Cognitive Science 32 (2008) 771–787

ISSN: 0364-0213 print / 1551-6709 online
DOI: 10.1080/03640210802067053

Bigrams and the Richness of the Stimulus

Xuân-Nga Cao Kam[a], Iglika Stoyneshka[a], Lidiya Tornyova[a], Janet D. Fodor[a], William G. Sakas[a,b]

[a]*PhD Program in Linguistics, The Graduate Center, City University of New York*
[b]*Department of Computer Science, Hunter College; and PhD Program in Computer Science and Linguistics, The Graduate Center, City University of New York*

Received 19 May 2006; received in revised form 3 September 2007; accepted 3 September 2007

Abstract

Recent challenges to Chomsky's *poverty of the stimulus* thesis for language acquisition suggest that children's primary data may carry "indirect evidence" about linguistic constructions despite containing no instances of them. Indirect evidence is claimed to suffice for grammar acquisition, without need for innate knowledge. This article reports experiments based on those of Reali and Christiansen (2005), who demonstrated that a simple bigram language model can induce the correct form of auxiliary inversion in certain complex questions. This article investigates the nature of the indirect evidence that supports this learning, and assesses how reliably it is available. Results confirm the original finding for one specific sentence type but show that the model's success is highly circumscribed. It performs poorly on inversion in related constructions in English and Dutch. Because other, more powerful statistical models have so far been shown to succeed only on the same limited subset of cases as the bigram model, it remains to be seen whether stimulus richness can be substantiated more generally.

Keywords: Poverty/richness of the stimulus; Language acquisition; Bigram model; Cross-entropy; Indirect evidence; Auxiliary inversion; Relative clauses; Interrogatives; Dutch; Do-support; Universal grammar

1. Introduction

There has been renewed interest in poverty of stimulus (POS) arguments for language acquisition (Ritter, 2002). Chomsky (1980, *et seq*) argued for the necessity of innate linguistic knowledge (*universal grammar* [UG]) on grounds that children master some language facts before being exposed to relevant exemplars. We focus here on evaluation of some recent findings that appear to undermine this form of argument for the POS.

Correspondence should be sent to Xuân-Nga Cao Kam, PhD Program in Linguistics, The Graduate Center, City University of New York, 365 Fifth Ave., New York, NY 10016. E-mail: xkam@gc.cuny.edu

A novel approach to evaluating—and rejecting—POS claims consists in showing that the language facts in question can be acquired from a corpus of child-directed speech by a very simple statistical learning algorithm with no access to prior knowledge of language structure.[1] Infants have been shown to be sensitive to statistical regularities in their input (Saffran, Aslin, & Newport, 1996), so if a simple statistical learner can pick up linguistic patterns appropriately without aid of UG, it would be implausible to maintain that children cannot.

Reali and Christiansen (2003, 2005), building on work by Lewis and Elman (2001), developed this approach in a study of the acquisition of auxiliary fronting in complex English sentences. They have shown that knowledge of which auxiliary to front is acquirable through frequency statistics over pairs of adjacent words (*bigrams*) in training corpus sentences. Their bigram learning model, trained on child-directed speech to 1-year-olds, was able to select grammatical sentences such as (1) over ungrammatical versions such as (2), although the corpus contained no sentences at all with this structure:

(1) Is the lady who is there eating?
(2) *Is the lady who there is eating?

We refer to the grammatical sentence type illustrated in (1) as a PIRC construction (a polar interrogative with a relative clause modifying its subject). Children's ability to discriminate between the correct and incorrect forms of PIRCs has become the classic example of stimulus poverty, cited often by POS adherents as a compelling illustration of knowledge in the absence of experience. The PIRC construction is, therefore, an important and relevant target for those like Reali and Christiansen who believe in the "richness of the stimulus."[2]

2. Bigram-based learnability of PIRCs

Reali and Christiansen (2005; henceforth R&C) tested a bigram model, a trigram model, and a neural network model. We focus here on the former because if a bigram-based model succeeds in acquiring PIRCs, it can be expected that the more powerful trigram and network models will do so too (but, see the discussion in section 6). In R&C's Experiment 1, a bigram model was trained on a corpus of 10,705 English utterances of adult speech directed to 13 through 21-month-old children, from a corpus of spontaneous adult–child conversations recorded and transcribed by Bernstein-Ratner (1984; available in the CHILDES database—see MacWhinney, 2000). More important, there were no instances of PIRCs in the corpus. Therefore, any information about PIRCs obtained by the model must have been derived from what R&C termed "indirect evidence" supplied by other sentence types.

The model's mastery of PIRCs was assessed on 100 pairs of test sentences similar to (1) and (2) above. Each pair consisted of a grammatical and a matched ungrammatical version, fitting the templates in (3):

(3) Grammatical Is NP {who|that} is A B?
Ungrammatical Is NP {who|that} A is B?,

where A and B are instantiated by VERB PHRASE, PARTICIPLE, NOUN PHRASE, PREPOSITIONAL PHRASE, ADJECTIVE PHRASE

None of the test sentences occurred in the training corpus. Not all the bigrams in the test sentences were in the corpus, but every unigram (word) in them was. The purpose of the experiment was to determine whether, in a forced-choice situation, the model could predict which sentence of a test pair was grammatical by projecting local (bigram) regularities from the corpus.

For each test sentence, R&C computed an estimated probability for each of its bigrams, from which they computed the *cross-entropy* of the sentence.[3] Cross-entropy can serve as a measure of the likelihood of the sentence occurring in the language domain from which the corpus is drawn. In R&C's grammaticality discrimination task, assuming that the more likely a sentence is to occur, the more likely it is to be grammatical, the test sentence version with the lower cross-entropy was chosen as the grammatical one.

One aspect of the method for estimating bigram probabilities is central to our discussion. A bigram consists of two adjacent unigrams. For a bigram that occurs in the corpus, the probability is standardly estimated by dividing the number of occurrences of the word pair in the corpus by the number of occurrences of its first word (= *maximum likelihood estimate*). For a bigram not occurring in the corpus, a variety of alternatives have been proposed for estimating probability (see Chen & Goodman, 1999). R&C employed an *interpolation smoothing technique,* which averages in the estimated probability of the second unigram. We will return to the consequences of this below. In what follows, when we use the term "bigram probability," we will mean the *smoothed* bigram probability, as defined by R&C.

The percentage of test sentence pairs for which the model selected the grammatical version is shown in Table 1. The model's performance was close to perfect: 96% of test pairs were correctly predicted. Based on this strong result, R&C concluded that "there is sufficiently rich statistical information available in child-directed speech for differentiating between correct and incorrect auxiliary (AUX) questions—even in the absence of any such constructions in the corpus" (p. 1008). In our study, we set ourselves the task of identifying *how* the model achieved this impressive level of performance. What information, precisely, was it picking up? Gaining some insight into this is of interest to both linguistics and learning theory, in view of the fact that there is no obvious transparent relation between word co-occurrences in the corpus and the abstract structural properties that govern correct application of auxiliary inversion. Indeed, even if exemplars of the correct PIRC form *had* been provided, it is not entirely clear how a bigram-based learner would extract relevant information from them. As Chomsky (1971) emphasized, auxiliary inversion is structure-dependent; it is sensitive to the hierarchical arrangement of phrases, not to linear sequences of words.

Table 1
Reali and Christiansen's (2005) Experiment 1: Number of sentences classified by the bigram model correctly or incorrectly as grammatical, or undecided

% Correct	% Incorrect	% Undecided
96	4	0

Table 2
Percentage of sentences classified by the bigram model correctly or incorrectly as grammatical or undecided in our six experiments

Experiments	Sentences Tested	% Correct	% Incorrect	% Undecided
1. Replication of Reali and Christiansen (2005)	100	87	13	0
2. Disambiguated *rel-* pronouns	100	18	36	46
3. Homography with determiner	100	18	37	45
4. Object-gap relative clause	100	35	15	50
5. *Do*-support	100	49	51	0
6. Verb inversion in Dutch	40	32.5	55	12.5

3. Understanding the bigram model's success

Before conducting new experiments, we replicated R&C's experiment to be sure that our training corpus, test sentences, and procedures were in accord with theirs. We then examined the specific bigrams that constituted each test sentence, to establish exactly what statistical information the model had access to and how it made use of it.

3.1. Experiment 1: Replication of R&C's result

We extracted 9,643 child-directed utterances by adults from the Bernstein-Ratner (1984) corpus. We manually constructed 100 pairs of test sentences conforming to R&C's templates in (3).[4] Employing R&C's formulae (refer to Footnote 3 above), we computed smoothed bigram probabilities and cross-entropies for all test sentences. The results, presented in the first row of Table 2, corroborate the success of the bigram model. (The "undecided" category in the last column of the table reflects cases where the two versions of a sentence were equal in cross-entropy. Rows 2 through 6 of the table refer to other experiments discussed below.) Clearly, the training corpus did provide sufficient information, expressible in the form of bigram counts, to permit the model to perform the sentence discrimination task to a high level of accuracy.

3.2. Which bigrams favor the grammatical sentences?

As a means of assessing the type of information the model had extracted from the corpus and put to work in judging the PIRCs, we looked at which bigrams in the test sentences contributed to the model's successful selection of the grammatical version.

In these materials, discrimination between the grammatical and ungrammatical sentences necessarily relies on just six distinguishing bigrams. All other bigrams appear in both sentences of the pair, so they cannot be a determining factor in choosing between versions. Consider the pair (4) and (5):

(4) Is the little boy who is crying hurt?
(5) *Is the little boy who crying is hurt?

Table 3
Distinguishing bigrams for the test sentence pair (4) and (5)

Test Sentences	Bigram1	Bigram2	Bigram3
(4) Grammatical	⟨*who is*⟩	⟨*is crying*⟩	⟨*crying hurt*⟩
(5) Ungrammatical	⟨*who crying*⟩	⟨*crying is*⟩	⟨*is hurt*⟩

The non-distinguishing bigrams in (4) and (5) are as follows: ⟨*is the*⟩, ⟨*the little*⟩, ⟨*little boy*⟩, and ⟨*boy who*⟩. The six distinguishing bigrams are shown in Table 3, where we have numbered them for ease of reference. Note that as a direct consequence of R&C's template in (3), bigram1-grammatical was ⟨*who is*⟩ or ⟨*that is*⟩ in every test pair. As we explain below, this gave the ⟨*who is*⟩ or ⟨*that is*⟩ bigram (hereafter ⟨*who|that is*⟩) the major influence in determining the model's performance on the discrimination task.

Table 4 shows the smoothed bigram probabilities for the sentence pair (4) and (5). In each cell, following R&C's formula, the first term of the sum is 0.5 of the unsmoothed bigram probability, and the second term is 0.5 of the probability of the second unigram. For bigrams not occurring in the corpus, the first term is zero, so the smoothing formula estimates probabilities based *solely* on the probability of the second unigram.

The relationships among these six bigrams are crucial to the model's choice. The templates in (3) entail that bigram2-grammatical and bigram1-ungrammatical always share the same second unigram (here, "crying"). For the pair (4) and (5), neither of these bigrams was in the corpus[5], so their smoothed probabilities are based solely on their second unigrams, which are identical. Hence, when the probabilities of the bigrams in each sentence version are multiplied together to give the sentence probability[6], these two bigrams cancel each other out and play no role in the model's choice. (The shading in this and other tables indicates bigrams that cancel out across the two sentence versions.) This is not an isolated case, but is typical of many of the test sentence pairs because only a low proportion (32%) of the bigrams in the test sentences occur in this (relatively small) corpus.

A comparable relationship holds between bigram3-grammatical and bigram3-ungrammatical. Their smoothing function is the same, so when neither bigram is attested in the corpus, they cancel and contribute nothing to the discrimination task. In (4) and (5), this

Table 4
Smoothed probabilities (× 100,000 for better visualization) for the six distinguishing bigrams in sentences (4) and (5)[a]

Test Sentences	Bigram1	Bigram2	Bigram3
(4) Grammatical	⟨*who is*⟩ $127.66 + 7.18 = 134.84$	⟨*is crying*⟩ $0 + .04 = .04$	⟨*crying hurt*⟩ $0 + .03 = .03$
(5) Ungrammatical	⟨*who crying*⟩ $0 + .04 = .04$	⟨*crying is*⟩ $0 + 7.18 = 7.18$	⟨*is hurt*⟩ $0 + .03 = .03$

[a] See text for explanation of shading.

leaves just the remaining two bigrams, bigram1-grammatical and bigram2-ungrammatical, to determine the model's decision. It is these bigrams that create its strong bias toward the grammatical sentence. In all test pairs, these bigrams share their second unigram, so they too would annul each other if neither bigram were attested. However, in this case, the bigram in the grammatical version is *always* attested. Template (3) entails that bigram1-grammatical is either ⟨*who is*⟩ or ⟨*that is*⟩ in every case. The training corpus contained 12 occurrences of ⟨*who is*⟩ and 23 occurrences of ⟨*that is*⟩.[7] Therefore, bigram1-grammatical is *guaranteed* to have a higher smoothed probability than bigram2-ungrammatical whenever the latter is not in the corpus, thus pushing the model toward correctly selecting the grammatical sentence. Only when bigram2-ungrammatical occurs in the corpus and its probability exceeds that of the ⟨*who*|*that is*⟩ bigram, could the model choose the ungrammatical test sentence.

These details concerning how the six distinguishing bigrams are likely to trade off against each other are important because they drive the outcomes of R&C's experiment, as well as our own. Out of the six distinguishing bigrams, only one (the ⟨*who*|*that is*⟩ bigram) features in *every* test pair; and in every case, this bigram is in the *grammatical* sentence. It thus serves as a "marker" for the grammatical version; and because the probability of ⟨*who*|*that is*⟩ is greater than that of many other bigrams, it often dominates the calculation. Of the 87 positive outcomes in our experiment, 80 (92%) can be traced specifically to the bigram ⟨*who*|*that is*⟩. The exact success rate varies slightly, of course, with the particular sentence pairs employed and the details of the corpus.[8] But in the training corpus in this experiment, which may well be typical in this respect, the various distributions of the other bigrams rarely outweighed the bias created by the constant presence of ⟨*who*|*that is*⟩ in every grammatical test sentence.

3.3. The source of the winning bigram

In addition to logging the incidence of correct sentence choices, our analysis thus established the nature of the corpus data on which those choices were based. A clear conclusion was that the ⟨*who*|*that is*⟩ bigram does the lion's share of the work in predicting the correct PIRC version. Despite the dependence of auxiliary inversion on abstract syntactic structure, it turns out that a simple cue consisting of two linearly adjacent words is all it takes to identify the correct form. In fact, this falls under a general recipe for success in bigram-based learning. A bigram model will have a good chance of performing well in any sentence discrimination task if there is a marker bigram (or more than 1) that (a) appears fairly systematically in the grammatical test sentences and not in the ungrammatical ones, and (b) has a high estimated probability relative to that of other bigrams. When these conditions are met, discrimination will probably succeed; when they are not, success is possible but cannot be counted on. To further evaluate bigram-based learning of PIRCs, therefore, attention must focus on the support provided by the corpus for the crucial ⟨*who*|*that is*⟩ bigram.

Because of the young age of the children to whom the adult sentences were uttered, it seemed surprising that the corpus would contain enough relative clauses to supply a substantial number of ⟨*who*|*that is*⟩ bigrams. A manual search for relative clauses revealed only 19 that contained an overt relative pronoun. None of these contained either ⟨*who is*⟩ or ⟨*that is*⟩. Instead, all 12 ⟨*who is*⟩ bigrams appeared in questions (e.g., "Who is in there?"), and all 23 ⟨*that is*⟩ bigrams had "that" as a deictic pronoun (e.g., "That is a rose."). Thus, the "who" or "that" of

a ⟨*who*|*that is*⟩ bigram was in every case merely a homograph of a relative pronoun.[9] In short, the ability of the model to select grammatical relative clause constructions in the experiment was due to a ⟨*who*|*that is*⟩ bigram that had nothing to do with relative clauses. The prevalence of the crucial bigram for discrimination of PIRCs was due instead to two "accidental" facts of English: that it has words that are homographs of relative pronouns, and that these words commonly occur immediately preceding "is." This raises a real possibility that bigram-based PIRC discrimination would be less successful for varieties of the PIRC construction that lack support from these or similar idiosyncrasies in the language. We conducted two experiments to confirm this, which we report in section 4 before moving on to examine, in section 5, the generality of what the bigram model is able to learn.

4. Without the "wrong" bigrams

In Experiment 2, we investigated the bigram model's performance on a language exactly like English, except that it lacked any homographs for relative pronouns. Although this is not a language in use by any language community, there is no reason to doubt that it is a normal learnable human language. There are many natural languages (e.g., Finnish, Hebrew, Yoruba) in which the relative pronoun does not look or sound like any other word.[10] The procedure was exactly as in Experiment 1, but all relative pronouns in the corpus and test sentences were labeled as either "who-rel" or "that-rel" in order to distinguish them from other occurrences of "who" and "that." The distinguishing bigrams in the test sentences, and their estimated probabilities, were exactly as for Experiment 1, except that "who-rel" or "that-rel" appeared in place of "who" and "that," respectively, as bigram1 in both sentence versions; and the first term of the estimated probability of bigram1-grammatical was now zero in all cases. The results, as expected, were not in favor of the bigram model (see the second row of Table 2). When it had to rely on genuine relative pronouns, the bigram model selected the grammatical version of the PIRC only 18% of the time. This is because, unlike English, this language lacked a simple bigram cue for making the correct choice. Although the ⟨*who-rel*|*that-rel is*⟩ bigram was present in all the grammatical test sentences, it never occurred in the corpus, so bigram1-grammatical no longer delivered a systematic boost to the grammatical version.

In Experiment 3, we examined a language in which relative pronouns were homographic with another item in the language but not helpfully so. We substituted the English determiner "the" for all the relative pronouns in the original corpus and test sentences, in order to check that not just any homography would facilitate selection of the grammatical version. Overlap between determiners and relative pronouns is not uncommon in natural languages; it occurs, for example, in German. But a determiner is not normally followed by an auxiliary verb, so this homography would not raise the estimated probability of the marker bigram that distinguishes grammatical from ungrammatical PIRCs.

The results are shown in the third row of Table 2. They are just as poor as for Experiment 2. This is not surprising. In both of these experiments, the model had no choice but to acquire its knowledge of relative clauses from actual relative clauses in the corpus, and the knowledge obtainable in that fashion was clearly insufficient. These two experiments thus confirm our diagnosis that the successful bigram-based performance on PIRCs in the previous studies

was in a sense a lucky fluke because it rested on information in the corpus that was actually irrelevant to the linguistic construction to be learned.[11] However, the results from Experiments 2 and 3 do not disconfirm the more important hypothesis of interest: that, contrary to the POS thesis, children may be able to acquire complex syntactic constructions on the basis of bigram evidence from other, simpler constructions. So far, these experiments merely present a reason for doubting that young children's linguistic input can, in the general case, provide bigram cues for complex constructions. When and whether those cues exist seems to depend on *ad hoc* facts of the target language, so that learnability is not guaranteed, unlike the more or less universal reliability of child language acquisition. Conceivably, however, a larger or "older" corpus (adult speech to older children) would offer a more adequate supply of genuine relative pronouns in ⟨*who*|*that is*⟩ bigrams, so that PIRC discrimination could succeed even without the crutch of homography. We return to issues about the richness of the corpus in section 6, but see Kam (2007) for data showing only modest improvements in discrimination performance with larger–older corpora, for PIRC varieties with object-gap relative clauses and *do*-support (which we show below can be difficult for bigram-based learning).

However, there is a deeper challenge for bigram-based PIRC learning, which is that there are many other varieties of PIRC (auxiliary inversion in relative clause constructions) that do not contain ⟨*who*|*that is*⟩ or any convenient marker bigram at all. So far, we have followed R&C in limiting the view to PIRCs defined by template (3), with a subject gap in the relative clause and *is* as the auxiliary (or copula) in both clauses. Despite receiving, by far, the most attention in the POS literature, this is merely an arbitrarily chosen instance of a much broader phenomenon. Auxiliary inversion in interrogatives can involve other auxiliaries such as "was," "must," or the "do" of do-support. The main clause and a relative clause may differ in their auxiliaries (e.g., "Must the boy who was crying go home?"), or one or both may have no auxiliary (e.g., "Must the boy who cried go home?"). There are also PIRCs in which the relative pronoun is followed not by an auxiliary but by the subject of the relative clause, as in "Is the boy who the girl is talking to going home?" where the object of the relative clause has been relativized. The relative clause subject can be any well-formed noun phrase, with considerable freedom as to its first word, so the relative pronoun would appear in different bigrams in different examples (e.g., ⟨*who the*⟩, ⟨*who Jim*⟩), diluting the likelihood that the relative pronoun is in a well-attested bigram in the corpus. In sum, in a more representative collection of PIRC test sentences, the bigrams in the grammatical version will be quite varied and might not offer a dependable basis for choosing correct sentences.

The claim that a bigram model can acquire valid linguistic knowledge of auxiliary inversion would be warranted only if its competence extends to this broader range of instances because they differ from template (3) only in details that do not bear on the basic linguistic pattern.[12] In the ideal case, a learning model would identify the broad generalization (i.e., invert the auxiliary in the main clause), which covers all the various instances. It is standardly assumed that human learners attain this generalization (although data on when they do so are sadly lacking in the language development literature). As a lesser achievement, a statistical model could be claimed to have acquired some significant linguistic knowledge of auxiliary inversion if it had identified distinguishing bigram(s) to cue the correct choice for each of the various subvarieties of PIRC, as the ⟨*who*|*that is*⟩ bigram does for *is-is* PIRCs. But the next three experiments that we report indicate that this is not so, at least for corpora comparable to

R&C's. The *is-is* PIRCs defined by (3) are tailor-made for a bigram language model, but it appears that they are exceptional. Results show that other PIRCs do not lend themselves so readily to bigram-based learning.

5. Extending the investigation to a wider range of PIRCs

As noted, question formation by auxiliary inversion occurs in many more contexts than fit the narrow specifications for test sentences in previous experiments. Our results so far indicate a dependence of bigram-based learning on one very specific bigram that is not present in the broader range of examples. Our next experiments investigate whether other bigram information is available to ground successful learning in all cases.

5.1. English PIRCs without a marker bigram

In Experiments 4 and 5, we examined PIRCs with object-gap relative clauses and PIRCs with lexical verbs needing do-support. We anticipated weaker performance on these than for the *is-is* PIRCs with subject-gap relatives tested in the previous experiments because in neither case is there a single characteristic distinguishing bigram in the grammatical sentence. But that is not decisive. It also needs to be established whether for these cases the input affords some confluence of minor cues, each one only a weak predictor but reinforcing each other in favor of the grammatical version. Not knowing in advance what these cues might be, our research strategy was to ascertain whether the learning model would succeed. If it did, an effort could be initiated to identify the cues responsible. If it did not, it could be concluded that such cues are not available.

For both experiments, 100 pairs of test sentences were constructed, and the method was as before. In the object-gap PIRC items, both clauses contained the auxiliary "is," but the relativized noun phrase was the object of the relative clause, as in the pair (6) and (7). For clarity, we have inserted brackets around the relative clause, and have indicated the trace (the underlying position) of the fronted auxiliary as well as the trace of *wh*-movement (the "gap") in the relative clause:

(6) Is$_i$ the wagon$_j$ [Ø$_j$ your sister is pushing t_j] t_i red?
(7) *Is$_i$ the wagon$_j$ [Ø$_j$ your sister t_i pushing t_j] is red?

The relative pronoun was phonologically null (Ø) in all test sentences. An overt relative pronoun ("who" or "that") could have been used, but it would not have been part of any distinguishing bigram, so the outcome would have been identical. As illustrated in (6) and (7), the relative pronoun in an object-gap relative clause is sandwiched between the modified noun and the subject of the relative clause, and these are identical across grammatical and ungrammatical versions. The distinguishing bigrams for (6) and (7) are shown in Table 5 with their estimated probabilities.

The patterning of the six bigrams in Table 5 is similar to the previous experiments with respect to the sameness of smoothing factors across versions, indicated by the shading. But here there is no distinguishing bigram that systematically appears in every pair. As a result,

Table 5
Smoothed probabilities (× 100,000) for the distinguishing bigrams in sentences (6) and (7)

Test Sentences	Bigram1	Bigram2	Bigram3
(6) Grammatical	⟨*sister is*⟩ 0 + 718.41 = 718.41	⟨*is pushing*⟩ 0 + 1.12 = 1.12	⟨*pushing red*⟩ 0 + 16.84 = 16.84
(7) Ungrammatical	⟨*sister pushing*⟩ 0 + 1.12 = 1.12	⟨*pushing is*⟩ 0 + 718.41 = 718.41	⟨*is red*⟩ 0 + 16.84 = 16.84

discrimination was not strongly biased toward either sentence version. The results, in the fifth line of Table 2, showed a mix of correct and incorrect choices, as well as some failures to choose either version for pairs in which all bigrams were absent from the corpus. Only 35% of test pairs were correctly distinguished. It appears, then, that there is not even any combination of cues in these object-gap PIRCs that tips the scales toward the grammatical version.

In the experiment on do-support PIRCs, we reverted to subject-gap relative clauses beginning with "who" or "that," but we used lexical main verbs instead of "is," as illustrated by the test pair (8) and (9). The main verb remained*in situ,* but a form of "do" appeared sentence-initially, and the corresponding main verb took an uninflected non-finite form ("want" and "play" in (8) and (9), respectively)[13]:

(8) Does$_i$ the boy [who plays the drum] t_i want a cookie?
(9) *Does$_i$ the boy [who t_i play the drum] wants a cookie?

These test pairs all have eight distinguishing bigrams, as Table 6 illustrates.

Because of the variation in verb finiteness, only two of the four cross-version bigram comparisons exhibit comparable smoothing factors. Therefore, the versions could never completely cancel out (except by coincidence), so it is predicted that undecided cases will not occur. Corresponding to the ⟨*who*|*that is*⟩ of the earlier experiments, the grammatical sentence in this case contains "who" or "that" followed by the lexical verb of the relative clause (e.g., ⟨*who plays*⟩ in (8)). Such bigrams are not likely to be very common in the corpus (even permitting homography, as we did in this experiment). In fact, none of the bigrams in (8) and (9) is particularly likely to have strong corpus support, so neither the grammatical nor the ungrammatical sentence version is expected to prevail. The results, shown in the fifth row of

Table 6
Smoothed probabilities (× 100,000) for the distinguishing bigrams in sentences (8) and (9)

Test Sentences	Bigram1	Bigram2	Bigram3	Bigram4
(8) Grammatical	⟨*who plays*⟩ 0 + 1.12 = 1.12	⟨*plays the*⟩ 0 + 1452.53 = 1,452.53	⟨*drum want*⟩ 0 + 315.42 = 315.42	⟨*want a*⟩ 355.87 + 1037.2 = 1,393.07
(9) Ungrammatical	⟨*who play*⟩ 0 + 55 = 55	⟨*play the*⟩ 0 + 1452.53 = 1,452.53	⟨*drum wants*⟩ 0 + 25.82 = 25.82	⟨*wants a*⟩ 0 + 1037.2 = 1,037.2

Table 2, comport with these predictions: roughly one half (49%) of the pairs were correctly discriminated. The bigram model was evidently not picking up any combination of cues as potent as the ⟨*who*|*that is*⟩ cue in the subtype of PIRC examples tested by R&C. This supports our conjecture that the positive results in the original experiments do not reflect a general grasp of the linguistic constraints on subject-auxiliary inversion.

5.2. Dutch PIRCs with main verb inversion

For our final experiment, we turned to another language in order to examine a subvariety of PIRC that does not occur in English. In some languages, the inversion process that forms questions applies more generally to all finite verbs, not just auxiliaries. This is the case in Germanic languages, including Dutch, as illustrated in example (10), which follows. The Dutch equivalent of English sentences like (8) thus does not need do-support.[14] Our goal is to determine the extent to which a bigram model is capable of extracting general patterns of sentence formation from a corpus. Testing it on Dutch, with its general pattern of verb inversion, can be informative in this regard. If it is to establish that bigrams provide a general basis for a statistical learnability argument against POS, a bigram model must be able to learn how to form complex questions in *any* language, even if the specific bigrams in those questions differ radically from one language to another. Testing this therefore demands the use of a real corpus of Dutch such as a Dutch-learning child would be exposed to. Thus, in Experiment 6 we did not merely change one controlled property of the English corpus as we did in Experiment 3, but started afresh with a Dutch corpus and Dutch sentences in order to allow any and all properties of the language input to contribute to the model's ability to discriminate grammatical from ungrammatical PIRCs.

The training corpus used in this experiment is known as the Groningen corpus (Bol, 1996), available in the CHILDES database. The children's ages ranged from 1 year, 5 months to 3 years, 7 months; and we chose adult utterances from among the earliest files in the corpus, covering a 4-month period for each child. This yielded 21,557 utterances of child-directed speech. The resulting corpus was thus both larger and somewhat older than our corpus of English child-directed speech, but this would tend to increase the chances of successful learning by the bigram model. Forty pairs of Dutch PIRC sentences, constructed with the assistance of a native speaker, were tested. In all items, both clauses contained a lexical verb only (no auxiliary).[15] Sentences (10) and (11) are typical of the test pairs:

(10) Wil$_i$ de baby [die op de nieuwe stoel zit] t_i een koekje?
Wants the baby that on the new chair sits a cookie?
"Does the baby that is sitting on the new chair want a cookie?"

(11) *Zit$_i$ de baby [die op de nieuwe stoel t_i] wil een koekje?
Sits the baby that on the new chair wants a cookie?
"Is the baby that sitting on the new chair wants a cookie?"

The eight distinguishing bigrams for (10) and (11) are shown in Table 7. (The unigram "-sent" in bigram1 is the sentence-initial marker, which is relevant in this experiment because the initial word differs across versions.)

Table 7
Smoothed probabilities (× 100,000) for the distinguishing bigrams in sentences (10) and (11)

Test Sentences	Bigram1	Bigram2	Bigram3	Bigram4
(10) Grammatical	⟨*-sent- wil*⟩	⟨*wil de*⟩	⟨*stoel zit*⟩	⟨*zit een*⟩
	368.82 + 177.54	135.14 + 896.31	675.68 + 134.35	1,964.29 + 998.99
	= 546.36	= 1,031.45	= 810.03	= 2,963.28
(11) Ungrammatical	⟨*-sent- zit*⟩	⟨*zit de*⟩	⟨*stoel wil*⟩	⟨*wil een*⟩
	173.97 + 134.35	892.86 + 896.31	0 + 177.54	270.27 + 998.99
	= 308.32	= 1,789.17	= 177.54	= 1,269.26

Smoothing factor similarities across the versions have the potential for cancellation of all the bigrams, and hence for some cannot-decide outcomes. In the example in Table 7 there happen to be no cancellations. The grammatical version was selected; but, in other cases, the ungrammatical version was chosen (e.g., the ungrammatical *"Lijkt de kok die moe maakt een cake?" was favored over the grammatical "Maakt de kok die moe lijkt een cake?" ["Is the cook who seems tired making a cake?"]). Overall, the model did not do well on these Dutch PIRCs; see row 6 of Table 2. Only 32.5% of the test pairs were correctly discriminated. Clearly, the Dutch training corpus did not offer any reliable indicators of grammaticality that could be encompassed in bigrams.

6. General discussion

R&C demonstrated that children's primary linguistic data affords information determining the correct form of one complex syntactic construction. As we have seen, this does not imply that the same will be true for every complex syntactic construction; so it does not by itself falsify the POS thesis in general. On the other side, our demonstration that for some syntactic constructions a bigram model does not find definitive information in a corpus of child-directed speech does not entail that other statistical models will equally fail to do so. Thus, the substantive issue of whether the input for syntax acquisition is rich or poor is not settled by the data presented here. Nevertheless, some general conclusions can be drawn from this exercise: conclusions about methodology, about future research needed, and about a possible role of UG.

Our methodological conclusion is that, at least when the goal is to shed light on human language acquisition, it should be standard practice for data-driven learning claims to be accompanied by an elucidation of the source of the knowledge acquired. The POS thesis is a thesis about first language acquisition by children. Its central importance to linguistics and psycholinguistics, and the reason it is still vigorously debated, is that it has strong implications for the *mechanisms* of human language acquisition. For this purpose, a learnability result is more telling if it is clear what precisely has been learned on the basis of what input evidence. This point is especially acute in the case of learning from "indirect" evidence, where the knowledge does not derive from explicit exemplars of the construction in question. In the case of PIRCs, once it was established which bigrams in the original *is-is* test sentences were responsible for the model's bias toward the correct version, it was easy to see that they were

the product of the criteria by which these sentences had been carved out as the particular subclass of PIRCs to be studied. In consequence, the finding of input richness for that class of items provides only the weakest encouragement for the belief that every other construction, when studied, will prove to have a comparably reliable statistical hallmark. Our Experiments 4 through 6 then showed that this is not so, at least under conditions comparable to those in R&C's experiment.

This leads us to our second conclusion, which is that these issues can be settled only by further studies that systematically vary the size and richness of the training corpus, in order to see whether there indeed exists some level of input information that a child might reasonably have access to and which does suffice for acquisition of general linguistic patterns without actually exemplifying them. A richer corpus might provide more varied examples of other related constructions, or it might include part-of-speech information, or even syntactic bracketing, which could be accessible to a learner via prosody or a previous round of learning. Kam (2007) represented a beginning on this large-scale, but essential, project. But what is also needed is an upgrading of the statistical power of the learning model—an escalation from bigrams to trigrams and other richer sources of co-occurrence information. R&C did test a trigram model, and also a much more powerful *simple recurrent network* with part-of-speech tags as input. However, the only results to date, even for these stronger learners, are for the original *is-is* subject-gap variety of PIRC, which we have seen is trivially acquirable by the simplest *n*-gram learner and is not a bellwether for human language acquisition at large. It remains to be seen whether the ability of, for example, a neural network will extend to a fuller range of PIRCs in English and other languages, and even perhaps to other constructions cited in the stimulus poverty versus richness debate.

Our third general conclusion is necessarily more speculative. It concerns the possible role of innate linguistic knowledge (UG), which has been the hostage to fortune in this debate since its inception. The need for UG to assist language acquisition is more or less complementary to the power of UG-free, data-driven learning; and as just observed, the latter will not be known until much more research has been done. But it is instructive to consider what the implications would be if future research, even with more sophisticated statistical mechanisms, were to confirm the mixed pattern of success observed in the experiments reported here, where one variety of a construction succumbed readily to a statistical learning algorithm, whereas others were resistant to it. If this were to emerge as a typical outcome, it could be concluded that the information provided by such data-driven computations might contribute to grammar acquisition by serving as a bootstrapping strategy for learners, but could not substitute for a grammar. If a child were to gather statistical data from a collection of utterances and then continue to rely on it indefinitely to guide his or her use of the language, without formulating a grammar rule, he or she would make egregious errors on the constructions for which the statistics do not point toward the grammatical version. Thus, even learning systems that are well-equipped to track corpus statistics may need to derive rules if the statistics predict the correct form of only some but not all sentence types. On these assumptions, then, learners may *start* by referencing probabilistic dependencies between words, but they would at some point make the transition to general grammatical rules; and it is possible that some sort of innate guidance, very like what linguists mean by UG, may prove essential in effecting that transition.

Notes

1. For a discussion of how this novel form of argument relates to earlier debates about stimulus poverty, see Kam, Stoyneshka, Tornyova, Fodor, and Sakas, 2007.
2. In addition to the work of Reali and Christiansen (2003, 2005), and Lewis and Elman (2001), computational studies of PIRCs that defend the richness of the stimulus include Clark and Eyraud (2006) and Perfors, Tenenbaum, and Regier (2006). For an extensive discussion of PIRCs in relation to stimulus poverty, see articles in Ritter (2002).
3. The formulae employed by Reali and Christiansen (1005), and also in our experiments, are as follows: where $c(x)$ is the count of x in the training corpus, N_s is the number of words (tokens) in the training corpus, N_T is the number of words in a test sentence, and λ is fixed at 0.5:
   Maximum likelihood probability of unigram w_i: $P_{ML}(w_i) = c(w_i)/N_s$
   Maximum likelihood probability of bigram $w_{i-1}w_i$: $P_{ML}(w_i|w_{i-1}) = c(w_{i-1}w_i)/c(w_{i-1})$
   Interpolated (smoothed) probability of bigram $w_{i-1}w_i$: $P_{interp}(w_i|w_{i-1}) = \lambda P_{ML}(w_i|w_{i-1}) + (1-\lambda) P_{ML}(w_i)$
   Cross-entropy of a test sentence s_T: $H(s_T) = -\frac{1}{N_T}\log_2 \prod_{i=2}^{N_T} P_{interp}(w_i|w_{i-1})$
4. Test sentences for all of our experiments are available at http://www.colag.cs.hunter.cuny.edu/pub/Bigrams_Richness_Experiments.zip
5. All statements in this article about which items appeared in the training corpus refer to our own corpus, modeled on Reali and Christiansen's (R&C; 2005) as noted earlier. Although there are minor discrepancies between the two, we believe they are sufficiently slight that our observations hold equally for R&C's experiment. Some corpus facts cited here may seem surprising (e.g., that the unigram "crying" was in the corpus, but the bigram ⟨*is crying*⟩ was not). On checking, we found that "crying" occurred following "you," "she's," and "he's" only. Following R&C's practice (possibly intended to mirror the inability of children of this age to recognize reduced forms), a form with a reduced auxiliary such as "she's" was treated as a single unigram.
6. We evaluated the bigram model on the basis of the cross-entropies of sentences, just as Reali and Christiansen (2005) did. However, our discussion of how individual bigram probabilities contributed to the comparison between grammatical and ungrammatical sentences is easier to follow in terms of the estimated probabilities of the sentences. This is taken to be the product of the probabilities of all the bigrams that compose the sentence. Cross-entropies and probabilities are intertranslatable (they are inversely proportional), so this expository decision has no effect on the facts reported.
7. These counts may seem modest, but comparatively speaking, the estimated probabilities of ⟨*who is*⟩ and ⟨*that is*⟩ were high. Their estimated probabilities were the second highest and seventh highest, respectively, of all 391 distinct distinguishing bigrams in the 100 test pairs.
8. In a subsequent run of this experiment with an arbitrarily different set of test sentences, the correct and incorrect outcomes were 84% and 16%, respectively, with no undecided situations. For runs in which the test sentences were deliberately constructed to favor

either bigram2-grammatical or bigram3-ungrammatical, the success rate rose to 93% and fell to 83%, respectively.

9. For children, who have access to the spoken but not the written forms, it is obviously homophony rather than homography that would be relevant; but to avoid switching back and forth between terms, we refer to homography throughout, even where it is strictly inappropriate as in discussion of child language acquisition. However, this may be more than a terminological issue. Possibly, "who" and "that" as interrogatives or deictics would have been prosodically distinguishable from "who" and "that" as relative pronouns in the original conversations, although not in the transcriptions employed in the experiments.
10. We chose to edit the English corpus in this way rather than shifting to a completely different language to isolate this one factor of homography from all the other syntactic differences (e.g., different word order) between English and another language that could influence the model's performance on PIRCs in uncontrolled ways.
11. 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, and which the learning model was deprived of in Experiments 2 and 3. This is an interesting possibility, but in the absence of child data relating to this topic, only speculation is possible. Experiment 1 suggests that a learner who conflated relative pronouns with superficially similar words in English would do well on PIRCs even without any experience of relative pronouns. However, the bootstrapping strategy has its drawbacks. It would result in a striking pattern of errors on other constructions until such time as the child attains an adult-like ability to distinguish the homographic forms one from another. For example, a "conflating" child might accept ungrammatical sentences with "this" misused as a relative pronoun in place of "that" (e.g., **Hug the boy this is crying*) or with the relative pronoun "who" triggering inversion (e.g., **Hug the boy who is the dog barking at*) because interrogative "who" does so. These are good testable predictions, but they do not resonate with currently familiar data on child language; and we know of no evidence that languages without homography for relative pronouns are difficult to learn.
12. However, some allowance might need to be made for unequal acquisition rates for different PIRC subtypes due to varying degrees of corpus support for their auxiliary verbs, for instance. Ambridge, Rowland, and Pine (2008) made a similar point concerning child acquisition.
13. Linguistic analyses of *do*-support constructions differ with respect to whether "do" is originally present and then moved, or is derivationally inserted to support a moved tense morpheme. We believe that nothing relevant to bigram-based learning hangs on this. Here we have adopted the *do*-movement analysis, as reflected in the traces indicated in (8) and (9).
14. Dutch does have *do*-support, but it applies primarily in the context of verb-phrase preposing. See van Kampen (1997) for linguistic references and a discussion of do-support in the acquisition of Dutch.
15. We would also have tested the Dutch equivalent of *is-is* PIRCs, but it was not feasible to do so because in their written form (without disambiguating prosodic phrasing), they

are structurally ambiguous in such a way that no learning algorithm could distinguish the grammatical and ungrammatical versions. The lexical verb PIRCs used in our experiment were disambiguated by the different argument structures of the verbs in the two clauses.

Acknowledgments

We are indebted to Marcel den Dikken for his advice on several aspects of this research and his invaluable assistance in creating the Dutch materials for Experiment 6. We are also grateful to Florencia Reali, Morten Christiansen, and to the anonymous reviewers for *Cognitive Science* for their comments and suggestions. We also received helpful feedback from audiences at the 18th annual City University of New York (CUNY) Conference on Human Sentence Processing, March 2005; and the Workshop on Psychocomputational Models of Human Language Acquisition at the meeting of the Association for Computational Linguistics, June 2005. This work began as a student research project by the first three authors, under the supervision of the last two authors. This research was supported, in part, by Grants 65398-00-34, 66443-00-35 and 66680-00-35 from the Professional Staff Congress of CUNY.

References

Ambridge, B., Rowland, C. F., & Pine, J. M. (2008). Is structure dependence an innate constraint? New experimental evidence from children's complex-question production. *Cognitive Science, 32,* 222–255.

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

Bol, G. W. (1996). Optional subjects in Dutch child language. In C. Koster & F. Wijnen (Eds.), *Proceedings of the Groningen Assembly on Language Acquisition* (pp. 125–135).

Chen, S. F., & Goodman, J. (1999). An empirical study of smoothing techniques for language modeling. *Computer Speech and Language, 13,* 359–394.

Chomsky, N. (1971). *Problems of knowledge and freedom*. New York: Pantheon Books.

Clark, A., & Eyraud, R. (2006). *Learning auxiliary fronting with grammatical inference*. Paper presented at the 10th conference on Computational Natural Language Learning, New York.

Kam, X.-N. C. (2007). Statistical induction in the acquisition of auxiliary-inversion. In *Proceedings of the 31st annual Boston University conference on language development* (pp. 345–357). Somerville, MA: Cascadilla.

Kam, X.-N. C., Stoyneshka, I., Tornyova, L., Fodor, J. D., & Sakas, W. G. (2007). *Bigram-based learning and the richness of the stimulus for language acquisition*. In LIBA (Linguistics in the Big Apple: CUNY/NYU Working Papers in Linguistics) at http://web.gc.cuny.edu/dept/lingua/liba/

Kampen, J. van (1997). First steps in Wh-movement. Delft, The Netherlands: Eburon.

Lewis, J. D., & Elman, J. L. (2001). Learnability and the statistical structure of language: Poverty of stimulus arguments revisited. In *Proceedings of the 26th annual Boston University conference on language development* (pp. 359–370). Somerville, MA: Cascadilla.

MacWhinney, B. (2000). *The CHILDES-project. Volume 2: Tools for analyzing talk: The database* (3rd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Perfors, A., Tenenbaum, J., & Regier, T. (2006, July). *Poverty of the stimulus? A rational approach*. Paper presented at the 28th annual conference of the Cognitive Science Society, Vancouver, Canada.

Reali, F., & Christiansen, M. H. (2003). Reappraising poverty of stimulus argument: A corpus analysis approach. In *Proceedings supplement of the 28th annual Boston University conference on language development*. Somerville, MA: Cascadilla.

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

Ritter, N. (Ed.). (2002). A review of the poverty of stimulus argument [Special issue]. *The Linguistic Review, 19*(1/2).

Saffran, J. R., Aslin, R., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science, 274,* 1926–1928.