

Bigram-based learning and the richness of the stimulus for language acquisition

**Xuân-Nga Cao Kam, Iglia Stoyneška, Lidiya Tornyova,
Janet Dean Fodor and William Gregory Sakas**

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 or specialized learning mechanisms. We report experiments based on those of Reali & Christiansen (2005), who demonstrated that a simple bigram language model can induce the correct form of auxiliary-inversion in certain complex English questions. We investigate the nature of the indirect evidence that supports this learning, to assess how generally it might be available. Results confirm the original finding but show that the model's success is highly circumscribed, resting on one particular word pair in the grammatical test sentences. The model performs poorly on inversion in related constructions in English and Dutch, which evidently do not afford effective cues accessible to the learning model. Other more powerful statistical models have so far been shown to succeed only on the same very limited subset of cases as the bigram model, which are clearly not indicative of broader success. The 'richness of the stimulus' for auxiliary-inversion thus remains unsubstantiated by this line of research to date.

1. Introduction

There has been renewed interest in the *poverty of the stimulus* argument (Chomsky 1980; see additional references in Ritter, 2002). Chomsky argued for the existence of innate linguistic knowledge (Universal Grammar, UG) on the ground that children show mastery of some properties of their target language before they have been exposed to relevant exemplars. His conclusion was that children must be biologically preprogrammed with knowledge of language facts unattested in their experience. The specific form in which this knowledge would be represented is not established by this argument, but UG is commonly taken to consist of a set of universal linguistic principles that interact with learners' observations of their particular target language. Other very different types of argument for biological specialization for language have also been proposed (species-specificity, a critical period for acquisition, rapid development of creole languages, etc.), but we will not address them here. We focus on evaluation of some findings that appear to undermine the poverty of stimulus argument.

The empirical foundations of the poverty of the stimulus (henceforth POS) thesis were decidedly slim when it was first propounded. Chomsky's case for it was based on informal intuitions about children's linguistic experience and the linguistic knowledge they come to have.

This was persuasive enough to defeat any behaviorist alternative to UG requiring extensive environmental shaping of linguistic abilities, and to energize an intense program of research on language structure and language acquisition. But the original commonsense evidence was not designed to withstand the impact of newer and more powerful UG-free models of learning (see Pereira, 2000), and very little additional evidence has been adduced since. What is remarkable, in view of the significance of the issues involved, is how little empirical work there has been in the intervening quarter century that has attempted to evaluate the truth of the POS thesis. Only recently has this important task begun to receive the attention it deserves.

It is clear what is needed. A relation has to be established between two kinds of data: evidence of the age at which children are exposed to some language fact, and evidence of the age at which children (ideally the same children) know that fact. But paired observations of this kind are rare. Recent research sparked by the increasing availability of corpora and computational techniques for searching them (MacWhinney, 2000) has brought the promise of more rigorous testing of the POS thesis, by making it easier to document what children say and what adults say to them, at what ages. However, even extensive corpus data are not well suited to definitively settling the POS issue. Proponents of POS may reasonably claim that corpus data overestimate the age at which learners have mastered some language fact F, since what children say in spontaneous daily talk, as registered in a corpus, is likely to be less advanced than what they *can* say (or understand) when the circumstances demand it, as in elicited production or comprehension experiments (Crain & Thornton, 1998). The age of mastery may also be overestimated by corpus data due to statistical problems resulting from the small samples typical of recordings of spontaneous child speech (Tomasello & Stahl, 2004). On the other hand, opponents of POS could note that similar problems attend estimates of the age of exposure to F, especially if exposure to a construction is taken to mean exposure to even a single instance of it. It is obviously impossible, for all practical purposes, to monitor every instance of F that a child hears prior to the established age of mastery of F, but the documented age of exposure is likely to be too high if based on child directed speech in corpora with the usual rather low sampling rate. The lag between first actual occurrence and first detected occurrence may be quite short for frequent constructions in adult speech, but could be several months for the relatively rare constructions that are likely to be the focus of POS debates (Tomasello & Stahl, 2004). Here too, when precision matters, experiments with children may fill in information that corpus studies cannot reasonably deliver. If children are taught a nonce word, or a syntactic construction in an artificial language, their exposure to it is under the control of the experimenter. But such studies represent a considerable expenditure of research effort, and have practical limitations of their own. Added to these problems of data collection are inevitable theoretical questions concerning the appropriate definitions of what should count as competence with respect to F (e.g., ability to imitate accurately) and what should count as exposure to it (e.g., whether overheard adult-to-adult conversation is a source for acquisition); extensive debate on these and related issues can be found in Ritter (2002).

A rare attempt to navigate these methodological shoals and pin a date on both exposure and attainment is a recent study by Lidz et al. (2003) of the phrasal status of the antecedent of the English anaphoric pronoun *one*, cited as an instance of POS by Baker (1978) and others since. Lidz et al. employed data from a comprehension experiment to establish the age of mastery of this fact (by 18 months), and corpus data to establish insufficiency of exposure at least until

about 5 years. Lidz et al.'s conclusion, in favor of POS, was that children's knowledge that the antecedent of *one* is phrasal could not have been learned, but this has since been challenged on various grounds by Akhtar et al. (2004), Regier & Gahl (2004) and Tomasello (2004), with response by Lidz & Waxman (2004). So far, then, it seems that the attempt to substantiate the POS claim for anaphoric *one* has not succeeded in convincing those who incline to the opposite view, and the debate remains open.¹

These continuing empirical uncertainties concerning the temporal relation between linguistic competence and linguistic experience are unfortunate because it seems likely that the shape of language research in the decades to come will be profoundly influenced by beliefs about whether or not the POS thesis is true. For this reason it is particularly newsworthy that a novel approach to evaluating POS claims has emerged in recent years which cuts right through these methodological complications. It turns attention away from the hunt for exemplars in learners' input and output. Instead, it offers a practical demonstration that innate specialization for language is not necessary for acquiring correct syntactic generalizations, *regardless of whether or not those generalizations are instantiated in the language the learner hears*. The demonstration consists in showing that the language facts in question can be acquired from a corpus of child-directed speech by a simple statistical learning algorithm with no access to any prior knowledge of language structure. Even young infants have been shown to be sensitive to statistical regularities in their input (Saffran et al., 1996; Saffran & Wilson, 2003), so if a simple statistical learner can generalize appropriately without aid of UG, it would be implausible to maintain that children cannot. In short, the research strategy of applying low-powered statistics in modeling language acquisition has the advantage that both sides of the POS debate may agree that whatever the statistical model can learn without aid of UG, children can too.

Note that for this purpose it is irrelevant whether or not the input contains any instances of the linguistic pattern that is attained, or even whether the investigator knows what facts the learning algorithm is responding to. It is proposed that children's input may provide 'indirect' evidence of the properties of a construction, which can be gleaned from other sentence types that children do hear. This indirect evidence would be missed by a traditional corpus study which searches only for examples of the target construction itself. The availability of indirect evidence sufficient for learning language fact F at a given age is established just by showing that F can be acquired by the statistical model from a corpus of adult speech to children at that age. Since the corpus is the total input to the model, not merely a sample, there is no room for concern that the facts of interest may have been acquired from sentences not in the corpus.

This approach could freely concede that exposure to some language fact F may follow mastery of F, if the facts should happen to fall out that way as empirical research proceeds. It is immune to such issues of time course because its demonstration of learnability offers an existence proof that the information is present in the input.² In this respect it breaks new ground,

¹ The POS thesis is that mastery may precede exposure. But conversely, in many cases children do not exhibit knowledge of a construction despite considerable exposure to it. In the absence of other explanations, this may suggest that a child needs to attain a certain maturational level in order to be able to take advantage of the input information provided; see Wexler (1999).

² This is a little too strong, since the corpus that provides the indirect information about fact F still needs to be representative of input to children at an age prior to their mastery of F, if the result is to be of psycholinguistic interest. But if it is assumed that the indirect cues come from sentences simpler than the F construction, which

surpassing conventional approaches by Pullum & Scholz (2002), Sampson (1997, 2002) and others to defending the ‘richness of the stimulus’ by pointing to instances of F in children’s input. The learnability demonstration nullifies not only the familiar POS claim that input examples of F are often lacking, but also the stronger claim that even if examples *are* available, learners could not represent them appropriately or project correct generalizations from them without guidance from UG. Thus, if stimulus poverty is to retain its status as a cornerstone of the argument for linguistic innateness, these demonstrations of learnability must be addressed.

An illustration of the ‘indirect evidence’ learnability argument against POS is provided by Reali & Christiansen (2003, 2005), building on work by Lewis & Elman (2001), who have applied this research strategy to the auxiliary-fronting construction in (1). Reali & Christiansen have shown that knowledge of this construction can be acquired by an extremely simple statistical learning model which refers only to pairs of adjacent words (bigrams) in sentences. Their bigram learning model, trained on a corpus of speech to one-year-olds, was able to select grammatical sentences such as (1) over ungrammatical versions such as (2), even though the corpus contained no sentences at all with the structure of (1).

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

In the discussion below we will refer to the grammatical sentence type illustrated in (1) as the *PIRC* construction, an abbreviation for *polar interrogative* with a *relative clause* modifying its subject. The ability of learners to discriminate between the correct and incorrect forms of auxiliary inversion³ in this construction was one of Chomsky’s first examples of stimulus poverty, and it has remained the classic example cited most often by POS adherents, presumably because it has been regarded as a compelling illustration of knowledge in the absence of experience. The centrality of PIRCs to the POS thesis is an important aspect of Reali & Christiansen’s challenge. The POS thesis cannot be defeated by falsifying it for any one construction, and it is not realistic to demand that it be disconfirmed for every construction; but if one of the most convincing cases succumbs to counterevidence, then there is *prima facie* reason to suspect that other cases would do so too if subjected to the same attention. Thus, there is a great deal at stake here. No doubt for this reason, the PIRC construction has been the focus of other recent ‘richness of the stimulus’ arguments also, by Clark & Eyraud (2006) and Perfors, Tenenbaum, & Regier (2006) as well as Lewis & Elman (2001), though these employ richer apparatus than simple bigram counts.

2. Bigram-based learnability of PIRCs

Reali & Christiansen (2005; henceforth R&C) report several experiments in which they tested a bigram language model, a trigram model, and a neural network model. We focus here on the former, since if a bigram-based model succeeds in acquiring PIRCs, it can reasonably be expected that the more powerful trigram and network models will do as well or better; we will

appear earlier or more abundantly in children’s input than F does, then it should be easier to demonstrate that mastery follows the availability of indirect cues than that it follows the availability of instances of F itself.

³ Following standard practice we refer to the inverting verbs as auxiliaries, though the examples often contain a copula (as in the relative clause of (1)/(2) above). See section 5.2 below on *do*-support of main verbs, and section 5.3 on inversion of main verbs.

comment on their performance briefly below. The bigram model was trained on a corpus of 10,705 utterances of child-directed speech extracted from a corpus of spontaneous adult-child conversations recorded and transcribed by Bernstein-Ratner (1984; available in the CHILDES database, MacWhinney, 2000). The children, whose native language was English, ranged in age from 13 to 21 months. Importantly, R&C note that there are no instances of the PIRC construction in the Bernstein-Ratner corpus, and hence none in the child-directed speech extracted from it. Therefore any information about PIRCs obtained from this corpus by the bigram model must be derived from other sentence types. The most likely candidates are simple (one-clause) polar interrogatives, and relative clauses in non-interrogative contexts. Our approximation of the R&C child-directed speech corpus (see below) contains 523 simple polar interrogatives, and 42 relative clauses in non-PIRC contexts.

In R&C's Experiment 1 the bigram model's knowledge of the PIRC construction was assessed by testing it on 100 pairs of test sentences similar to (1) and (2) above, generated semi-automatically from words occurring in the corpus. (It will be important below that words were individuated solely on the basis of their orthographic form since the training corpus was not tagged for part of speech.) Each pair of test sentences consisted of a grammatical and a matched ungrammatical version of a PIRC construction, 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 VP, PARTICIPLE, NP, PP, ADJP, etc.

Examples (1) and (2) above fit these templates and constituted one of R&C's test pairs. All the test sentences were novel: none of them occurred in the training corpus. Not all of the bigrams that constituted those sentences were in the corpus either, though every unigram (word) in them was. The goal was to have the model predict the grammaticality status of novel sentences, by projecting local regularities in the corpus such as captured in the bigram statistics. A value was assigned to each sentence of a test pair, based on the bigram statistics garnered by the model, and the sentence with the value that showed it to be more similar to those in the corpus was taken as the model's prediction of the grammatical form.

The value that R&C computed for each test sentences was its *cross-entropy*, which is a measure of the likelihood of that sentence occurring in the language domain from which the corpus is drawn (as predicted by the bigram language model). Specifically, the probability of each bigram in a test sentence was estimated (with smoothing; see below). The product of the estimated probabilities for all the bigrams in a sentence yields an estimated probability for the whole sentence. The negative log of the estimated sentence probability, adjusted for sentence length, gives its cross-entropy. Cross-entropy is inversely correlated with probability; hence a *low* cross-entropy is an indicator of higher likelihood that the sentence in question would occur in a language domain of which the corpus is a representative sample. 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 is a reasonable candidate for being the grammatical one, in R&C's forced choice test situation.

One point in particular concerning the method for estimating bigram probabilities will be central to our discussion below. A bigram consists of two adjacent words (unigrams) in a corpus.

The probability of the bigram is defined as the probability of its second word given its first word. For a bigram that occurs in the corpus, this can be estimated by counting occurrences of the word pair in the corpus and dividing by the number of occurrences of its first word (= Maximum Likelihood Estimate). For a bigram that does not occur in the corpus, some other means of estimating its probability is needed. A variety of alternatives have been proposed in the literature. R&C employed an *interpolation smoothing technique*, which makes use of the estimated probability of the second unigram, based on its frequency in the corpus. It is important to note that R&C’s smoothing formula applies to all bigrams, whether they occur in the corpus or not, giving equal weight to the bigram probability (which may or may not be zero), plus the probability of the second unigram (which is never zero, since only unigrams occurring in the corpus were used in the test sentences). To avoid confusion it is important to note that in the discussion that follows, when we say “bigram probability” we will mean the *smoothed* bigram probability (i.e., including the smoothing factor based on the second unigram). See Jurafsky & Martin (2000) for general discussion and motivation of n-gram formulae. The formulae employed by R&C, 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 : $H(s_T) = -\frac{1}{N_T} \log_2 \prod_{i=2}^{N_T} P_{interp}(w_i | w_{i-1})$

The bigram model’s prediction accuracy can be assessed as the percentage of test sentence pairs for which it selects the grammatical version. The result for R&C’s Experiment 1 is shown in Table 1.

Table 1

R&C’s Experiment 1: Percentage of sentences classified by the bigram model correctly or incorrectly as grammatical.

	% Correct	% Incorrect
R&C’s Experiment 1	96	4

The model’s performance was close to perfect: the 96% of test pairs correctly predicted was far higher than chance, and the mean cross-entropy of the set of all grammatical test sentences was significantly lower than that of the ungrammatical test sentences.⁴ On the basis of this strong

⁴ R&C’s Experiment 2 also tested their bigram language model on the PIRC construction, but on a smaller scale. There were just six test sentence pairs, based on the six sentences tested with children by Crain & Nakayama (1987). The training corpus was as in Experiment 1. The results were similar to those of Experiment 1: all six sentence pairs were correctly classified by the bigram model and the cross-entropy comparison for grammatical versus ungrammatical versions was statistically reliable. We do not discuss the details here.

positive result, R&C concluded that “these results indicate that it is possible to distinguish between grammatical and ungrammatical AUX-questions based on the indirect statistical information in a noisy child-directed speech corpus containing no explicit examples of such constructions.” More generally, they concluded that “there is sufficiently rich statistical information available indirectly in child-directed speech for generating correct complex aux-questions – even in the absence of any such constructions in the corpus”.

The bigram model’s performance in this experiment is impressive. We set ourselves the task of identifying *how* the model achieves its success. For example, we wanted to know: which bigrams in the grammatical and ungrammatical sentences resulted in the lower cross-entropy for the former; which sentences in the corpus of child-directed speech on which the model was trained provided those bigrams, and with what frequency; whether the relevant statistics are robust or are sensitive to small changes in the content of the corpus. These questions need to be asked. It is true, as we noted in section 1, that a learnability argument against POS goes through even if no-one knows what cues the learning model is picking up from the input; the model’s success is sufficient to show that they exist. But what those cues are is nevertheless important, for at least two reasons. One is that the bigram model’s level of attainment is startling from a linguistic perspective. Chomsky’s emphasis on structure-dependence as the basis for choosing the correct version of auxiliary inversion highlighted the fact that the correct rule refers to the hierarchical relationships in the syntactic structure of the word string prior to inversion.⁵ The evidence that the bigram model can extract the correct pattern for PIRCs seems to imply either that the model is computing hierarchical structure from the bigrams, or else that the aux-inversion rule is not after all defined over hierarchical structure. The latter conclusion would call for some almost unthinkable revision of current linguistic theory, since auxiliary inversion makes reference to phrasal structure in all current linguistic frameworks. Thus, it is in the interest of every linguist to understand *how* the bigram model does what it does.

A second reason for wanting to *understand* the bigram model’s performance is to be able to estimate which, and how many, other proposed POS cases are also susceptible to bigram-based learning. As we have noted, demolishing just one potential example of POS leaves the general POS thesis unharmed. But some examples are more potent than others. The potential significance of R&C’s result is all the greater because in the case of PIRCs there is no obvious transparent relation between word co-occurrences in the corpus and the structural relation that must be acquired. If the bigram model can succeed in such a case, it can be expected to perform as well or better in other cases where the structural fact to be acquired is more overtly reflected at the word level. Thus, in order to gauge the impact of R&C’s finding on linguistic theory and the status of UG, we need to find out how indicative the PIRC construction is. Only if it is a genuine straw in the wind, a harbinger of many other such learning achievements, will the conclusion follow that special language-specific principles are not needed to guide syntax acquisition.

In the following sections we report the method and outcomes of our investigation. To anticipate: we find that R&C’s finding is replicable but is very restricted in scope. The bigram model succeeds for the specific sub-type of PIRC specified by (3), with “is” as the sole auxiliary in both clauses, and relativization of the subject of the subordinate clause. We will call this the

⁵ Strictly speaking there is no auxiliary inversion *rule* in current transformational treatments. The same facts are ascribed instead to general principles and constraints; see Fodor & Crowther (2002).

is-is PIRC construction.⁶ It is not known whether the bigram model’s good performance will generalize to the range of similar constructions with other auxiliaries, or with mixed or multiple auxiliaries.⁷ Our data show that it does not extend to PIRCs with *do*-support, or PIRCs with relative clauses that have object gaps rather than subject gaps. This shows that the corpus does not supply adequate bigram information that pertains either directly or indirectly to these other sub-types of the auxiliary inversion construction. The bigram model also fails for a comparable corpus of Dutch child-directed speech in which, unlike English, the main verb can invert with the subject when there is no auxiliary in the sentence. From these findings, and the reasons for them which we uncover below, it can be concluded that the success of R&C’s bigram model for the *is-is* type of PIRC in English is a mere happenstance, which offers no encouragement for thinking that other constructions or other languages will also be learnable without the aid of UG.

3. Understanding the bigram model’s success

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

Before conducting new bigram experiments, we replicated R&C’s experiment to make sure that the training corpus, the test sentences, and the computation of cross-entropy that we were employing were in accord with theirs. Following R&C, we used the Bernstein-Ratner corpus (the version that was not tagged for part of speech). We extracted from it all and only the utterances by adults, and deleted those which seemed almost certainly to have been addressed to other adults; the training corpus was thus limited to adult-to-child speech as in R&C’s experiment. This yielded a set of 9,643 utterances, similar to R&C’s corpus from the same source. We manually constructed 100 pairs of test sentences conforming to R&C’s templates in (3) above; 40 pairs had relative pronoun “who”, 60 pairs had “that”.⁸ We computed smoothed bigram probabilities, and cross-entropies for all test sentences, using the same formulae as R&C (see above). We then examined whether the cross-entropy of the grammatical sentence of a test pair was lower than that of its matched ungrammatical sentence; if so, we counted this as selection of the grammatical version.

⁶ By our definition above, the PIRC construction is not limited to *is-is* versions. We have restricted our discussion to *is-is* forms so far because R&C confined their experiments almost exclusively to them. In this, R&C were following Chomsky’s lead, and also that of Crain & Nakayama (1987) who tested primarily the *is-is* variety with children. Our own results reported below extend to a broader range of PIRCs. Even more varieties of complex interrogatives with aux-inversion have been noted in the POS literature, e.g., complex wh-questions, and questions with an adverbial clause rather than a relative clause (as in *When the little boy is crying, is he unhappy?*). Whether the presence of such examples in children’s input is relevant to the acquisition of sentences like (1) has been the subject of debate; see the papers by Pullum & Scholz, Sampson, and others in Ritter (2002).

⁷ In R&C’s Experiment 2, one of the six items tested (based on an example in Crain & Nakayama’s 1987 psycholinguistic study) had two auxiliaries in the relative clause; another had “was” in one clause and “is” in the other. Both items were correctly judged by the bigram language model. Crain & Nakayama’s Experiment 2 tested “can” and “should” with children, but R&C did not test these auxiliaries in their bigram experiments. For additional child data, see Ambridge, Rowland, & Pine (in press) for an experiment in which children made occasional errors on PIRCs with “can”. There appear to be no other empirical data on children’s abilities with respect to other subvarieties of PIRC.

⁸ It may not be proper to refer to the word “that” introducing a relative clause as a relative pronoun, but for convenience here we will do so.

The results, presented in the first line of Table 2, though not quite as impressive as R&C's, clearly corroborate the success of the bigram model for sentences of this type. (For convenience, we present the data from all of our experiments in the same table; experiments 2-6 will be discussed individually below.) The "undecided" category in the last column of the table reflects cases where the two versions of a sentence were equal in cross-entropy. Figure 1 portrays the data in Table 2 graphically. (All test sentences are available at http://www.colag.cs.hunter.cuny.edu/pub/Bigrams_Richness_Experiments.zip.)

Table 2

Percentage of sentences classified by the bigram model correctly or incorrectly as grammatical, or undecided, in the six experiments reported here.

	sentences tested	% correct	% incorrect	% undecided
Expt1 - Replication of R&C	100	87	13	0
Expt2 - Disambiguated rel pronouns	100	18	36	46
Expt3 - Homography with determiner	100	18	37	45
Expt4 - Object gap relative clause	100	35	15	50
Expt5 - <i>Do</i> -support	100	49	51	0
Expt6 - Verb inversion in Dutch	40	32.5	55	12.5

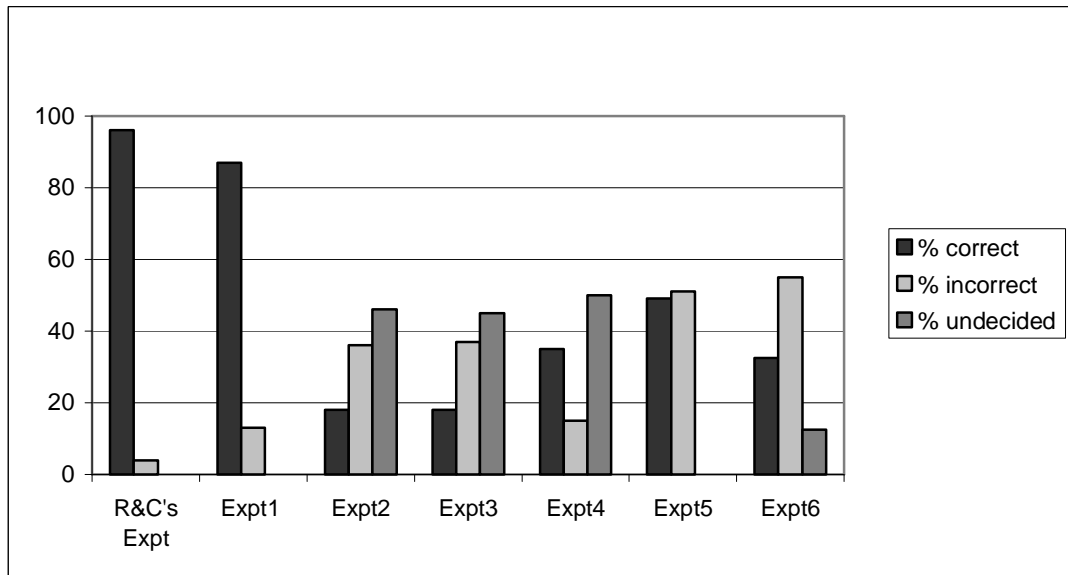


Figure 1. Percentage of sentences classified by the bigram model correctly or incorrectly as grammatical, or undecided, in R&C's Experiment 1 and the six experiments reported here.

With these data in hand, we were able to look more closely into how the bigram model selects the correct sentence.

3.2. Which bigrams favor the grammatical sentences?

Discrimination between the grammatical and ungrammatical sentences in a test pair defined by (3) 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.⁹ Consider the pair (4) and (5), from among our test items.

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

We present their distinguishing bigrams (by order of their appearance in the sentences) in Table 3, where we have numbered them for ease of reference. (For example, *<bigram1-grammatical>* is the first distinguishing bigram in the grammatical sentence.) The non-distinguishing bigrams in these sentences are: *<is the>*, *<the little>*, *<little boy>*, and *<boy who>*.

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

Test sentences	Bigram1	Bigram2	Bigram3
(4) Grammatical	<i><who is></i>	<i><is crying></i>	<i><crying hurt></i>
(5) Ungrammatical	<i><who crying></i>	<i><crying is></i>	<i><is hurt></i>

As noted, all unigrams in the test sentences occurred in the corpus, though not all of the bigrams did. For those that did not, the bigram probability is estimated based *solely* on the estimated probability of the second unigram (see section 2). As a direct consequence of R&C's templates in (3) by which these test sentences were created, bigram1-grammatical was *<who is>* or *<that is>* in every test pair, and both of these bigrams did occur in the corpus.¹⁰ As we explain below, this gave the *<who is>* or *<that is>* bigram (which we abbreviate as *<who/that is>* in what follows) the greatest influence on performance in the sentence discrimination task. Table 4 below shows the smoothed bigram probabilities for the sentence pair (4)/(5). In each cell of the table, 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, following R&C's smoothing formula which gives equal weight to both. For better visualization, the table presents the probabilities multiplied here by 100,000.

⁹ This includes all bigrams containing the initial and final sentence boundary markers. Since these are not distinguishing, we omit them from the data analyses; but see discussion of Experiment 6 below.

¹⁰ All statements in sections 3, 4 and 5 about the contents of the training corpus refer to our own corpus, modeled on R&C's as noted above. We believe that any discrepancies between the two are sufficiently slight that our factual statements here can be taken to hold equally for R&C's experiments.

Table 4

Smoothed probabilities (x 100,000) for the six distinguishing bigrams in sentences (4) and (5). (See text for explanation of shading.)

	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

The relationships among these six bigrams are crucial to the outcome of the experiment, yet they are to some extent determined by the experimental design. As a consequence of the templates in (3) that define the test sentences, bigram2-grammatical in (4) and bigram1-ungrammatical in (5) have the same second unigram (here: “crying”). Since in this case it happens that neither bigram occurs in the corpus, the smoothed probability of each is based solely on that second unigram.¹¹ Hence, when the probabilities of the bigrams in each sentence version are multiplied together to give the estimated sentence probability,¹² these two bigrams effectively cancel each other out and play no role in the model’s selection of one version. (We use shading in Table 4 and subsequent tables to indicate 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 rather few bigrams in the test sentences do occur in this (relatively small) corpus. For test pairs where one or other of these two bigrams does occur, the one in the grammatical test sentence is likely to outweigh the one in the ungrammatical version (here: <who crying>) which is a legal English sequence only in rare contexts (e.g., *He’s a man who crying amuses*) and so is unlikely to occur in the corpus.¹³ Aggregated over the set of test sentences, this can be expected

¹¹ The fact that the unigram “crying” appears in the corpus but the bigram <is crying> does not is perhaps surprising. We checked and found that “crying” occurs following “you”, “she’s” and “he’s” only. Note that a reduced auxiliary verb as in “she’s” was not treated as an independent unigram, following R&C’s practice (which possibly was intended to mirror the inability of children of this age to recognize reduced forms). This may, however, have resulted in the exclusion of some potentially relevant examples, such as *What’s that animal we saw at the zoo yesterday?*, which occurred in the training corpus and might (depending on its proper analysis) be a PIRC. (All three examples of child-directed PIRCs cited by Pullum & Scholz, 2002, in their empirical assessment of POS had a reduced “is” in “where’s”.)

¹² We evaluated the bigram model on the basis of the cross-entropies of sentences, just as R&C 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. The probability of a sentence is simply the product of the probabilities of all the bigrams that compose the sentence. This expository decision has no effect on the facts reported, since cross-entropies and probabilities are intertranslatable; they are inversely proportional. The sentence of a test pair with the higher probability (lower cross-entropy) was taken to be the one selected by the model as grammatical.

¹³ As this illustrates, a bigram that is illicit in a test sentence may nevertheless occur in the corpus as part of a different construction. This is especially so in the present experiments because words were individuated orthographically, as noted above. For example, the non-finite passive/adjective “hurt” in (4)/(5) is not distinguished in the corpus statistics from the active verb “hurt”, which has other privileges of occurrence (e.g., it occurred in the sentence *You might hurt the doggie*).

to tilt the bigram model's choice towards the grammatical version. A comparable but opposite relationship holds between bigram3-grammatical and bigram3-ungrammatical. When neither is attested in the corpus, their estimated probability depends on their second unigram, which by template (3) is always identical, so they balance each other out in the discrimination task and contribute nothing. But in this case, when one is attested it is more likely to be the one in the ungrammatical sentence. Neither is illicit in English, but bigram3-grammatical (here: <cry^{ing} hurt>) consists of the last word of a relative clause followed by a non-finite predicate, an uncommon sequence in English (except in PIRCs, which are not represented in the corpus).¹⁴ Although this word sequence might appear in the corpus in some other guise (e.g., *Too much crying hurt her eyes*), on balance it is probably less likely than bigram3-ungrammatical, which is part of a 'normal' finite predicate. Thus, when these two bigrams do not fully cancel out, the comparison between them will typically (though not necessarily) tilt the model's choice toward the *ungrammatical* version.

In short: Of the six distinguishing bigrams in test sentence pairs built according to the design templates, four match up across sentence versions in such a way that no *systematic* advantage is expected for either the grammatical or the ungrammatical version. If the bigram model's performance were based on these four bigrams, it would not consistently select the grammatical version. (Our data confirm this; the success rate would be 16%, with 36% incorrect and 48% no-choice.) It is thus the remaining two bigrams, bigram1-grammatical and bigram2-ungrammatical, which create the model's strong bias toward the grammatical sentence. These bigrams also necessarily share their second unigram, so they too would annul each other in the sentence discrimination task if neither bigram were attested in the corpus. However, in this case the bigram in the grammatical version is *always* attested. Template (3) entails that bigram1-grammatical is a fixed form, either <who is> or <that is> in every case. The training corpus used in these experiments does contain these bigrams: there are 12 occurrences of <who is> and 23 occurrences of <that is>. ¹⁵ Therefore bigram1-grammatical is *guaranteed* to have a higher smoothed probability than bigram2-ungrammatical whenever the latter is not in the corpus, and that will push the bigram model toward correctly selecting the grammatical sentence. Whether bigram2-ungrammatical does occur in the corpus or not varies considerably from one test sentence to another. This bigram consists of the last word of a relative clause followed by "is". In (5) the relative clause ends in a verb, so it is unlikely to be followed by "is" (except, e.g., in *Crying is bad for your eyes*). On the other hand, the relative clause may end in a noun, and that noun followed by *is* may be a quite frequent bigram in the corpus (e.g., <baby is>).¹⁶ If its probability exceeds that of the <who/that is> bigram, the model could choose the ungrammatical

¹⁴ For some test sentences, e.g., those in which the relative clause ends in a noun and the matrix predicate consists of a prepositional phrase once the "is" has been fronted in the grammatical version (e.g., *Is the box that is wrapped in blue paper for Paul's birthday?*), bigram3-grammatical could be a common word sequence such as <paper for>, which could have a better chance of being attested in the corpus than an example like <cry^{ing} hurt> in sentence (4).

¹⁵ These corpus frequencies may appear quite modest, but comparatively speaking the estimated probabilities of <who is> and <that is> are high. Their estimated probabilities are the 2nd highest and 7th highest, respectively, of all 391 distinct distinguishing bigrams in the 100 test pairs.

¹⁶ For example, this was the case in two of the four sentence pairs for which the bigram model preferred the ungrammatical version in R&C's experiment (**Is the jacket that on the chair is lovely?* and **Is the dog that on the chair is black?*) The other two ungrammatical items that were incorrectly selected in that experiment were **Is the lady who here is drinking?* and **Is the alligator that there is red?*. Here too, the bigram2-ungrammatical (<here is> and <there is> respectively) happened to be a very frequent word sequence which could outweigh bigram1-grammatical.

test sentence. However, by contrast with the variability of bigram2-ungrammatical, the reliable presence of the <who/that is> bigram in the grammatical sentence gives a steady boost to the grammatical sentence, enough that it emerges as the winner in many cases.

We have delved into these details concerning how the six distinguishing bigrams are likely to trade off against each other because they drive the outcomes of R&C's experiment and our own. Though seemingly trivial in themselves, they have a powerful effect because they apply very broadly to all the test materials created from R&C's templates. One and only one bigram (bigram1-grammatical: <who/that is>) features among the six distinguishing bigrams in every test pair, and in every case this bigram is in the *grammatical* sentence of the pair. It thus serves as a 'marker' for the grammatical version. Since it occurs with greater frequency in this corpus than many other bigrams, it very often dominates the calculation. This is why the bigram model has such a robust preference for the grammatical PIRC.

This analysis sheds light on our numerical data. For 47 of the 100 test pairs, <who/that is> was the only one of the distinguishing bigrams that was attested in the corpus, so the probability of the grammatical test sentence was necessarily higher than that of the ungrammatical test sentence and the bigram model always selected the grammatical version. In another 23 cases, one or more of the distinguishing bigrams in the ungrammatical version were attested, but <who/that is> was the only one of the three distinguishing bigrams in the grammatical version that was attested, and in 15 (= 65%) of those cases its probability was high enough to defeat the ungrammatical version. In another 18 cases, one or both of the other distinguishing bigrams in the grammatical sentence were attested also, but the probability of <who/that is> was sufficiently high that it would have defeated the ungrammatical version even without their assistance. In total, then, 80 (= 92%) of the 87 positive outcomes can be traced specifically to the bigram <who/that is>. The exact success rate in such an experiment will of course vary somewhat with the particular sentence pairs employed in the test phase as well as with the details of the corpus.¹⁷ But for the training corpus in this experiment, which may well be typical in this respect, it is clear that 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*

A clear conclusion from our replication of R&C's Experiment 1 is that the <who/that is> bigram does the lion's share of the work in predicting the correct version of PIRC sentences. This follows from a more general recipe for success for a bigram learning model: a bigram model will have its best chance of performing well in a sentence discrimination task if there is a bigram (or more than one) which (i) appears fairly systematically in the grammatical test sentences and not in the ungrammatical ones (such as the <who/that is> bigram in the case of PIRCs), and (ii) has a high estimated probability with respect to the training corpus relative to that of other bigrams. When these conditions are met, discrimination will probably succeed;

¹⁷ 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 cases. 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. These differences are small compared to the effect of <who/that is>, but are in the anticipated directions.

when they are not, success is possible but cannot be counted on.

To further evaluate R&C's finding, therefore, attention must focus on the <who/that is> bigram. It was powerful in guiding the correct discriminations for PIRCs because it was always present (by design) in the grammatical test sentence and was quite frequent in the corpus. In view of the young age of the children to whom the utterances in the corpus were addressed, it seemed surprising to us that the corpus contained enough relative clauses to supply all these <who/that is> bigrams. Indeed, a search for relative clauses revealed only 19 that contained an overt relative pronoun (4 with "who", all of which were subject-gap relatives; 15 with "that", of which 9 were subject-gap and 6 object-gap relatives). None of these contained a <who is> or <that is> bigram; the relative pronoun was followed by a lexical verb, such as "lives" in *I found the doggie that lives in this house*, or by a nominal in an object-gap relative such as *I saw somebody that you like*.¹⁸ The source of the <who/that is> bigram therefore could not have been relative clauses. Instead, we found that all 12 <who is> tokens appeared in questions (e.g., *Who is in there?*), and all 23 <that is> tokens occurred with the "that" as a deictic pronoun (e.g., *That is a rose*).¹⁹ Thus the "who" or "that" of a <who/that is> bigram was in every case merely a homograph of a relative pronoun.²⁰ In other words, the ability of the bigram model to predict PIRCs, which rests primarily on the existence of the <who/that is> bigram in the corpus, was due to a <who/that is> bigram that had nothing to do with relative clauses.

Summary: We set out to uncover the linguistic relationship between the evidence provided by the corpus, and the grammatical discriminations made possible by that evidence. We did so in order to be in a position to assess whether a similar relationship would hold in other potential cases of stimulus poverty, in which case the bigram model might very well succeed for them also. We found no grounds for supposing that the model succeeded with PIRCs because it was responding in some way to the hierarchical structure of the test sentences, as would be implied by Chomsky's claim that structure dependence is the key to acquisition of correct auxiliary inversion. Rather, the supportive relationship between the training corpus and the test items rested on the linear adjacency of just two words. The potency of those two words was found to be due to an accidental fact of English, or rather, to *two* accidental facts. One is that the English language contains words that are not relative pronouns but have the same orthographic form as the relative pronouns. The second is that these other words quite commonly occur immediately

¹⁸ One had "who's" with contracted *is*, but a contracted auxiliary was not analyzed as a separate unigram, following R&C's practice as noted above (footnote 11), so this did not count as a <who is> bigram.

¹⁹ One of the 23 <that is> bigrams was possibly a deictic determiner in a disfluent sentence. The unigram "that" also occurred as a complementizer, as in *Tell grammy that you're gonna come and swim in her lake*, and as a determiner as in *Look at that dolly*, but in these roles it was never followed by "is". The complementizer or determiner "that" does nevertheless have an effect on the bigram probability calculations since it is included in the denominator in calculating the Maximum Likelihood Estimate of bigram probability (see section 2). For example, if the corpus contains many instances of complementizer "that" followed by something other than "is", this lowers the estimated probability of the "that is" bigram.

²⁰ 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 will refer to homography throughout, even where it is strictly inappropriate as in our brief excursions into discussing child language acquisition. Note, though, that this may be more than a terminological issue. Quite possibly, "who" and "that" as interrogatives or deictics would have been prosodically distinguishable from "who" and "that" as relative pronouns in the original conversations, though not in the transcriptions in the corpus used in the experiments.

preceding “is”.²¹ Hence the PIRC construction holds no promise of success on other constructions, or even on PIRC constructions in other languages if they lack these idiosyncrasies. There are many learnable natural languages (e.g., Finnish, Hebrew, Yoruba) in which the relative pronoun does not look or sound like any other word.²² There are also languages in which relative pronouns are homographs of other words than in English; for example, in German many relative pronouns have the same form as definite determiners. It can be anticipated, therefore, that the bigram model would be less capable of discriminating PIRC constructions in such languages – though it might do well on other languages, such as French, which have homography (and homophony) not unlike that of English. We conducted two experiments to confirm this, which we report in Section 4 before moving on to examine the breadth of the bigram model’s learning ability in Section 5.

4. Without the ‘wrong’ bigrams

It is predicted that without the facilitating effect of the ‘wrong’ <*who/that is*> bigram in the grammatical version, the bigram model would be unable to discriminate between grammatical and ungrammatical PIRCS. In one experiment, the overlap between relative pronouns and other forms in English was eliminated; relative pronouns were labeled as such in order to disambiguate them. In the other experiment, there was overlap with another form, but it did not contribute to the probability estimate for the critical bigram in the grammatical PIRC. The language tested in that experiment was identical to English except that the relative pronoun was a homograph of a determiner.

4.1. Experiment 2: Disambiguating the relative pronouns

Starting with the same corpus as in Experiment 1, we investigated the bigram model’s performance on a language exactly like English except lacking the English surface similarities between relative pronouns and interrogative and deictic pronouns. For this purpose we repeated the experiment as before after labeling all relative pronouns in the corpus and the test sentences 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 (illustrated in Table 4 above), except that “who-rel” or “that-rel” appeared in place of “who” or “that” in bigram1 in both sentence versions, and the first term

²¹ It is also crucial to the model’s success for *is-is* PIRCS that there is no invariant morpheme *X* at the end of every subject noun phrase (or at the end of every relative clause, like “de” in Chinese.). If there were, the distinguishing bigram <*X is*> would function as a ‘marker’ for the ungrammatical version in every test pair, showing that the main clause “is” has not been fronted. If this bigram were substantially present in the corpus (e.g., in declarative sentences), it might raise the estimated probability of ungrammatical PIRC versions over that of grammatical versions even despite the <*who/that is*> bigram in the latter. In other words, another important contributor to the model’s success for English is the language-specific fact that, contrary to this, there is almost no limit to what the final word of an English relative clause may be.

²² The linguistics literature provides no definitive count of such languages, but an informal poll conducted through Linguist List also yielded Avestan, Haida, Hungarian, Kambara, Kiswahili, Malay, Scots Gaelic, Thai and Zulu among languages whose relative pronouns are not phonologically identical to interrogative pronouns or other morphemes in the language. (See also extensive information on relative clause markers and pronouns in de Vries, 2002.) The generations of children who have acquired these languages thus were unable to benefit from overlaps with interrogative sentence bigrams. (Caution: Our informal survey did not establish whether all these languages have PIRC-creating verb/auxiliary inversion.)

of the estimated probability of bigram1-grammatical was always 0 since the relative pronouns never preceded “is” in the corpus. The results, as expected, were not in favor of the bigram model. They contrasted strongly with the results of Experiment 1; see the second line of Table 2 above, which shows the percentage of sentences in Experiment 2 that were correctly or incorrectly classified by the bigram model as grammatical, and the percentage in which it had no basis for choosing one or the other.

With these disambiguated relative pronouns, the bigram model failed to select the grammatical version of the PIRC for 82% of the test pairs. This is as expected on our diagnosis of what makes for success in bigram-based discrimination. Unlike the original corpus, the language with unambiguous relative pronouns lacks any bigram that is both attested in the corpus and appears more systematically in the grammatical test sentences than in the ungrammatical ones. The <who-rel/that-rel is> bigram was in all the grammatical test sentences but never occurred in the corpus, whereas unlabeled <who/that is> bigrams (in which the “who|that” was not a relative pronoun) occurred in the corpus but not in the test sentences. The outcome was therefore more varied than in Experiment 1. When none of the six distinguishing bigrams in a test pair were in the corpus (as in the case of *Is the little boy who-rel is crying hurt?* versus **Is the little boy who-rel crying is hurt?*), their smoothed bigram probabilities all canceled out across the grammatical and ungrammatical test sentences (including bigram1-grammatical and bigram2-ungrammatical, unlike Experiment 1 where that never occurred), so the bigram model had no basis for choosing either version. The high number of undecided outcomes in this experiment is attributable to the low incidence of many of the distinguishing bigrams in this relatively small corpus. When one or more of the five distinguishing bigrams other than bigram1-grammatical did occur in the corpus, the outcome depended on whether they contributed more to the grammatical or the ungrammatical sentence; there was no systematic preference for the grammatical version. (A test pair that was correctly discriminated was *Is the man who-rel is in the pool swimming?* versus **Is the man who-rel in the pool is swimming?*. A misclassified example was **Is the little house that-rel behind the tree is the doghouse*, which was wrongly selected over *Is the little house that-rel is behind the tree the doghouse?*). The disparity between these results and those of Experiment 1 (and of R&C’s Experiment 1) exposes the considerable impact of the ‘wrong’ <who/that is> bigrams in creating the positive outcomes of those previous experiments.

Natural languages not infrequently exhibit homography among their lexical items, including their functional categories (‘closed class items’ such as prepositions, complementizers, determiners and particles), although the forms with multiple functions differ from one language to another (e.g., “so” in English, “-no” in Japanese). The difference between success in Experiment 1 and failure in Experiment 2 shows that such homography can boost the probabilities of influential bigrams. Conceivably, then, the right moral to draw from these experiments is not that the bigram model’s success in Experiment 1 was spurious, but that homography (really homophony, of course) can be a useful bootstrapping device for learners which they should exploit whenever possible. This is an interesting possibility. Is it what children do? And if so, does it help them in acquiring PIRCs? Drawing inferences for human language acquisition from the performance of abstract computational models is of course a tricky matter, but this idea is certainly worth thinking through. Since there are, to the best of our knowledge, no child data on this topic, we must consider both possibilities.

Imagine, then, a child who relies on bigram statistics to predict the correct forms of syntactic constructions in the target language. Consider first a ‘non-conflating’ child who, when initially exposed to relative clauses, is able to distinguish relative pronouns from other pronouns, even ones that sound similar, on the basis of their distribution or prosody. Such a child would be in the situation of the bigram model in our Experiment 2 with unambiguous relative pronouns: lacking a robust cue for the grammatical form, the child’s performance if tested on PIRCs would be poor. Now suppose instead that children learning English do at first conflate relative pronouns with deictic and interrogative pronouns. Analogy with the bigram model indicates that they would benefit from this, in that they would do well on discriminating grammatical from ungrammatical PIRCs even without any experience of relative pronouns. However, this bootstrapping strategy would predict a striking pattern of errors on other constructions until such time as the child eventually attains an adult-like ability to distinguish the various subtypes of pronouns from one another. For example, a ‘conflating’ child could presumably accept ungrammatical sentences with “this” mis-used as a relative pronoun in place of “that” (e.g., **Hug the boy this is crying*), or sentences in which relative pronoun “who” wrongly triggers inversion (e.g., **Hug the boy who is the dog barking at*) because interrogative “who” does so.²³ Thus, some testable empirical predictions flow from the suggestion that there is no stimulus poverty for children’s learning of English PIRCs because of the abundance of “who” and “that” in constructions other than relative clauses.

To summarize: Our Experiment 2 data confirm that the bigram model succeeded on PIRCs in Experiment 1 by basing its evaluations of relative clauses on facts about questions and demonstrative expressions. A strategy of bootstrapping one construction via superficially similar words that occur in other constructions appears to be a mixed blessing, yet without this the bigram model was unable to find any indirect evidence for PIRCs. It would be of considerable interest to know whether children do make homograph-conflating mistakes such as illustrated above, in contexts where relatives, interrogatives and deictics do not behave alike. We will proceed along another track here. English does at least offer ‘wrong’ bigrams which children might – or might not – take advantage of in discriminating PIRCs. But since not all languages do so, we next consider a language in which relative pronouns are unhelpfully homographic with other items in the language.

4.2. Experiment 3: Homography with a determiner

²³ In R&C’s Experiment 3 a connectionist model (a simple recurrent network) was tested on *is-is* PIRCs by the predict-the-next-word procedure. The training corpus was the same as for their Experiment 1 except that each word was replaced by one of 14 part of speech tags. This form of input is highly conflating. There was a single tag “PRON” for all pronouns in the corpus and test sentences, so relative pronouns were conflated with every other subclass of pronoun, not just interrogatives and deictics but also personal pronouns such as “she”, “our”, etc. The SRN performed well; it predicted V (a verb/auxiliary) more strongly than any other part of speech following a sequence such as V DET N PRON... (corresponding to an English sentence fragment such as “Is the boy who...”). The SRN’s experience of pronoun-verb sequences of all kinds (e.g., “She sings”, “What was that?”) could strengthen its expectation of V following PRON, leading to success in the PIRC test items. But again, this would generate a host of errors on other sentences. Represented simply as the part of speech categories, ungrammatical sentences such as **It/she/this did you say?* would be as acceptable as *What did you say?*; and **I see the boy him is crying* would be as acceptable as *I see the boy who is crying*. In our own studies with tagged input (Kam, 2007), we use fine-grained tagging (over 100 categories) to avoid these potential problems. (An SRN study by Lewis & Elman, 2001, used non-tagged input; see section 6 below.)

Not all languages are like English with respect to the double fact that relative pronouns have homographs, and the homographs often occur in the same local contexts as relative pronouns. It is this that raises the estimated probability of the *<who/that is>* bigram which biases the bigram model toward the grammatical PIRC. It can be expected that the bigram model would fare less well with other languages, even languages whose relative pronouns do have homographs, if the syntactic category of the homographs is not such that they can be followed by a verb.

A definite determiner is a good candidate for this role. It is a functional (closed-class) item with high corpus frequency, and, as noted, there are natural languages with relative pronouns identical in form to definite determiners, but a definite determiner is not likely to be followed by a verb. In Experiment 3, therefore, we substituted the word “the” for all the relative pronouns in the original (unlabeled) corpus and test sentences, to check that this homography does *not* help the bigram model to distinguish the grammatical and ungrammatical versions of PIRC constructions. Note that we chose to edit the English corpus in this way rather than turning to a natural language such as German that exhibits this kind of homography. We did so in order to isolate this one factor of homography from all the other syntactic differences between English and another language (e.g., in the case of German: verb-second word order in main clauses; verb-final order in subordinate clauses; case and agreement features on relative pronouns and determiners; etc.) which could also influence the model’s performance on PIRCs in uncontrolled ways. (Our experiment on Dutch, reported below, shows that such differences do indeed impinge on PIRC performance. The possibility that other natural languages offer other bigram cues not available in English is addressed in that experiment.) It is true that English with relative pronouns pronounced like “the” is not a language that is spoken by anyone, but there is no reason to doubt that it is a learnable human language.

The distinguishing bigrams in the test sentences were as illustrated in Table 4 for Experiment 1, except for bigram1-grammatical and bigram1-ungrammatical which both had “the” in place of “who” or “that”. The estimated probabilities of those two bigrams therefore differ from Table 4. The first term of the estimated probability of bigram1-grammatical, which was always *<the is>*, was 0 in all cases. Bigram1-ungrammatical, containing “the” in place of “who” or “that” varied across sentences; its estimated probability was generally low. Since only this bigram differed in estimated probability from Experiment 2, it is predicted that discrimination task outcomes will be quite similar to those of Experiment 2, with few correct selections and frequent inability to choose. The results are shown in the third line of Table 2. They are indeed just as poor as for Experiment 2. The reason for this failure is also similar to that for Experiment 2. Bigram1-grammatical, which is *<the is>* in Experiment 3, systematically appears in the grammatical version and not the ungrammatical version of every test sentence pair, but it is not a useful ‘marker’ for the grammatical version because it does not occur in the corpus. Outcomes therefore fluctuate between unsystematic selection of grammatical or ungrammatical versions when some of the other distinguishing bigrams do occur in the corpus, and “undecided” responses when all six are unattested.

Thus it is confirmed that the bigram model does not benefit from just any overlap between relative pronouns and other words in the language. The high performance level in R&C’s Experiment 1 and in our replication of it rests on a peculiar confluence of facts about this

particular construction in English. Straying from this situation even in small details leaves the bigram model with no cues, direct or indirect, for predicting the grammatical form of PIRCs. This strengthens the notion that the bigram model's success for English PIRCs does not augur similar learning achievements for other constructions or other languages. To evaluate this, we extended our investigation to a wider range of PIRC constructions.

As we did so, it became clear that our general recipe for bigram-based learning success could be made more precise. First, as noted above, the bigram (or bigrams) responsible for correct sentence discrimination must be in the grammatical version. Two adjacent words that should *not* co-occur (e.g., <*the of*>) would be a clear indication of ungrammaticality to a human adult language user; but for the bigram model of these experiments, the *non*-occurrence in the corpus of such a word sequence has no more import than the non-occurrence in the corpus of any legitimate but unlikely word combination (e.g., *green ant*). Second, in the ideal case a marker bigram for the grammatical sentence would consist of two function words (functional categories; closed class words). Unlike lexical categories (nouns, verbs, etc.), these items appear in many sentences that otherwise differ greatly in their content. Hence a single bigram consisting of a pair of function words (like *who* and *is*) can do a great deal of work; it can flag a grammatical construction through almost unlimited variation in sentence meaning and vocabulary. This is not the case if either or both of the unigrams is a lexical category (e.g., <*book is*>, <*who jump*>). Finally, for optimum usefulness, the two words that compose the crucial bigram must reflect in some fashion, however indirectly, a linguistically relevant fact about the grammatical version. The <*who/that is*> bigram does this particularly well. The relative pronoun which is its first unigram proves that the bigram is recording a fact about the relative clause rather than the main clause. The adjacent finite auxiliary proves that there has been no auxiliary fronting from that clause. In the forced choice situation of these experiments, this entails that the auxiliary in the main clause *has* been fronted.

Turning now to additional variants of the PIRC construction, we find that for one reason or another they lack bigrams which satisfy these criteria of recurrence and linguistic informativeness. In one case (object-gap relative clauses), even when “who|that” and “is” are both present in a sentence they are not adjacent, so they fall beyond the scope of any one bigram. In another case (lexical verbs needing *do*-support), the word that follows “who|that” is a lexical verb rather than a function word, so the marker bigram is different for each test sentence and all those bigrams would have to appear in the corpus for successful performance. (See section 6 for discussion of a potential solution.) Thus, while *is-is* PIRCs are perfectly tailor-made for bigram-based learning, it appears that these other subtypes of the English PIRC construction do not lend themselves to it at all well. Therefore we predict that they will not be well judged by the bigram model. We now report two experiments which document this. In these experiments we revert to the original corpus as in Experiment 1 (with no labeling or replacement of relative pronouns), in order to put aside now the issues of homography and ‘wrong’ bigrams; to the extent that those are helpful, they are available once again to the learner in Experiments 4 and 5. The learning failures we observe in these experiments are therefore independent of those in Experiments 2 and 3.

5. Extending the investigation to more PIRCs

So far we have followed the original bigram-based learning experiment by R&C in limiting view to PIRCs which fit template (3) above, with *is* as the finite auxiliary (or copula) in both clauses and subject relativization in the relative clause. But though this sentence type has received most attention in the POS literature, it is linguistically just one arbitrarily chosen instance of a much broader phenomenon. Auxiliary inversion in interrogatives can involve other auxiliaries such as “was”, or “must”, or the “do” of *do*-support. In multi-clause examples the clauses may differ in their auxiliaries (e.g., *Must the boy who was shouting go home?*), or one or both clauses may have no auxiliary (e.g., *Must the boy who shouted go home?*). There are also varieties of PIRC in which the relative pronoun is followed not by any auxiliary or verb, but by the subject of the relative clause, as in *Is the girl who the boy is talking to trying to run away?* in which it is the object of the relative clause that is relativized. Since the subject of the relative clause can be any well-formed noun phrase, with considerable freedom as to its first word, the bigram containing the relative pronoun will vary across examples (e.g., $\langle who\ the \rangle$, $\langle who\ Jim \rangle$, $\langle who\ every \rangle$) diluting the likelihood that the corpus will contain the bigram containing the relative pronoun. Note that this is so even for an *is-is* PIRC, if the gap in its relative clause is in some position other than the subject. In a more representative collection of PIRC test sentences, therefore, the bigrams in the grammatical version will be quite varied. Outcomes of the discrimination task can therefore be expected to be correspondingly more variable than for a uniform set of test items with $\langle who/that\ is \rangle$ in every one. It is conceivable of course that each subvariety of PIRC will have its own distinguishing bigram or combination of bigrams that play the role that the $\langle who/that\ is \rangle$ bigram plays in *is-is* subject-gap PIRCs. However, our anatomization of what a bigram model needs in order to succeed suggests that this is not so, and our empirical data confirm this.

5.1. Experiment 4: Object-gap relative clauses

The method was as in the previous experiments. 100 pairs of PIRC test sentences were constructed in which both clauses contained the auxiliary “is”. But this time the relativized noun phrase was the object of the relative clause (the direct object of the verb, as in (6) and (7) below, or its indirect object, or the object of a preposition).²⁴ The trace of the object is shown as t_j in the examples below, coindexed with the phonologically null relative pronoun \emptyset_j , which is coindexed with the head noun (here: *wagon*) that is modified by the relative clause. The trace of the fronted auxiliary in (6) and (7) is shown here as t_i , coindexed with the moved auxiliary *is*.

- (6) Is_i the wagon $_j$ [\emptyset_j your sister is pushing t_j] t_i red?
 (7) * Is_i the wagon $_j$ [\emptyset_j your sister t_i pushing t_j] is red?

The relative pronoun was phonologically null in all test sentences. An overt relative pronoun (“who” or “that”) could have been used, but it would have made no difference to the results because it would not have been included in any of the distinguishing bigrams for these object-gap relative clause constructions. Examples (6)/(7) are typical in this respect. They are identical from the sentence beginning until after the subject of the relative clause; the bigrams that distinguish them do not start until the word “sister”. Exactly the same would be true if they had

²⁴ The test sentences in R&C’s experiments all had subject-gap relative clauses, except for one item (derived from Crain & Nakayama’s 1987 child language study) in their Experiment 2, whose gap was the object of a postverbal preposition. It was correctly classified by the bigram model.

contained an overt relative pronoun at the position of \emptyset_j . So in contrast to the three previous experiments, the relative pronoun plays no part in discriminating between the grammatical and ungrammatical versions of object-gap PIRCs. In consequence, there is no distinguishing bigram which signals that it belongs to the relative clause, hence no recurrent bigram that conveys information about whether the auxiliary in the relative clause moved out or stayed in place.

The distinguishing bigrams for examples (6) and (7) are shown in Table 5, with their estimated probabilities.

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

	Bigram1	Bigram2	Bigram3
(6) Grammatical	< <i>sister is</i> > 0 + 718.41 = 718.41	< <i>is pushing</i> > 0 + 1.12 = 1.12	< <i>pushing red</i> > 0 + 16.84 = 16.84
(7) Ungrammatical	< <i>sister pushing</i> > 0 + 1.12 = 1.12	< <i>pushing is</i> > 0 + 718.41 = 718.41	< <i>is red</i> > 0 + 16.84 = 16.84

With respect to the smoothing factors, the patterning of the six bigrams in Table 5 is similar to the previous experiments: the second unigram of each bigram in the grammatical sentence matches one in the ungrammatical sentence, so the smoothing factor will always balance out across the two versions when bigrams are not attested in the corpus (as shown by the shading in the table), creating undecided situations. When the bigram model does make a choice, which version is preferred will depend on non-systematic facts concerning whether and how often each of the six bigrams occurs in the corpus. There is no distinguishing bigram here that systematically appears in most or all of the grammatical sentence versions. In bigram1-grammatical and bigram2-grammatical, the auxiliary “is” is flanked by lexical categories, which vary from test pair to test pair. Bigram3-grammatical, as usual, does little to assist the grammatical version. Thus, no bigram gives a consistent advantage to the grammatical sentence.

This profile of the bigrams involved in object-gap PIRCs predicts that for these test sentences there will be both correct and incorrect choices, as well as some failures to choose when all bigrams in a test pair are absent from the corpus. This is what was observed; see the results in the fourth line of Table 2 above: only 35% of test pairs were correctly distinguished. The pair (6)/(7), with no attested distinguishing bigrams (see Table 5), is one instance of a tie. The sentence *Is the dessert the kid is eating good?* is an instance of correct selection; its bigram2-grammatical (<*is eating*>) was in the corpus while the other five bigrams were not. Overall, as predicted, the bigram model does not perform reliably on object-gap PIRCs, for lack of a distinguishing bigram that is both informative and recurrent, to tip the scales toward the grammatical version.

5.2. Experiment 5: PIRCs with do-support

In this experiment we returned to subject-gap relative clauses, but used lexical main verbs in place of the “is” of the previous experiments. 100 pairs of test sentences with properties as illustrated in (8) and (9) were constructed.

- (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?

Both members of a pair contained a subject-gap relative clause beginning with “who” or “that”. (In 46 pairs the relative pronoun was “who”; in the remainder it was “that”.) In both members of a pair, each clause contained a lexical main verb which needed the support of an auxiliary *do* to create the interrogative form. In (8) and (9) we show traces of *do*-movement as *t_i*.²⁵ The distinguishing bigrams for sentences (8) and (9) are shown in Table 6. (There are four distinguishing bigrams in these test sentences for reasons given below.)

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

	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

Note that the verb in the clause with which the “do” is associated is non-finite, showing no number agreement with its subject (which was always singular in the test sentences). This is evident in the contrast between finite “plays” in (8) versus non-finite “play” in (9), and similarly for finite “wants” in (9) versus non-finite “want” in (8). For all our test sentences this difference in finiteness had an observable effect on the form of the verb. This is why more bigrams differ between the two versions of these *do*-support items than for the sentence types tested in Experiments 1-4 where fronting the “is” did not alter the form of the predicate that remained in place.

The fact that there are four distinct verb forms in these sentences (*play/want*; finite/nonfinite) also entails that not all of the distinguishing bigrams match up pairwise across the versions. The smoothing factor is identical in only two cases: for bigram2-grammatical and bigram2-ungrammatical, and for bigram4-grammatical and bigram4-ungrammatical. Hence, the estimated probabilities of the two sentence versions are bound to differ (coincidence aside) regardless of whether or not any of the distinguishing bigrams occur in the corpus. It is therefore predicted that in this experiment, unlike Experiments 2-4, there should be very few, if any, undecided cases.

²⁵ 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. Here, for expository convenience, we will presume the movement analysis; we believe that nothing relevant to bigram-based learning hangs on this assumption.

When the bigram model does make a choice, there is no basis for expecting either the grammatical or the ungrammatical version to prevail in this experiment. None of the distinguishing bigrams in these sentences is likely to recur in many test pairs, since they all contain a content word. This is true even for bigrams containing the relative pronoun. Therefore the bigram model has no particular bigram(s) that it can count on to favor the grammatical version.²⁶

The results, shown in the fifth line of Table 2 above, conform with these expectations: discrimination is at chance and there are no ties. The test pair (8)/(9) was among those that were incorrectly judged, because the product of the smoothed bigram probabilities (see Table 6) happens to be higher in the ungrammatical version than in the grammatical one. A grammatical sentence that was correctly selected is *Does the man who goes to the beach need sandals?*

Experiments 4 and 5 targeted PIRC varieties that were selected in order to test the validity of our hypothesis about the kinds of bigrams that the bigram language model thrives on. For a variety of syntactic reasons, the grammatical versions of PIRCs with object-gaps and PIRCs with *do*-support do not contain any adjacent pair of recurrent words that can serve as a marker for the grammatical version, like the *<who/that is>* bigram in the original *is-is* subject-gap test sentences. The negative results of these experiments thus support our conjecture that the positive results for the *is-is* PIRCs do not reflect any general grasp of linguistic constraints on subject-auxiliary inversion. As soon as the test sentences are allowed to reflect a range of variation more typical of the English language, the bigram model loses its edge. This has obvious bearing on whether statistical learning has been shown to compensate for the purported poverty of the stimulus for child syntax acquisition. Chomsky's POS thesis pertains equally to every construction in every learnable natural language; but we have found that even within a single language, it is only in a small proportion of cases that bigram-based learning is able to tap indirect evidence in the corpus to substitute for the lack of direct exemplars. Thus bigram-based modeling leaves stimulus poverty as an open issue. There are two possibilities: that the properties of PIRCs are not in general derivable from indirect evidence in a corpus of sentences as word strings; or that the information is present in the corpus but these computations over bigrams are not powerful enough to extract it. We come back to this in the general discussion, after reporting one final experiment in which we turned to another language in order to examine a subvariety of PIRC that does not occur in English.

5.3. Experiment 6: Dutch PIRCs with lexical verb fronting

In some languages, the inversion process that occurs in English questions is more general: it may apply to all finite verbs, not just auxiliaries, and it may (or must) apply in declarative sentences as well as questions. This is the case in Germanic languages, including Dutch among others; see example (10) below. We are interested in determining the extent to which a bigram model is capable of extracting general patterns of sentence formation from a corpus. Testing it on

²⁶ It could be expected that bigram1-grammatical would be more frequent in the corpus than bigram1-ungrammatical, since the verb "play" in the latter is non-finite, which is illicit following the relative pronoun which is its subject. However, this is another place where homography is relevant. The non-finite verb is morphologically (orthographically) indistinguishable from a plural finite verb, so the word sequence *<who play>* could indeed occur in the corpus. For 16 of our 100 test pairs, bigram1-ungrammatical did occur in the corpus as a plural.

Dutch, with its general pattern of verb inversion, can be informative in this regard. This experiment also addresses the question of whether the bigram model's failures in the previous experiments are in some way peculiar to English. That would still be seriously troublesome for the hypothesis that there is no stimulus poverty for bigram-based acquisition of PIRCs; but if *only* English lacked indirect cues to PIRC structure, some excuse for it might perhaps be found. This would evidently be more difficult if the bigram model failed on PIRCs in other languages too. On the other hand, for establishing that bigrams provide a general basis for the learnability argument against POS, it would suffice to show that there are different cues in different languages, so that the bigram model has some basis for learning how to form complex questions in any language even if the specific bigrams in those questions differ radically from one language to another. To test 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 to contribute to the bigram model's task of discriminating grammatical from ungrammatical PIRCs.

Dutch has a PIRC construction that is similar to English, except that no Dutch equivalent of *do*-support is needed because lexical verbs can be fronted.²⁷ The corpus used in this experiment is known as the Groningen corpus (Bol, 1996), which is available in the CHILDES database. It is a record of spontaneous conversation between adults and seven Dutch children in informal home settings similar to those of the Bernstein-Ratner English corpus. The children were from 1;05 to 3;07 so it was not possible to match the ages of the children exactly to those in the Bernstein-Ratner corpus, but we chose from among the earliest files in the corpus, covering a 4-month period for each child, from 20 to 23 months. This yielded 21,557 utterances of adult child-directed speech. The resulting corpus was thus both larger and somewhat 'older' than the corpus of English child-directed speech, but this would tend to increase the chances of successful learning by the bigram model (see section 6).

40 pairs of Dutch PIRCs were tested. These were constructed, with the assistance of a native speaker, as for the previous experiments except that we followed the constraints of Dutch syntax, e.g., word order was SVO in main clauses and SOV in embedded clauses, and lexical verbs were fronted in questions without auxiliaries. Dutch has two relative pronouns: "die", which is more frequent, is used when the noun head is 'common gender' or plural; "dat" is used when the head is a singular neuter noun.²⁸ In 35 of our test sentence pairs the relative pronoun was "die"; in 5 it was "dat". Sentences (10) and (11) are typical of the test pairs. For clarity, we have inserted brackets around the relative clause, and have indicated the trace (the underlying position) of the fronted verb.

- (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?

²⁷ Dutch does have *do*-support but it applies primarily in the context of VP-preposing. See van Kampen (1997) for linguistic references and discussion of *do*-support in the acquisition of Dutch.

²⁸ There is homography in Dutch not unlike that in English. Dutch has "die" and "dat" not only as relative pronouns but also as demonstrative determiners (e.g., *Die auto is mooi*, 'That car is beautiful') and as demonstrative pronouns (e.g., *Dat is een mooie auto*, 'That is a beautiful car'); "dat" can also be a complementizer (e.g. *Ik weet dat jij mij leuk vindt*, 'I know that you like me').

‘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?’

In all the test sentences, as in these examples, both clauses contained a lexical verb only (no auxiliary). In half of the test pairs the main clause had a transitive verb with its object and the relative clause had an intransitive verb with an adjunct or secondary predicate. In the other pairs the main clause had an intransitive verb with an adjunct or secondary predicate and the relative clause had a transitive verb with its object. It was not feasible to test the Dutch equivalent of *is-is* PIRCs, because they are structurally ambiguous (in written form) in a way that would make it impossible for any learning algorithm to distinguish the grammatical and ungrammatical versions. Prior to question formation, the first “is” would be at the end of the relative clause that modifies the subject, and the second “is” would immediately follow the subject (i.e., would follow the relative clause). So the two instances of “is” would be adjacent, and it would be unclear which of them had then been fronted to form the interrogative. The word string would be the same in both versions, even though the structural position of the trace would differ between them. For example, corresponding to the declarative (12), both the grammatical and the ungrammatical PIRCs would be transcribed as: *Is de jongen die in de kamer is roodharig?*.

- (12) [*De jongen [die in de kamer is] is roodharig*]
The boy who in the room is is red-haired
‘The boy who is in the room is red-haired.’

In spoken Dutch, such as the children were exposed to, the two structures for this interrogative word string would almost certainly be disambiguated by a prosodic break at the end of the relative clause, which would reveal which verb had moved and which had stayed in place. In future work it would be very interesting to employ a training corpus with prosodic boundaries annotated. For present purposes, however, we disambiguated the grammatical and ungrammatical versions by using test sentences in which the relative and matrix clauses contained lexical verbs that differed in argument structure, as in (10)/(11). Note that (10) is coherent only if the fronted verb “wil” (‘wants’) comes from the matrix clause, while in (11) the fronted verb “zit” (‘sits’) can only be construed as having moved (improperly) from the relative clause. (Otherwise, i.e., on the contrary analyses with argument structure violations, these sentences would have the incoherent interpretation: ‘Is the baby that wants on the new chair sitting a cookie?’)

The distinguishing bigrams for (10) and (11) are shown in Table 7. There are eight distinguishing bigrams here. Because the verbs of the two clauses in a sentence differ, the grammatical and ungrammatical sentence versions differ not only at the end of the relative clause but also at the beginning of the sentence where the fronted verb occurs. So in this experiment the initial sentence marker (-sent-) must be included in the analysis; it is the first unigram in a bigram in which it is followed by the fronted verb (here: “wil” or “zit”).

Table 7

Smoothed probabilities (x 100,000) for the distinguishing bigrams in sentences (10) and (11).

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

All of these distinguishing bigrams contain a lexical category, as in the English *do*-support examples. (Note that “wil” in these examples is a main verb, not an auxiliary.) Also, like the English object-gap examples, the relative pronoun is not adjacent to the critical verb in the relative clause, since the arguments and adjuncts of the verb normally intervene in Dutch between the relative pronoun at the beginning of clause and the verb at the end of it.²⁹ This has the further consequence that the relative pronoun does not appear in the distinguishing bigrams, since it is flanked in both versions by the same words (the noun that is modified, and the first word of the argument or adjunct of the verb). Thus, by the standards that have emerged from the preceding experiments, this array of distinguishing bigrams does not look promising for the bigram model. However, a main goal of the exercise is to see whether the Dutch sentences contain other useful cues, which are not anticipated by our analysis above. Even if Dutch has nothing as robust as the <*who/that is*> bigram in English *is-is* PIRCs, there might perhaps be some confluence of minor cues, each only weakly predicting but reinforcing each other. Not knowing in advance what these cues might be, the research strategy is to ascertain whether bigram-based learning is successful. If not, it can be concluded that such cues are not available; while if it is, an effort can be initiated to identify them.

The results, shown in the sixth line of Table 2 above, show that the bigram model does *not* do well on these Dutch PIRCs. It chose the grammatical version in only 32.5% of test pairs. (For example, it chose correctly between (10) and (11), but chose the ungrammatical **Lijkt de kok die moe maakt een cake?* over the grammatical *Maakt de kok die moe lijkt een cake?* ‘Is the cook who seems tired making a cake?’.) This poor performance makes it clear not only that the bigrams as illustrated in Table 7 are indeed ineffective, as anticipated, but also that these sentences do not offer the bigram model any *other* indicator of grammaticality. An additional finding is that there were only a few ties (12.5%) between the grammatical and ungrammatical versions in this experiment. The second unigrams of the distinguishing bigrams match exactly across the grammatical and ungrammatical versions, as can be seen in the example in Table 7, so the smoothing terms are identical and they would all balance out if none of the distinguishing bigrams occurs in the corpus. (This differs from the English *do*-support examples, as discussed

²⁹ Subsequent to Experiment 6 we re-ran the Dutch materials omitting the adjuncts in the intransitive relative clauses, so that the relative pronoun and the clause-final verb were in most cases adjacent, and together constituted a distinguishing bigram. Some sentence pairs were judged differently than in Experiment 6, but the overall success rate was exactly the same as in Experiment 6.

above.) This suggests, contrary to fact, that there would be a fairly high proportion of “undecided” responses in this experiment. However, examination of the Dutch test items shows that their distinguishing bigrams were well-represented in the corpus: 29% of them occurred in the corpus, compared with only 19% in the English *do*-support experiment. The reasons for this are not difficult to discern. The Dutch corpus was more than twice the size of the English one. Also, the fact that Dutch word order principles allow verbs to precede or follow their objects or adjuncts, and to precede or follow their subjects even in declarative sentences, means that Dutch is much richer than English with respect to bigrams that relate a verb with a determiner or noun or adverb. Hence, more bigrams in the Dutch test sentences are attested, and fewer sentence choices rest solely on the smoothing factors.

In short: the Dutch results, both the lack of preference for the grammatical version, and the low proportion of “undecided” responses, are in accord with our general analysis of the bigram model’s capabilities and limitations. It is also quite telling that the model exhibited no noticeable increase in accuracy with the shift to the larger Dutch corpus. A larger corpus provides a greater opportunity for the model to pick up statistical trends even if they are quite subtle. From the fact that it did not do so, it may fairly be concluded that there are no useful cues to grammaticality in the bigram composition of Dutch PIRCs. Hence Dutch PIRCs join all except the *is-is* subject-gap PIRCs in English as candidates for POS status for a learner with only the limited resources of bigram statistics. Chomsky’s argument that innate linguistic knowledge (UG) is needed to supplement input information in the acquisition of PIRCs thus remains essentially untouched by the bigram-based learnability approach.

6. General discussion

A demonstration that children’s primary linguistic data affords information determining the correct form of one complex syntactic construction does not imply that the same will be true for every complex syntactic construction; so it cannot by itself falsify the POS thesis. Conversely, a demonstration like ours, that for some syntactic constructions a bigram language 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 this debate does not settle the substantive issue of whether the input for syntax acquisition is rich or poor. Nevertheless, some general conclusions can be drawn: conclusions about methodology, about prospects for future research along these lines, and about what role UG might still play.

Our methodological conclusion is that it should be standard practice for data-driven learning claims to be accompanied by an elucidation of the source of their abilities, particularly when the goal is to shed light on human language learning. The POS thesis is, after all, a thesis about first language acquisition by children. Its central importance to linguistics and psycholinguistics, and the reason it is still vigorously debated after all these years, is that it has strong implications for the mechanisms of human language acquisition. In section 1 we noted that a rebuttal of POS based on statistical learning capabilities can make its point even in purely ‘black box’ mode, i.e., even if it is unknown what input information the learning system is picking up on. However, a learnability result is more revealing if the black box is opened up, to provide a glimpse of the epistemic relation between what the learner ends up knowing and where that knowledge is coming from among the observable facts of the corpus. The need for this is obviously especially

acute in the case of learning from indirect evidence, where the knowledge does not come from explicit examples of that construction.

In the present case, once we established which bigrams in the *is-is* test sentences were responsible for the model's bias toward the grammatical version, it was easy to see that this bias could not extend to other instances of (what is arguably³⁰) the very same linguistic generalization. So it became clear that the success for the *is-is* variety is in some sense a fluke of the surface lexical properties of the particular target sentences: the grammatical version happens to contain a pair of highly frequent (closed-class) words, occurring adjacent to each other at precisely the locus in the grammatical word string that would have been disrupted if auxiliary inversion had been applied incorrectly. It follows from the very nature of bigram-based learning that this is the ideal profile, the surest route to a strong bias in the direction of the grammatical form. So the impressively positive results of R&C's experiments are understandable.

Once uncovered, this characteristic of the *is-is* PIRCs could be recognized as the product of the criterion by which these PIRCs were separated off as the particular subclass to be studied. The *<who/that is>* bigram was part of the template defining the target sentences; but it is not inherent to question formation, either in general or even just within English. Therefore the finding of input richness for *is-is* makes no dent in the argument for input poverty for all the other kinds of PIRCs that exist in natural languages. At best, the results for *is-is* might offer some mild encouragement for the belief that every construction, when studied, will prove to have its *own* characteristic statistical hallmark. As it happens, our experiments show that this is not so. But suppose for the moment that it were true. It would imply that any learning system relying solely on bigrams would acquire a language in small slivers. Broad generalizations would go undetected wherever subcases of a pattern differ in their local surface details, as some PIRCs differ from others. Our PIRC results, when looked into more closely, thus present a clear working example of the close connection between learning from superficial word combinations, and learning only small-scope subgeneralizations.

Our second general conclusion is that if *is-is* learning is set aside, because of its demonstrably narrow compass, the learnability rebuttal of POS, whose significant potential we discussed in section 1, remains unsubstantiated at present. To the best of our knowledge, there has been no demonstration of the learnability of PIRCs in general, or of any other complex syntactic construction, from real-life input to children by a UG-free learning model that clearly does not overstep the computational resources of a normal preschool child. This remains open as a challenge for the statistical learning research community. A growing number of studies are converging on this goal, but so far none meets all these criteria.³¹ Even the neural network

³⁰ That there is a generalization to be captured about auxiliary inversion is recognized even in a framework such as Construction Grammar which does not emphasize broad cross-constructional principles. Adele Goldberg (p.c.; see also Goldberg & del Giudice, 2005) suggests that auxiliary inversions in different contexts can be regarded as a unitary phenomenon if they invariably co-occur in natural languages. This criterion is perhaps not satisfied by inversion in questions and inversion in (for example) counterfactuals (*Were she here, all would be well*); English retains the former but may be losing the latter. But despite a lack of sufficient data we conjecture that it *is* satisfied by PIRCs with a subject-gap relative and PIRCs with an object-gap relative (for languages that permit gaps of both kinds). We know of no data, however, which indicate whether children treat these as related.

³¹ The realistic psychological resources condition excludes interesting work by Clark and Eyraud (2006) and Chater and Vitanyi (in press). It may or may not exclude Perfors et al. (2006) and neural network results such as those of

studies of PIRCs, by R&C and Lewis & Elman (2001), have so far tested only *is-is* PIRCs with subject-gap relatives, which we have shown are not indicative of broader success. So it has not yet been demonstrated that the networks' ability to discriminate grammatical and ungrammatical PIRCs generalizes to the full range of examples relevant to disproving stimulus poverty for PIRCs.

One way for future learnability research to go about meeting this challenge would be to shift up to a larger or 'older' corpus. A larger corpus would supply more accurate statistics and place less reliance on the smoothing technique to fill in missing data. This might permit a bigram model to select the grammatical form of PIRC types with object gaps or *do*-support or main-verb fronting. However, some caution may be in order until this has been demonstrated, because the explication of Experiments 4 - 6 showed that the difficulty in learning these PIRC varieties (as opposed to those of Experiments 2 and 3) was not low corpus frequency of bigrams that would have been effective if attested. Rather, it was the fact that the *target sentences* contain no distinguishing bigram that would systematically bias toward the grammatical form. (Empirical results support this: recent research shows that even increasing the size of the corpus tenfold yields only modest improvements in performance, to approximately 70%; see Kam, 2007.) Performance would be enhanced by shifting to a larger, richer or more representative corpus only if the other PIRC types prove to be discriminable by aggregating a multitude of minor regularities in the corpus.

A corpus of adult speech to older children might contain more indirect evidence about PIRCs than speech to children under 2 years old as in the Bernstein-Ratner corpus. Any corpus is fair game for demonstrating learning from indirect evidence as long as it contains no explicit instances of the construction being tested for, and precedes the age at which children have acquired that construction. R&C's project was very ambitious in using such an 'early' corpus, since there is no evidence from child studies that even *is-is* PIRCs are within the competence of children before the age of 3;2 (the youngest child in Crain & Nakayama's 1987 experiment). Of course children younger than that may have the relevant linguistic knowledge even though testing them in such a way as to reveal it may not be feasible. Hence, improving the bigram model's performance by shifting to an 'older' corpus has the practical drawback that it could become re-entangled with traditional POS disagreements about establishing ages of exposure and ages of mastery. In our latest work we have raised the age in corpus data to 8 years, finding some improvement on object-gap and *do*-support PIRCs, but still no more than 70% correct (see Kam, 2007, for details). Corpus data for children older than this are hard to come by, and whether they are necessary is difficult to know, in view of the complete absence of data in the child language literature on when children control auxiliary inversion in fully general form across all relevant contexts.

Another natural step for improving performance would be to move up from the level of individual words to that of lexical categories such as Noun and Verb. An analysis similar to the bigram analysis could be conducted over sentences that have been coded into strings of such categories by part-of-speech tagging. This has two potential advantages. It could increase the chance of capturing true linguistic generalizations, which are not formulated in terms of

Elman (1993), Lewis and Elman (2001) and Frank et al. (ms.); but the latter did not employ real-life corpora of child-directed speech.

particular word sequences but in terms of groupings of more abstract syntactic categories. The second advantage is that it increases the corpus counts of the ‘content’ words, so that they can begin to take on some of the load of discriminating grammatical from ungrammatical sentences, which previously relied primarily on functional categories as in the <who/that is> bigram. (There is a trade-off, of course, between this advantage and the loss of precision due to aggregating words into categories, as noted in footnote 23 above.) For the sentences in (4)/(5), for instance, whereas the corpus may contain neither <is crying> nor <crying is>, it is very likely to contain <v:aux&3S part-PROG> and in greater numbers than <part-PROG v:aux&3S>. In other experiments we have quantified the advantage this confers on the troublesome subtypes of English PIRCs with object-gaps and *do*-support. Performance improved, though it still did not exceed 70% correct, and some of the improvement was in reducing the number of undecided test pairs rather than in increasing the proportion of correct decisions, which actually declined for object-gap PIRCs; see Kam (2007) for details. Chang et al. (2006) also found a decline in performance for category strings as opposed to word strings, for word order production with a different (enriched) bigram-based model.

Finally, an obvious next move for improving performance would be to strengthen the power of the learning algorithm. We focused here on the bigram language model because if it had succeeded, the evidence of the richness of the stimulus would have been compelling. The fact that it was able to acquire only a very restricted subset of PIRCs means that there is more work to be done to evaluate the richness hypothesis. As we have noted, a trigram language model and neural network models have been applied to the task of PIRC learning, and these are more powerful than a bigram model. So far they have been put to the test only on *is-is* subject-gap PIRCs, which we have shown can be discriminated on the basis of a simple local cue, unlike the other equally relevant varieties of PIRC that we have examined. Since it does not imply knowledge of auxiliary-inversion in general, success on *is-is* subject-gap PIRCs is a very weak measure of acquisition. However, it is not out of the question that in future experiments such models will be shown to be capable of acquiring the full range of PIRCs (without overgeneralizing in other respects; see section 4.1).

Our third general conclusion concerns the possible role of innate linguistic knowledge (UG), which has been the hostage to fortune in the stimulus poverty/richness debate since its inception. The point we make here is necessarily hypothetical, since the need for UG to assist language acquisition is more or less complementary to the power of UG-free data-driven learning, and we have just observed that the latter will not be known until more research has been done. But it is instructive to consider what the implications would be if future research with more sophisticated statistical mechanisms were to confirm the mixed pattern of success observed in our experiments, where some varieties of a construction succumbed readily to a statistical learning algorithm while others were highly 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.

Specifically: corpus statistics could help a learner guess that some word strings are grammatical and some are ungrammatical, even if neither have actually occurred in the child’s input so far. A non-occurring string like *Is the little boy who is crying hurt?* would ‘sound good’

on the basis of bigram data, while a non-occurring string like *Is the little boy who crying is hurt?* would ‘sound bad’. Judgments such as this – though only a surmise based on bigram data – might then feed into the child’s formulation of the rule (or parameter setting) for auxiliary inversion, just as informants’ grammaticality judgments are used by linguists as a basis for uncovering the grammar of a language. Once the child has established, on these grounds, a rule about which auxiliary verb is fronted in an *is-is* PIRC, that rule may have broader applicability, predicting which auxiliary is fronted in an object-gap PIRC, for example, even if the child has never heard one.

There is no proof that this is so, but it does offer a plausible and positive role for the kinds of information that a simple word-based statistical learner could pick up (though committed linguistic nativists would deny that rejection of ungrammatical PIRCs requires any input at all). However, what *cannot* be the case is that a child gathers bigram data from the corpus and then continues to rely on it indefinitely to guide question formation *instead* of formulating a grammar rule. A learning system which did that would make egregious errors on object-gap and *do*-support PIRCs, for which bigram statistics do not point the learner reliably toward the grammatical version. To state it more broadly: Even learning systems that are equipped to track corpus statistics must derive rules, if the statistics predict the correct form of only some but not all sentence types in the language.

If something like this is correct, another point may then follow. If a child has to deduce the correct form of a *do*-support object-gap PIRC on the basis of a general rule for auxiliary inversion acquired from *is-is* subject-gap PIRCs, then that child must (a) grasp that in some relevant sense these qualify as the same construction, and (b) know how to establish structural parallels between the two so that the rule devised for one of them *can* be applied to the other. This constitutes quite sophisticated abstract knowledge, since the two forms may not be any more alike superficially than forms which do *not* qualify as the same construction. But it is not clear that this abstract knowledge could itself have been acquired from experience. On these assumptions, then, although learners may start by referencing simple probabilistic dependencies between words, it appears that they must at some point make the transition to general grammatical rules, and that something very like what linguists mean by Universal Grammar may be needed to guide this transition.

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 and Morten Christiansen for their comments and suggestions. We also received helpful feedback from audiences at the 18th Annual CUNY Conference on Human Sentence Processing, March 2005, and the Workshop on Psychocomputational Models of Human Language Acquisition at ACL-2005. This work began as a student research project by the first three authors, under the supervision of the last two authors. It will be presented in greater detail and extended in Kam’s Ph.D. dissertation, in progress. A shorter version of the present paper will be published shortly in *Cognitive Science* (Kam et al., in press). This research

was supported in part by grants 65398-00-34, 66443-00-35 and 66680-00-35 from the Professional Staff Congress of the City University of New York.

References

- Akhtar, N., Callanan, M., Pullum, G. K., & Scholz, B. C. (2004). Learning antecedents for anaphoric *one*. *Cognition*, *93*, 141–145.
- Ambridge, B., Rowland, C. F., & Pine, J. M. (in press). Is structure dependence an innate constraint? New experimental evidence from children's complex-question production. *Cognitive Science*.
- Baker, C. L. (1978). *Introduction to generative transformational syntax*. Englewood Cliffs, NJ: Prentice Hall.
- Bernstein-Ratner, N. (1984). Patterns of vowel modification in motherese. *Journal of Child Language*, *11*, 557–578.
- Chang, F., Lieven, E., & Tomasello, M. (2006). Using child utterances to evaluate syntax acquisition algorithms. In *Proceedings of the 28th annual conference of the cognitive science society*. Vancouver, Canada.
- Chater, N., & Vitanyi, P. (in press). 'Ideal learning' of natural language: positive results about learning from positive evidence. *Journal of mathematical psychology*.
- Chomsky, N. (1980). *Rules and representations*. Oxford: Basil Blackwell.
- Clark, A., & Eyraud, R. (2006). Learning auxiliary fronting with grammatical inference. Presented at the *tenth conference on computational natural language learning*. New York.
- Crain, S., & Nakayama, M. (1987). Structure dependence in grammar formation. *Language*, *63* (3), 522–543.
- Crain, S., & Thornton, R. (1998). *Investigations in universal grammar: a guide to experiments in the acquisition of syntax and semantics*. Cambridge, MA.: MIT Press.
- Elman, J. L. (1993). Learning and development in neural networks: the importance of starting small. *Cognition*, *48*, 71–99.
- Fodor, J.D., & Crowther, C. (2002). Understanding stimulus poverty arguments. *The linguistic review*, *19*, 105–145.
- Frank, R., Mathis, D., & Badecker, W. (unpublished ms., 2004). The acquisition of anaphora by simple recurrent networks.
- Goldberg, A. E., & Del Giudice, A. (2005). Subject auxiliary inversion: a natural category. *Linguistics review*, *24*, 411–428.
- Jurafsky, D., & Martin, J. H. (2000). *Speech and language processing*. Upper Saddle River, NJ: Prentice Hall.
- 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 Press.
- Kam, X.-N. C., Stoynezhka, I., Tornyoova, L., Fodor, J. D., & Sakas, W. G. (in press). Bigrams and the richness of the stimulus. *Cognitive science*.
- 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 Press.
- Lidz, J., Waxman, S., & Freedman, J. (2003). What infants know about syntax but couldn't have

- learned: experimental evidence for syntactic structure at 18 months, *Cognition*, 89, B65–B73.
- Lidz, J., & Waxman, S. (2004). Reaffirming the poverty of the stimulus argument: a reply to the replies. *Cognition*, 93, 157–165.
- MacWhinney, B. (2000). *The CHILDES-Project* (3rd edition). Volume 2: *Tools for analyzing talk: the database*. Hillsdale, NJ: Erlbaum.
- Pereira, F. (2000). Formal grammar and information theory: Together again? *Philosophical transactions of the royal society*, 358, 1239–1253.
- Perfors, A., Tenenbaum, J., & Regier, T. (2006). Poverty of the stimulus? a rational approach. *28th annual conference of the cognitive science society*. Vancouver, Canada.
- Pullum, G. K., & Scholz, B. C. (2002). Empirical assessment of stimulus poverty arguments. *The linguistic review*, 19, 9–50.
- Real, 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*.
- Real, F., & Christiansen, M. H. (2005). Uncovering the richness of the stimulus: structure dependence and indirect statistical evidence. *Cognitive Science*, 29, 1007–1028.
- Regier, T., & Gahl, S. (2004). Learning the unlearnable: the role of missing evidence. *Cognition*, 93, 147–155.
- 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.
- Saffran, J. R., & Wilson, D. P. (2003). From syllables to syntax: Multi-level statistical learning by 12-month-old infants. *Infancy*, 4, 273–284.
- Sampson, G. (1997). *Educating eve: the 'language instinct' debate*. New York: Cassell.
- Sampson, G. (2002). Exploring the richness of the stimulus. *The linguistic review*, 19, 73–104.
- Tomasello, M. (2004). Syntax or semantics? Response to Lidz et al. *Cognition*, 93, 139–140.
- Tomasello, M., & Stahl, D. (2004). Sampling children's spontaneous speech: How much is enough? *Journal of child language*, 31, 101–121.
- Vries, M. de (2002). *The syntax of relativization*. Ph.D dissertation. University of Amsterdam.
- Wexler, K. (1999). Maturation and growth of grammar. In W. C. Ritchie, & T. K. Bhatia (Eds.), *Handbook of child language acquisition*. San Diego, CA: Academic Press.