both declaratives and corresponding interrogatives were assumed, for good reasons, to be derived from common underlying forms that yield the basic shared semantic interpretations of the paired constructions. These expressions differ only by a lexical property that in some structures 'attracts' the verbal auxiliary to the front: for example, in (i) but not in the semantically similar expression (ii):

(i) he asked 'are men happy?'

(ii) he asked whether men are happy

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, it does not depend on the sentence structure, and consequently is called *structure-independent*. In contrast, rule H2 refers to the abstract label 'noun phrase', a grouping of words into phrases, and consequently is called *structure-dependent*. In this case, the crucial domain-specific factor (1) is the *structure dependence of rules* (as is stressed in all the published work regarding this topic, see, e.g., Chomsky 1968, 1971, 1975, 1980).[9]

We can describe the examples we have covered in terms of two principles, which happen to overlap in the case of subject relative clauses. For the V case, the pairing (or raising) indeed 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 (see Chomsky 1965, 1968; Berwick 1985, for further discussion).[10] For all cases, the descriptive principle is that subject relative clauses act as 'islands', 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). Such 'island constraints' have been studied since the early 1960s.[11] Tentatively, we can take these two principles to be factor (1) principles, that is, part of antecedent domain-specific knowledge. However, at least the first principle might reasonably be regarded as a factor (4) principle, reducing to minimal search, a natural principle of computational efficiency. We will not explore the source of the second principle here, but it has been the topic of a good deal of inquiry, which also seeks to reduce it substantially to locality and related principles that might fall within a general notion of efficient computation that is language- or possibly even organism-independent.

2.3 An Optimal General Framework

How can we construct a system that will cover the empirical examples in the previous section, while minimizing the contribution of domain-dependent factors (1)? We first note that in order to satisfy conditions I and II above, such a system must yield an

[9] The 1980 publication includes a section explicitly headed 'Structure Dependence of Linguistic Rules', p. 39; in this regard, note also that Crain and Nakayama (1987: 522) concluded that their experiments 'support Chomsky's contention that children unerringly hypothesize structure-dependent *rules*.' [our emphasis].

[10] Different patterns of fMRI brain activation have been evidenced when a subject monitors structure-dependent (and therefore grammatically realistic) rules, as opposed to rules applying to words in a fixed position in a sentence (and therefore grammatically impossible) (Musso, Moro, et al. 2003). Further fMRI evidence showing this difference has recently being published by Pallier, Devauchelle, and Dehaene (2011; see also Moro's commentary in the same issue).

[11] Note that this restriction to subject relative clauses is presumably part of some broader principle; for the initial observation, see Chomsky (1962: 38–47), followed by Ross's more general account (1967: 2–13), and many improvements since. V-pairing is more constrained than wh- pairing because it evidently requires a kind of adjacency; see section 2.5 for further discussion of this constraint, which holds much more generally for lexical items that are 'atoms' for computation in the sense discussed directly below.

infinite number of discrete, structured pairs. While there are many specific methods for accomplishing this, since the latter part of the nineteenth century it has been known that any approach will incorporate some primitive combinatory operation that forms larger elements out of smaller ones, whether this is done via a Peano-style axiom system, a Fregean ancestral, a Lambek-style calculus with 'valences', or by some other means. Call this basic operation Merge.

At a minimum, Merge takes as input two available syntactic objects X, Y, each an 'atom' for computation (drawn from the lexicon), or else constructed by Merge from such atoms, and from these constructs a new, extended object, Z.[12] In the simplest case, X and Y are unchanged and unordered by the merge operation, so that Merge (X, Y) can be taken to be just the unordered set {X, Y}. We will refer to the condition that X and Y are unchanged as the 'no-tampering condition' (NTC), a general principle of efficient computation. Imposing an order on X and Y requires additional computation, which, it appears, does not belong within the syntactic-semantic component of language. There is substantial reason to suppose that ordering is a reflex of the process of *externalization* of structures by the sensory-motor system, and does not enter into the core processes of syntax and semantics that we are considering here (see Berwick and Chomsky 2009, and Chomsky's separate contribution to this volume).

Anything beyond the simplest case of Merge(X,Y) = {X,Y} requires additional stipulations and more complex computations, and therefore is to be rejected unless it receives adequate empirical support.

If X is a lexical item and Y any syntactic object (SO), then the output of Merge is the set {X, SO} with SO traditionally called the *complement* of X. As a simple example, 'see the man,' with part-of-speech labels v, det, n, traditionally written as a Verb Phrase consisting of the verb 'see' and its Noun Phrase complement 'the man', can for expository convenience be represented as {v, {det, n}}. Since Merge can apply to its own output, without limit, it generates an infinite number of discrete, structured expressions.

Each syntactic object X formed by the repeated application of Merge has properties that enter into further computation, including semantic/phonetic interpretation: a verb phrase VP functions differently from a noun phrase NP. In the best case, this information about X will be contained in a single designated element of X, its *label*, which can be located by a search procedure as the computation involving X proceeds. In the best case, the search procedure will be optimal, hence plausibly an instance of factor (4). We will put aside for the moment the interesting question of optimal labeling algorithms, noting only that in the simple case of lexical item ('head') H and complement XP, {H, XP}, the optimal minimal search algorithm will locate H

[12] There may be justification for an additional operation of *pair-Merge* that forms ordered pairs. For some discussion of this point, see N. Chomsky (2009, and this volume). In the best case, we can reduce the number of merged items to exactly two; see Kayne (1984) for evidence on this point.

as the label, thus v in {v, {det, n}}, a Verb Phrase. (More on this in Chomsky's separate contribution to this volume.)

Let us take Y to be 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, then without stipulation we have two possibilities: either Y is not a term of X, what is called *external* Merge (EM); or else Y is a term of X, what is called *internal* Merge (IM). In both cases the outputs are {X,Y}. External Merge typically underlies argument structure, as in *see the man* with 'the man' the Noun Phrase object of 'see' in the Verb Phrase {X,Y} (omitting irrelevant details). Internal Merge typically underlies non-argument structure (discourse, scope related, and the like). For example, in topicalization constructions such as 'him, John really admires n', an intonation peak is placed on the 'new' information, 'him', which is associated via Internal Merge (IM) with the position marked by n, where it receives its semantic role by External Merge (EM). This contrasts with the construction without IM operating, namely, 'John really admires him', with normal intonation.[13]

IM yields two *copies* of Y in {X,Y}, one copy internal to X, and the other external to X, in what is sometimes called the 'copy theory of movement'. Note that the fact that both copies appear, unchanged, follows from the optimal computational constraint NTC: it would require extra computational work to delete either one. Thus there is no need to explain the existence of copies, since they in effect 'come for free.' What would require explanation is a ban on copies. Furthermore, contrary to common misunderstandings, there is no operation of 'forming copies' or 'remerging copies.' Rather, the copy theory of movement follows from principles of minimal computation.

Suppose, for example, we have the structure (9a) below. Taking Y = *what* and X = the syntactic object corresponding to the structure of (9a), with Y a term of X, and applying internal Merge, we obtain the output (9b), where *what* is in the so-called 'Specifier position' of Comp:[14]

(9a) $[_{\text{Comp}}$ [you wrote *what*]]

(9b) $[_{\text{Spec}}$ *what* $[_{\text{Comp}}$ [you wrote *what*]]]

It is apparent that internal Merge—the special case where either X or Y is a term of the other—yields pairs, or 'raising' constructions of the kind discussed earlier in section 2.2: the structurally lower occurrence of *what* in (9b) is in its proper position for

[13] This distinction between structures formed via EM and those formed by IM is sometimes called the 'duality of semantics,' and is presumably part of UG. Relying on it, the child knows that in such structures as 'what eagles eat' (as in 'I know what eagles eat'), etc., 'what' is displaced from the underlying structure formed solely by EM that yields the correct interpretation of 'what' as the object of 'eat.' More complex systems that bar IM and instead add new mechanisms have to provide a more intricate account of this pervasive and quite salient duality property of semantics, which has to be captured in some way in any adequate theory of language. There are some apparent departures, presumably reflecting our current lack of understanding.

[14] This 'specifier' position itself may well be eliminable. See section 2.5 and Chomsky's separate contribution in this volume. This possibility does not bear on the discussion in this section. Here, 'Comp' stands for the 'complementizer', sometimes overt, as in 'it seems *that* John wrote something'.

interpretation (as an argument of 'wrote'), while the structurally higher occurrence of *what* is in the position where it is 'pronounced' (and, furthermore, interpreted as an operator ranging over the construction, so that the interpretation is roughly 'for which thing x, you wrote the thing x'). Thus this formulation meets requirement II. Given the two descriptive principles mentioned earlier, one for 'atoms' and the other for all phrases, IM generates a structured object that provides precisely the proper positions for interpretation.[15]

Importantly, having Merge operate freely, including both EM and IM, is the simplest option. It would require some specific stipulation to rule out either IM or EM. And it would require further stipulation to develop new mechanisms to achieve the same results as in computation with unrestricted Merge. Such stipulations to construct pairings enrich UG, the domain-specific factor (1), and therefore require empirical evidence. What would be needed is evidence for the double stipulation of barring IM (or EM) and adding new descriptive technology to replace what IM and EM do without stipulation. Lacking such empirical evidence, we keep to the simplest Merge-based system.

As with (9a), (1a) may now be rewritten as (10), with two copies of 'can', the structurally lower copy indicating the proper place for interpretation associated with *eat*, and the structurally higher one indicating the position for pronunciation:

(10) [can [eagles that fly] <u>can</u> eat]]

The relation between the (1a,b) pairs is thus established via the IM operation and the resulting copies.[16] Note that the v notation used earlier for exposition may now be seen to be more than just an expository convenience. Understood as a copy, not a notational device, it captures the pairing in what appears to be an optimal way. (10) exhibits the syntactic structure transferred to the language components responsible both for articulating and interpreting the syntactic form. It is at this latter stage that explicit pronunciation of the second occurrence of 'can' is suppressed.[17]

[15] Since the No Tampering Condition NTC does not permit any manipulation of the structure X, the only possible operation is to *raise* Y from within X; *lowering* Y into X is barred. Thus without stipulation the duality of semantics is determined in the right way: the structurally higher position is not the position where argument structure is determined but instead has to be the operator position, which also conforms, automatically, to the structural notion of 'c-command' determining scope, as necessary for independent reasons—as in standard quantification theory notation. (See also Chomsky, this volume.)

[16] We adopt here and elsewhere the convention of underlining the unpronounced lower copy.

[17] There is some (arguably marginal) evidence from child language studies (Nakamura and Crain, 1987; Ambridge et al., 2008) that there could be a presumptively performance tendency to repeat this second occurrence, so-called 'aux-doubling', a fact lending additional credence to the copy theory. Further, there are interesting cases where some residue of the lower copy is retained in pronunciation, for example, if the copy is in a position where an affix requires it. Even so, the overwhelming phenomenon is deletion of the lower copy, for reasons that are discussed in Berwick and Chomsky (2011): it saves considerable duplicated neural-mental and articulatory computation. It seems to be the case that there is no language that 'pronounces' the full set of copies, e.g., in 'which picture of John did you say Bill told Mary Tom took' the fully spelled-out other copies would amount to (at least) something like, 'which picture of John did you

Systematically running through examples (2)–(9), we can now readily check that in each case the copying account automatically fills in the legitimate locations for v, dv, a, or wh interpretation, meeting our requirements (I) and (II), and most of (III). For example, in (6a), repeated below as (11), 'happy' is interpreted properly in its position after the predicate 'is':

(11) [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.

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'.[18] Given this, all of the criteria (I)–(IV) are satisfied.[19]

These are straightforward examples. But copying can also account for far more complex cases, where, for example, quantificational structure cannot simply be read directly off surface word order, another potentially serious POS problem. For instance, in (12a) below, 'which of his pictures' is understood to be the object of 'likes', analogous to 'one of his pictures' in (13). The copying account renders (12a) as (12b), with the copy 'one of his pictures' in exactly the correct position for interpretation. Further, the quantifier-variable relationship between 'every' and 'his' in (12a) is understood to be the same as that in (13), since the answer to (12a) can be 'his first one' (different for every painter, exactly as it is for one of the interpretations of (13)). No such answer is possible for the structurally very similar (14). Here too the correct structure is supplied by (12b). In contrast, in (14) 'one of his pictures' does not fall within the scope of 'every painter', the right result.

(12a) [which of his pictures] did they persuade the museum that [[every painter] likes best?]

(12b) [which of his pictures] did they persuade the museum that [[every painter] likes [which of his pictures] best?]

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

say [which picture of John] Bill told Mary [which picture of John] Tom took [which picture of John]'. (There are some claims about Afrikaans which assert that this particular language violates this principle, but we put these to one side here.) In fact, in examples like these, sometimes called 'successive-cyclic movement', the position of the unpronounced copy is often marked by some device—morphology, special agreement, or in a case discussed by Torrego in Spanish (1984), V-raising. V-raising meets the standard conditions as outlined in the main text, as expected.

[18] For interesting data and arguments showing the crucial importance of 'islands' and why these cannot be extracted from statistical data see also the chapter by Kam and Fodor in this volume.

[19] We leave open the possibility that there might be some language-independent principles related to island constraints, as discussed in Berwick and Weinberg (1984).

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

A wide range of similar cases involving such 'reconstruction effects' are readily accommodated by the copying account, all within this very restricted UG.

2.4 Other Explanatory Attempts

Since the first expository examples were formulated, there have been attempts to formulate alternative partitionings of factors (1)–(4), distinct from the account given in section 2.3. In this section we review three of the most recent such approaches in light of our criteria listed in section 2.2. In general, while these recent alternatives also strive to reduce the linguistic domain-specific factor (1), the right methodological goal, we shall see that they all fail. For one thing, they leave the principle of the structure dependence of linguistic rules untouched. Further, some aim only to generate the correct polar interrogative sentence strings, rather than addressing the only real question of linguistic interest, which is generating the correct structures for interpretation along with correct pairings, as we emphasized in section 2.1. Those that do aim to get the right pairings, sometimes implicitly, still fail to do so, as we shall show. Finally, in general they do not address the broader cross-linguistic and empirical examples and cannot generate the attested broader patterns of correct and incorrect pairings.

2.4.1 *Clark and Eyraud (Clark and Eyraud 2007; Clark, Eyraud, and Habrard 2008; Clark 2010); hereafter, CE*

We begin by considering a string-based approach that was motivated by considering some of Zellig Harris's proposals on 'discovery procedures' for grammars. CE advance an inference algorithm for grammars that, given positive examples such as (15a) and (15b) below, generalizes to a much larger derivable set of sentences that includes examples such as (15c), while correctly excluding ungrammatical examples such as (15d).

(15a) men are happy.

(15b) are men happy?

(15c) are men who are tall happy?

(15d)* are men who tall are happy?

Briefly, the method weakens 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. E.g., given 'the man died' and 'the man who is hungry died', we can conclude that the strings 'the

978–0–19–959033–9 02-Piatell-c02-drv Piatelli (Typeset by SPi) 32 of 309 June 22, 2012 11:6

man' and 'the man who is hungry' are substitutable for one other in these sentences, and therefore are substitutable in all sentences; similarly, given a new sentence, 'the man is hungry', we may use the congruence of 'the man' and 'the man who is hungry', to substitute for 'the man', yielding '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 of course does not extend existential substitutability to universal substitutability. (The punctuation marks at the end of the example sentences are actually crucial for the operation of the algorithm; see Clark & Eryaud, 2007.) 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 in the same congruence class according to the two simple strings given above. This yields an account of sentence structure, 'how words are grouped into phrases'. It is this extension that does the work in CE's system of generalizing to examples that have never been encountered by a learner—that is, generating novel strings. But it is evident that these notions collapse at once.

CE themselves remark that weak substitutability will 'overgenerate radically' and on 'more realistic samples this algorithm would eventually start to generate even the incorrect forms of polar questions'. That is true, but misleading. The problems do not arise only 'eventually' and with 'more realistic samples', but rather at once and with very simple ones. E.g., from the examples 'eagles eat apples' and 'eagles eat', we conclude that 'eat' is in the same class as 'eat apples', so that substituting 'eat apples' for 'eat' yields the ill-formed string, 'eagles eat apples apples.' Note that 'eat' and 'eat apples' are both verb phrases, but cannot be substituted for each other in 'eagles—apples.' In fact, virtually no two phrases will be substitutable for each other in all texts. Similar elementary examples yield incorrect forms for polar sentences. Thus, from 'can eagles fly' and 'eagles fly' we conclude that 'can eagles' and 'eagles' are in the same congruence class, yielding the polar question 'can can eagles fly'.

To take another example, consider the following simple sequence. It yields an ungrammatical sentence derivation (square brackets are introduced for readability, '≡' denotes 'is weakly substitutable for'):

(16) *does he think* [*well*]?

(17) *does he think* [*hitting is nice*]?; ∴ *well* ≡ *hitting is nice*

Accordingly, given the sentence 'is he well?', we may substitute 'hitting is nice' for 'well' to yield the invalid string 'is he hitting is nice'. In short, it is easy to construct many simple counter-examples like this that violate the weak generative capacity of English. As has long been known, such an approach cannot get off the ground, even with the simplest cases.

As we stressed earlier, the question of interest is generating the right structures for interpretation, along with the proper pairings. The simplest way we know

of—virtually without assumptions—is the one we just sketched (which when spelled out carefully, incorporates the basic assumptions of note 5).[20]

To summarize, CE develop an approach that fails even for the simplest examples and completely avoids the original problem, and of course does not even address the question of why the principles at work generalize broadly, it seems universally. There seems to be no way to remedy the irrelevance of CE's proposal while keeping to anything like their general approach.

2.4.2 *Perfors, Tenenbaum, and Regier (2011), PTR: Bayesian model selection of context-free grammars*

PTR also consider the key question of domain-specific vs domain-general knowledge in language acquisition, but from a different perspective and with a very different way of partitioning factors (1)–(4). We review their approach briefly before turning to its evaluation.

Factor (1), prior, domain-specific linguistic knowledge:
For PTR, this consists of a series of crucial stipulations:

(i) Sentence words are assigned unambiguous parts of speech.

(ii) In particular, PTR represent a sentence such as 'eagles that can fly eat' as the part-of-speech sequence 'n comp aux v vi'; the sentence 'eagles eat' as 'n v'; 'eagles can eat' as 'n aux v'; 'can eagles eat' as 'aux n vi'; and 'eagles are happy' as 'n aux adj'.

Here, the part-of-speech label 'n' denotes any noun; 'comp', the 'complementizer' that introduces embedded S's, typically the word 'that'; and 'adj', any adjective. For PTR's analysis, 'aux' denotes any auxiliary verb (including *can*, *do*, *will*, and the copula in its varied uses; thus appearing twice in 'is the child being obstinate'); 'vi' denotes any verb taken to be uninflected for tense, e.g., 'eat' in (5b); and 'v' any inflected verb, e.g., 'fly' in (5b). Note that 'fly' and 'eat' are actually ambiguous as to whether they are inflected or not, but PTR assume this choice to have been resolved in the required way *before* the analysis proceeds, by some means that they do not discuss. We note that the CHILDES training corpus they use does not in fact typically distinguish 'v' and 'vi'; the novel tag 'vi', which plays a crucial role in the analysis, has been introduced by PTR as a stipulation.

(iii) All the phrases S, NP, VP, IP, etc. required to build a context-free grammar to cover at least the sentences in the training corpus (PTR further assume as given the correct phrase boundaries for the part-of-speech sequences in the training

[20] Clark (2010) extends the distributional approach to include an explicit notion of structure, thereby remediating the issue of addressing only weak generative capacity, but as far as we are able to determine, as of yet there are no results that yield the desired auxiliary verb POS results.

corpus); a (stochastic) context-free grammar that can parse all the sentences of the training corpus; and a finite-state (right-linear, regular) grammar usually derived from the Context Free Grammar (CFG) that can also parse all the sentences of the training corpus.[21] Note in particular that PTR's system does not learn any particular grammar rules; these too are stipulated.

Factor (2), domain-general knowledge:

PTR assume a Bayesian model selection procedure that can choose among the three grammar types, picking the one with the largest posterior probability ('most likely') given the corpus. This probability is in turn the product of two factors, (i) the prior probability of a grammar, $P(G)$, essentially a measure of the grammar's size, with larger grammars being less likely; and (ii) the likelihood of a grammar-corpus pair, which is the conditional probability of generating (parsing) the given corpus given the grammar, $P(\text{corpus}|G)$. The 'best' likelihood $P(\text{corpus}|G)$ is found by attempting to maximize $P(\text{corpus}|G)$, by altering the initial uniform probabilities of the antecedently stipulated CFG or Finite-State Grammar (FSG).[22]

Factor (3), external stimuli:

PTR use a 'training' set of 2336 'sentence types' selected from the CHILDES Adam corpus. As mentioned, actual sentence words are replaced with pre-assigned part-of-speech tags; certain sentences have been removed from the corpus.[23]

Two basic elements are learned by PTR's system: (1) the re-estimated probabilities for the context-free or finite-state rules that (locally) maximize the likelihood of a corpus, grammar pair; and (2) which of the stipulated types of grammar (memorized sentence list, FSG, or CFG) yields the highest posterior probability. In two cases, PTR's method does construct grammar rules *per se*. On approach conducts an automatic "local search" from a given hand-specified context-free grammar (and its corresponding finite-state grammar). The second attempts to carry out a (partially) global, automatic search of the space of possible FSG's, while also calculating the posterior probability of the resulting grammars.

[21] PTR also posit a third 'grammar' type, which consists of simply a memorized list of the sentences (for them, part-of-speech sequences) in the corpus. Some versions of PTR's analyses start from a hand-built context-free grammar (CFG) and then carry out a 'local search' in the space of grammars around this starting point, to see whether this alters their Bayesian selection of CFGs over Finite State Grammars (FSGs). It does not. But we should note as do PTR that there is no mechanical inference procedure provided for constructing CFGs generally; even for FSGs the problem is known to be NP-hard.

[22] PTR include a third factor, the probability of the particular grammar type, T, (i.e, memorized list, finite-state/regular, or context-free), but since these probabilities are all set to be the same, as PTR note, the T value does not alter the relative final posterior probability calculation. The maximization of P(corpus|grammar) is done by a local hill-climbing search method known as the 'inside-outside' algorithm; the details here are not relevant except to note as PTR do that this method is not guaranteed to find a global maximum.

[23] 'The most grammatically complex sentence types are removed…' specifically, (PTR fn. 5), 'Removed types included topicalized sentences (66 individual utterances), sentences containing subordinate phrases (845), sentential complements (1636), conjunctions (634), serial verb constructions (460), and ungrammatical sentences (443)'. For example, PTR exclude the sentence with the subordinate clause, 'are you as tall as Mommy' (Adam02.txt, example 1595).

Specifically, PTR argue that the Bayesian calculus works out so as to rank stochastic context-free grammars with higher posterior probabilities—a 'better fit to the corpus'—than the two other choices which they take to lack hierarchical structure, establishing that this latter property of natural language is learnable without having to posit it a priori.[24]

PTR claim two main results. First, PTR conclude that 'a learner equipped with the capacity to explicitly represent both linear and hierarchical grammars—but without any initial bias to prefer either in the domain of language—can infer that the hierarchical grammar is a better fit'. Second, PTR assert that their 'best' (most probable) context-free grammars exhibit 'mastery' of the auxiliary system: '…we show that the hierarchical grammar favored by the model—unlike the other grammars it considers—masters auxiliary fronting, even when no direct evidence to that effect is available in the input data.' (p. 313).[25]

However, as we show directly, PTR do not establish either of these results, and in particular have not confronted the original POS problem at all.

2.4.2.1 *Learnability of hierarchical structure?* Consider first the question of the learnability of hierarchical structure, given PTR's three choices for covering the given corpus: a memorized finite list of part-of-speech sequences; a (stochastic) context-free grammar; and a regular (finite-state) right-linear grammar derived from the covering context-free grammar.

We may immediately exclude (as they also do) the finite list option as a viable option for any realistic learning model for natural languages. The finite list 'grammar' simply memorizes each particular part-of-speech sequence in the corpus as a special case. Not only is each sentence then totally unrelated to every other sentence—the next sequence could even come from a completely unrelated language, such as German

[24] PTR argue that their three-way choice is a reasonable starting point, though they agree these sorts of grammars are inadequate as models for human language. They also agree that a child does not actually follow this procedure. What is crucial to them is that such completely general statistical Bayesian procedure *can* converge on the hypothesis that the grammar has hierarchical structure. But these three possibilities verge on straw-man possibilities—by their own admission, they are not alternatives a child would actually entertain. The finite memorized set is not even worth considering for elementary memory reasons. Further, as we note in the main text, FSGs do yield hierarchical structure, unless we add an extra assumption of strict associativity. We are then left with two choices, not three, both with hierarchical structure, and, it has been known from the foundational work in formal language theory, finite-state grammars will in general be much larger than CFGs generating the same regular language (see Meyer and Fischer 1969; Berwick 1985). So FSGs are easily eliminated as possible candidates for natural languages, as was familiar from the earliest work in the field. Furthermore, ordinary CFGs are also ruled out, for reasons understood 40 years ago when they were eliminated in favor of X-bar theories of phrase structure (Chomsky, 1970; Jackendoff, 1972). The basic conclusions are carried over to all work we know of making use of phrase structure grammar—which can apparently be eliminated in favor of the most elementary combinatorial operation along the lines discussed above.

[25] 'We argue that phenomena such as children's mastery of auxiliary fronting are not sufficient to require that the innate knowledge constraining generalization in language acquisition be language-specific. Rather it could be based on more general-purpose systems of representation and inductive biases that favor the construction of simpler representations over more complex ones.' (p. 311)

or Chinese—but storage quickly grows far beyond any conceivable memory capacity (and in the limit, is of course impossible). Furthermore, there is no way to address either of the basic conditions (I) or (II) above.

That leaves only the context-free and regular grammars as real candidates. Assuming language to be infinite, as do PTR, then there must be some operation that eventually applies to its own output, that is, *recursively*, or some logical equivalent like a Fregean ancestral. The sequence of applications of this operation always fixes some hierarchical structure (one notion of strong generation), which is not to be confused with the (weakly generated) *string* that is produced. E.g., assuming f to be a successor operation, applied to a single element a, we obtain the structured object $f(\ldots(f(f(a)))\ldots)$ along with the weakly generated string a^n. The operation applies to the output of a previous application of that same operation, and so on. Note that hierarchical structure will always be produced when generating infinite languages, even in the finite-state case, though we can add an extra operation that removes it, such as right associativity in the previous example. Similarly, for CFGs, an operation to remove structure can be added, leaving the (weakly generated) 'terminal string'. Thus *both* of PTR's remaining options generate hierarchical structure, and so there is actually no choice as to whether language is to be represented with hierarchical structure or not. The only question that PTR actually address is whether context-free grammars are to be preferred to finite-state grammars—both inducing hierarchical structure, both long known to be inadequate as descriptions for natural language, as PTR themselves note (p. 329)—while excluding by stipulation the simpler and apparently much more adequate systems described above in section 2.3.

PTR assume that if a grammar produces hierarchical structures, then rules must be structure-dependent. But this is an error. Thus given the structure (1a) (= [can [eagles that v* fly] v eat]]), we are free to interpret 'can' in the position v with a structure-dependent rule, or to interpret it in the position v* with a structure-independent rule. That was the original problem. It makes no difference whether structure is innately determined or learned, or if the latter, how it is learned. In all of these cases, the original POS problem remains unaffected.[26]

The confusion between hierarchical structure and structure dependence of rules appears throughout their paper. Thus they state, 'Henceforth, when we say that 'language has hierarchical phrase structure' we mean, more precisely, that the rules of syntax are defined over hierarchical phrase-structure representations rather than a

[26] An additional error is PTR's conflation of hierarchical structure with analysis into phrases. Thus suppose we have the following (non-hierarchical) sequence of phrases: [$_S$ [$_{AUX}$ is] [$_{NP}$the eagle] [$_{PP}$ in the air]]]. A *structure-dependent* rule can refer to the phrase-names S, Aux, NP, etc., remaining blind to the particular word tokens 'is', 'the', etc. and can front 'the eagle'. A *structure-independent* rule would ignore the brackets. Nothing in this presumes that the bracketing *must* be hierarchical, though of course it *may* be (and in fact generally is). The essential point is that grammatical operations make reference to *phrases*, rather than individual words; the 'hierarchical' addition is just that, PTR's own addition.

flat linear sequence of words. Is the knowledge that language is organized in this way innate?' (p. 7–8).

But having hierarchical phrase structure does not entail that rules are defined over these structures. Rather, the question remains open. That was exactly the point of the original POS problem, which was originally posed on the assumption that structure *is* hierarchic.

Elsewhere they ask: 'is it [that the rules of syntax are defined over hierarchical phrase-structure representations] a part of the initial state of the language acquisition system and thus a necessary feature of any possible hypothesis that the learner will consider?' (p. 309) They do not address this question, contrary to what they assert. Rather, they consider an entirely different question: is hierarchical structure innate or acquired? They claim to show that it is acquired, but they do not address this question either; rather, they beg the question by considering only a choice between two systems, both hierarchic (putting aside the inconceivable list option). And again, the answer to the question they beg leaves the POS problem unchanged. PTR do not address the original POS question regarding the learnability of the *structure dependence* of grammatical rules, as published in all the accounts regarding this topic (Chomsky 1968, 1971, 1975, 1980).

PTR go on to say that 'This question [learnability of hierarchical structure] has been the target of stimulus poverty arguments in the context of a number of different syntactic phenomena, but perhaps most famously auxiliary-fronted interrogatives in English' (p. 309). However, this is incorrect. The question has always been whether *rules* are structure-dependent, not whether language is hierarchical; the POS question remains as before, the choice between *using* structure or ignoring it when hypothesizing rules, regardless of whether children have to learn that language is hierarchical or not.[27]

[27] PTR base their misconstrual on a single sentence from informal discussion in an international conference: 'We quote at some length from one of Chomsky's most accessible statements of this argument, in his debates with Piaget about the origins of knowledge' (Piattelli-Palmarini 1980). It is rather odd to take a sentence from informal discussion (not incidentally with Piaget) when so much is available in the very same conference proceedings, and in print elsewhere, directly refuting their misinterpretation. But even the passage they quote is clear in context. It refers to the suggestion that, if there is hierarchical structure, that would somehow solve the POS problem. It wouldn't, because while the child acquiring language *can* use the structure, giving the right result, the 'left-most' property is of course just as readily available as an induction base.

Furthermore, and more significantly, a few pages later (p. 124), Chomsky points out that the examples that were discussed (and that PTR rely on) 'are misleading in one important respect', namely, they are presented as if they are a list of properties of UG, but the important point is that "this list of properties forms a highly integrated theory … [They] flow from a kind of common concept, an integrated theory of what the system is like. This seems to me exactly what we should hope to discover: that there is in the general initial cognitive state a subsystem (that we are calling [UG] for language) which has a specific integrated character and which in effect is the genetic program for a specific organ … It is evidently not possible now to spell it out in terms of nucleotides, although I don't see why someone couldn't do it, in principle'. The structure-dependent hypothesis discussed is one fragment of that integrated theory, which, we have suggested, can be reduced to much simpler terms that apply much more generally, and in crucial respects may be language- or even organism-independent.

As we have seen, PTR's proposals are irrelevant to the original question of structure-dependence, and are also irrelevant to the new question they raise of learnability of hierarchical structure. Furthermore, as already discussed, the conclusion that language is hierarchically structured follows virtually without assumptions, so the question they pose (and beg) does not arise.

In short, PTR do not deal with the only problem that had been posed: how the proper pairings are acquired by the child, in accord with the universal patterning of data as described in examples (5)–(8). PTR do not even answer the question as to why we should expect the acquired grammar to be a CFG in the face of overwhelming evidence that CFGs make far too many unwarranted stipulations; for example, there is no reason to choose the rule VP → V NP rather than VP → N PP. These are among the considerations that led to X-bar theory forty years ago.[28] The merge-based system described in section 2.3 is simpler—requires fewer factor (1) language-specific stipulations—than PTR's 'best' CFG with its hundreds of rules. It also yields the required pairings straightforwardly, and appears to deal appropriately with the cross-linguistic examples and constraints that PTR's stipulations do not even address.

2.4.2.2 *Reali and Christiansen (2005); RC: learning from bigrams and trigrams*[29]
Besides PTR's Bayesian method, others have offered statistically-based proposals for solving the POS problem for yes-no questions. We consider just one representative example here, a recent model by Reali and Christiansen (2005), hereafter RC. As summarized in a critique of this model, 'knowledge of which auxiliary to front is acquirable through frequency statistics over pairs of adjacent words (bigrams) in training corpus sentences.' (Kam, Stoyneshka, Tornyova, Fodor, and Sakas 2008: 722).

RC's method is straightforward. Like PTR, RC use a corpus of child-directed speech from CHILDES as their test input data, but in this case, actual words, not just parts of speech; in this sense their approach is less stipulative than PTR's. This becomes the training data to calculate the frequency of word pairs. Given this, one can then calculate an overall sentence likelihood, even for previously unseen sequence of word pairs.[30] This sentence likelihood is then used to select between opposing 'test sentence

[28] It can be shown that, while PTR's system using tree substitution can correctly match a few of the correct and incorrect patterns for auxiliary-verb inversion, it fails on many others, both in terms of weak generative capacity as well as in terms of assigned parse trees, because the CFG rules sometimes interact together to yield the wrong results.

[29] See also the critique offered by Kam and Fodor, this volume and Berwick, Pietroski, Yankama, and Chomsky 'Poverty of the Stimulus Revisited' (2011). This recent work and Kam's previous work (2007, 2009) show that RC's kind of statistical analysis can be extended to trigrams (sequences of three words) without success. Serious additional problems arise with the recurrent neural networks proposed by RC.

[30] RC used *cross-entropy* as this likelihood measure, roughly like taking the product of each word pair bigram, but in a log-transformed space (and so turning the product of probabilities into a sum). If we denote by $P(w_i|w_{i-1})$ the conditional probability of the word sequence $w_{i-1}w_i$, then the cross-entropy of a sentence N words long is $-1/N \log_2 \Sigma_{i=2}{}^{N} (w_i|w_{i-1})$ As Kam et al. (2008: 784, and Kam and Fodor this volume) note, 'cross-entropies and probabilities are intertranslatable (they are inversely proportional).' Further, if bigram count for a particular pair is 0, then RC use a smoothing method based on unigram likelihoods (the frequencies for the words considered by themselves).

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 will have a greater likelihood than those with incorrect auxiliary fronting.

RC's (2005) Experiment 1 demonstrates that on 100 test pairs, so-called polar interrogatives with subject relative clauses (PIRCs), the bigram method successfully chooses the correct form 96 percent of the time (as restated by Kam et al., 2008: 773, Table 1). RC go on to demonstrate that simple recurrent neural networks, SRNs (Lewis and Elman 2001), can be trained on the same data, replicating this performance.

RC also extended their bigram approach to a *trigram* model. That is, they calculated sentence likelihoods according to the frequencies of three-word sequences, rather than just two-word sequences. Kam (2007, 2009; Kam and Fodor this volume) also addressed this question, and confirmed that trigram performance essentially tracked that of bigrams, and succeeded (or failed) for the same reason that bigrams did. We can also confirm (see Table 2.1 below), that the homophony issue (between pronouns and complementizers in English) that Kam et al. discovered was the real source of the apparent success of RC's bigrams to discriminate grammatical from ungrammatical auxiliary inverted forms arises as well in the trigrams case, so moving to trigrams or beyond will not help. Berwick et al. (2011) further argues that a neural network is essentially an emulation of the bigram analysis, and so also fails.

To test RC's trigram approach, we used child-directed utterances by adults 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.

We first first checked to see whether we could replicate the Reali and Christiansen (2005) results using trigrams. Using their trigram estimation method, we again tested whether trigram statistics could successfully discriminate between the same 100 grammatical and ungrammatical test sentence pairs used by RC. The results are given in row 1 of Table 2.1: 95 percent correct, 5 percent incorrect, and 0 undecided. This is quite close, but not exactly the same as Reali and Christiansen's (2005) results. Reali

Table 2.1. Percentage of test sentences classified correctly vs incorrectly as grammatical or undecided, using RC's trigram analysis and Kam et al.'s

Experiments	Sentences Tested	% Correct	% Incorrect	% Undecided
1. Replication of Reali and Christiansen (2005), trigram test	100	95	5	0
2. Disambiguated *rel*-pronouns, trigram test, using Reali and Christiansen (2005) test sentences	100	74	11	15

and Christiansen found that the trigram analysis gave *exactly* the same results as the bigram analysis, incorrectly judging the same four ungrammatical sentences as more likely than their grammatical counterparts. Our replication made a single additional mistake, classifying *is the box that there is open* as more likely than *is the box that is there open.* In this last case, aside from one exception, all trigrams in both the grammatical and ungrammatical sentences are 0, so the trigram values are actually estimated from bigram and unigram values. The sole exception is what makes the difference: the ungrammatical form contains a trigram with frequency 1, 'is open end-of-sentence', while the corresponding grammatical form has displaced 'is' to the front of the sentence, so this trigram does not occur. This single 'winning trigram' in the ungrammatical form is enough to make the ungrammatical form more likely under the trigram analysis than its grammatical counterpart.

We then applied Kam et al.'s methodology regarding the effect of the homographic forms for *who* and *that* to the trigram case. We replaced each occurrence of *who* and *that* in the Reali and Christiansen test sentence data where these words introduced relative clauses with the new forms *who-rel* and *that-rel*. The training data remained as before, the Bernstein corpus used by Reali and Christiansen. We then applied the trigram calculation to classify the revised test sentences, with the results shown in row 2 of Table 2.1: 74 percent correct, 11 percent incorrect, 15 percent undecided. This is some improvement over the bigram results, of about 5 percent, but is still well below the close to perfect results found when 'winning bigrams' like *who-is* or *that-is* are not excluded. Thus, while trigrams boost performance slightly, the high accuracy for both bigrams and trigrams in discriminating between grammatical and ungrammatical sentences seems to be due to exactly the effect Kam et al. found: the accidental homophony between pronouns and complementizers in English, rather than anything to do with the yes-no question construction itself.

In short, moving to a trigram model does not help solve this POS problem; indeed, this problem too has been restricted to weak generative capacity, and has omitted the central issue of valid pairings.

Ultimately, the flaw here, parallel to that of CE, is that what the child (and adult) comes to know, as we have seen in our discussion is indeed based on the structure dependence of rules, whether acquired or innate, and this knowledge cannot be replicated by simply examining string sequence frequencies.

More broadly, the bigram analysis makes no attempt to construct pairings. The bigram analysis takes any examples such as (1a,b) as a string of word pairs, with the declarative pairs unrelated to the corresponding interrogatives, thus avoiding the central issue of a semantic connection between a declarative sentence and its corresponding interrogative.

Further, the bigram analysis does not cover the desired range of cases in (6)–(15). Finally, the RC bigram analysis is not the simplest possible, since it demands a factor (2) probability calculation that does not otherwise seem to be required (and for longer

sentences becomes increasingly difficult). To be sure, it has been argued elsewhere (Saffran, Aslin, and Newport 1996; Hauser, Aslin, and Newport 2001) that such a facility might be available as part of some more general cognitive competence, even in other animals (though it has yet to be demonstrated that such high-precision numerical calculations are readily available). But as we have seen, there is a simpler alternative that gets the correct answers yet does not invoke any such likelihood calculation at all.

2.5 Conclusion: What POS Questions Remain?

Much progress has been made in the past half century in reducing the richness and complexity of the postulated innate language-specific properties, thus overcoming certain POS problems and laying a sounder basis for addressing further questions that arise within the biolinguistic framework: questions of acquisition/development, evolution, brain–language relations, and the like. Examples since the 1960s include the elimination of phrase structure grammar (PSG) with all of its complexity and stipulations, progressive simplification of transformational rules and finally their reduction to the same primitive and arguably minimal operation that yields the core of PSG properties, and much else. Needless to say, a great deal remains unexplained. And even if reduced, POS problems always remain, as apparent factor (1) elements are accounted for in terms of general principles—in the best case, natural law.

A good illustration is the example we have been considering: V-raising. As we discussed, there is a natural account in terms of minimal search, possibly a principle of computational efficiency that falls within laws of nature: namely, a clause-initial element C(omplementizer) that determines the category of the expression (declarative, interrogative, and the like) attracts the closest verbal element, where 'distance' is measured *structurally*, not linearly, the only possibility in a Merge-based system in which linear order is part of the mapping to the sensorimotor interface, hence does not enter into core syntactic/semantic computations. This proposal is based on a number of assumptions, some reasonable on empirical and conceptual grounds, but some of them illegitimate within the minimal Merge-based framework, hence stipulative, facts that have not hitherto been recognized. Note that any such stipulation amounts to a real POS problem, since it is a particular fact about language that must be antecedently available. As mentioned at the outset, we would like to eliminate such stipulations if possible, eliminating the POS problem that arises. While we cannot pursue the matter here in any depth, the general situation may be easily stated.

Abstracting from many important details, consider the syntactic object (18), exhibiting the basic structure of (1b), 'eagles that fly can eat':

(18) [C [$_{\text{AuxP}}$ Subject [$_{\text{AuxP}}$Aux VP]]

Here the subject, called the 'specifier of AuxP' (SPEC-AuxP), is 'eagles that fly,' the inner Aux-phrase is 'can VP,' and the VP is 'eat.' Aux is the head of the AuxP, carrying

all information about it that is relevant for further computation (its *label*). C searches for the closest label, and finds Aux, which it raises to C. This is essentially the analysis in traditional grammar, though the terminology and framework are quite different. Further, it appears to capture the basic facts in the simplest way. But it is illegitimate on our assumptions, since the very notion of a Specifier position (SPEC) is an illegitimate borrowing from phrase structure grammar which has long been abandoned, for good reasons (see Chomsky 2012–forthcoming, and in this volume).

On any account, putting the particular terminology to one side, the Subject is merged with AuxP to form a subject-predicate construction. In our terms, oversimplified here for expository reasons, the operation yields {subject, AuxP}. The Subject also has a label, say N (for simplicity): 'eagles that fly' is an NP. But the minimal search procedure that seeks the closest label in (1b) runs into an ambiguity: should it select N or Aux? The problem does not arise in the structure (24); here Aux is the closest label, by stipulation. But we have no basis for the stipulation that the Subject is the Specifier of the auxiliary phrase, SPEC-AuxP, rather than, say, AuxP being the Specifier of the Subject, SPEC-subject. The stipulation illustrated in (24) is therefore illegitimate. The learner must somehow know the right choice to pursue; this is a real POS question that one should try to eliminate, deriving the correct choice from more general principles that are already required.

Note that the same problem arises in any structure of the form A = {XP, YP} where neither XP nor YP is a lexical item (a head). A proposed solution should address the basic Aux/V-raising problems; and reciprocally, provide evidence for the assumptions on which they are based. They must be integrated within broader principles that hold for structures of the form A more generally. There is much to say about this topic, but this is not the place. These considerations about the special case of Aux/V-raising do, however, suggest ways to address a variety of problems that have resisted principled, non-stipulated analysis, while also opening new and quite intriguing questions. That is exactly the kind of outcome we should look forward to when principled solutions are sought for POS problems.

We cannot proceed with the matter here, but it is worth observing that the initial question proposed as a didactically simple illustration of the general POS problem has in this way opened up many productive lines of inquiry when addressed in the manner that is typical of questions of biology and linguistic science generally, as we discussed briefly at the outset.