CONTRIBUTIONS OF STATISTICAL INDUCTION

TO MODELS OF SYNTAX ACQUISITION

By

XUÂN-NGA KAM

A dissertation submitted to the Graduate Faculty in Linguistics in partial fulfillment of the requirements for the degree of Doctor of Philosophy, The City University of New York.

2009

This manuscript has been read and accepted for the
Graduate Faculty in Linguistics in satisfaction of the
dissertation requirement for the degree of Doctor of Philosophy.

Janet Dean Fodor

Date — Chair of Examining Committee

Gita Martohardjono

Date — Executive Officer

Janet Dean Fodor

Martin Chodorow

William Gregory Sakas
Supervisory Committee

THE CITY UNIVERSITY OF NEW YORK

Abstract

CONTRIBUTIONS OF STATISTICAL INDUCTION

TO MODELS OF SYNTAX ACQUISITION

By

XUÂN-NGA KAM

Adviser: Professor Janet Dean Fodor

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, with the deeper implication that innate knowledge is not needed for grammar acquisition. Reali & Christiansen (2005) demonstrated that a simple bigram model trained on child-directed speech can induce the correct form of auxiliary inversion in certain complex English questions (e.g., *Is the boy who is crying hurt?*). The significance of this achievement is called into question, however, by Experiments 1– 6 reported here, which show that the success is highly circumscribed, resting on one particular bigram (<*who is*> or <*that is*>) in the grammatical test sentences. The model performs poorly on inversion in related constructions in English and Dutch, which do not afford effective cues accessible to a bigram analysis.

Performance improved modestly when learning resources were added in Experiments 7–15: the learning algorithm was upgraded to a trigram model, corpus size was increased, part-of-speech information was provided. Even so, there were no circumstances in which auxiliary inversion was well-discriminated across other variants (with *do*-support, with object-gap relatives). This suggests that the *n*-gram models were not capturing the linguistic generalization that unites the various instances of auxiliary inversion.

This weak performance is unsurprising, since the *n*-gram learners had no access to information about phrase-structure. Chomsky (1980) emphasized the significance of 'structure dependence' for correct application of the auxiliary-inversion rule. Experiments 16–18 provided some partial phrase-structure information relevant to the task. When noun phrases in the corpus and test sentences were surrounded by NP brackets, performance was extremely poor. But replacing each (maximal) noun phrase by the symbol NP finally yielded success across all three sub-cases of auxiliary inversion tested.

Consequently, based on the results to date, the *n*-gram challenge to stimulus poverty and UG remains unsubstantiated. However, if it can be shown in future work that an *n*-gram model is capable of assigning phrase-structure to word-strings, there are grounds for anticipating that it could succeed in extracting the general pattern of auxiliary-inversion.

Acknowledgements

Though this dissertation was my work for many years, I could have never undertaken and finished it without the support of many people. First, I am indebted to Janet Dean Fodor for her tremendous help throughout this process. This research benefited so much from her knowledge and critical thinking. Professor Fodor is an inspirational person by her dedication, intelligence and patience. It was an honor to work with her. I also would like to thank my other two committee members, William Sakas and Martin Chodorow who always gave me good feedback and advice and were always available to discuss any issues that I might have. I am grateful to Marcel den Dikken for his expertise in Dutch and his prompt replies to my questions and to my two colleagues, Iglika Stoyneshka and Lidiya Tornyova who contributed to parts of this research. The linguistics faculty at the Graduate Center is composed of amazing individuals who broadened my horizon in linguistics and made me a better scholar. Further afield, my thanks also go to Florencia Reali, Morten Christiansen, Elissa Newport and Ben Ambridge for their comments and suggestions.

I am grateful to Professors Gita Martohardjono, Elaine Klein, Ricardo Otheguy, Bob Vago, Margaret Wade-Lewis, James Hala and George Otte for giving me the opportunity to teach a subject that I love and to work in their research labs to expand my skills. I am also indebted to Dr. Claude Vogel who trusted and valued my work when I was a novice in the computational industry.

My fellow students at the Graduate Center (in particular, Rocio, Amit, Ylana, Iglika, Lidiya, Ronit, Yukiko, Stephanie, Leigh, Rebecca, Agustina, Michele, Ingrid and Tomonori) need recognition because they understand better than anybody my frustrations, my joys and my resilience to finish this dissertation. We could empathize with each other and it provided so much comfort so thank you all!

My other friends in New York and around the world (a special thought for the 'gang' in Paris) cannot be forgotten. Thank you for checking up on me, encouraging me, and keeping me grounded. Your friendship was inestimable in tough times.

Finally, I would like to dedicate this work to some people who are dear to my heart. First, my parents and my brother for loving me and standing behind me in all circumstances. Thank you, mom, for giving me the opportunity of a better education and life. I am also grateful to Olivier for his unconditional love. My days are filled with happiness thanks to you. My last thank is for Claude Kam. You are an example of kindness and generosity and I am proud to be part of your life.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction and literature review

1.1. Language acquisition: divergent approaches

How children acquire language is a central topic of research in psycholinguistics. And yet over the years, the answer to that question has divided the field. Two prominent perspectives on the topic have long confronted each other: the innatist versus the data-driven position.

Chomsky (1957 and since) has argued that language is acquired on the basis of innate linguistic knowledge, often known as *universal grammar* (UG). This has been the dominant approach in modern linguistics for many years. One of Chomsky's arguments for innate pre-tuning for language is called the *argument from the poverty of the stimulus*. The claim is that for some complex grammatical constructions in the adult language, the primary linguistic data available to learners do not contain any overt evidence about the generative rule or principle that licenses that construction. However, children do at some point acquire knowledge of these constructions. Chomsky argued that innate linguistic knowledge must fill in the information that is missing from the child's data.

Exactly what innate information is supplied depends on the linguistic theory that is assumed. It differs, for example, between transformational grammar (Chomsky, 1957 et seq.) and lexical-functional grammar (LFG; Kaplan & Bresnan, 1982); within transformational grammar, there has also been an evolution from a rule-based model (Chomsky, 1965) to a Principles and Parameters (P&P; Chomsky, 1981) model, and from there to the Minimalist Program (Chomsky, 1995).

The theory of innate linguistic knowledge (henceforth UG theory) gained many supporters and its influence has been considerable. However, as with many dominant paradigms, it also drew criticism: in particular some researchers have challenged the existence of innate linguistic knowledge as an essential foundation for language acquisition and have argued that grammatical generalizations can be induced from readily accessible language observations.

Hence a different approach to language acquisition has emerged, with increasing intensity in recent years, whose main focus is data-driven learning: how the child can piece together a grammar based solely on observable facts available in the child's linguistic environment. Despite recognizing the mental reality of grammar, this harks back in some ways to behaviorist approaches to language in the first half of the 20th century. This framework has been propelled recently by the increasing number and range of electronic language corpora, especially corpora of child-directed speech, available to the research community. Furthermore, more powerful computational resources facilitate the modeling of complex language learning processes. The current research will assess the potential viability of data-driven learning. Therefore I will now briefly review the major studies relevant to the discussion, first considering experimental psychological data and then the computational modeling results.

In the psychological study of child language development, for many years the emphasis was on the extent to which UG was available to children as evidenced in early mastery with few errors in such phenomena as word-order patterns, null-subjects, long

distance movement, binding and so forth. Recently there has been a move towards a "usage-based" approach to language acquisition (Tomasello, 2003, among others). The proposal is that children store the utterances that they hear and from these they gradually build up ("piece-meal learning") a knowledge of the constructions of the language.

The development of this theory of natural language acquisition has proceeded in parallel with an innovative series of psychological studies of the acquisition of artificial language materials, for which innate knowledge is presumably unavailable. By careful design of the 'language', the experimenters can fully control the information available to learners for acquiring the language facts. Unlike naturalistic studies of child language development, this makes it possible to compare very precisely what a child succeeded in learning against exactly the information afforded by the input sample. An influential early work of this kind was the study by Saffran et al. (1996) which employed a constructed language whose elements were words (there was no structure at the syntactic level). The relationships between successive syllables were controlled such that the transitional probability between one syllable and the next was high in some cases and low in others. The former were regarded as comparable to within-word sequences, and the latter to between-word divisions. After exposure to the training sample, the children were presented with some syllable sequences which did constitute words in the artificial language and some sequences which had occurred in the input sample but did not constitute words. The result was that the children listened significantly longer to the non-word sequences, indicating that they had been able to track the transitional probabilities between syllables, of a kind that would facilitate word boundary recognition in a natural

language. This finding was particularly dramatic because of the young age of the subjects (8-months old) and the brevity of the exposure to the training sample (2 minutes). The authors concluded that “Our results raise the intriguing possibility that infants possess experience-dependent mechanisms that may be powerful enough to support not only word segmentation but also the acquisition of other aspects of language.” (p. 4)

Several studies ensued which extended this approach to the acquisition of syntactic patterns in artificial languages. An important example is the study by Thompson and Newport (2007). In this case, the subjects were not infants but adults (undergraduates). The language consisted of a set of ‘sentences’ generated by a phrase-structure grammar. The sentences had the structure ABCDEF, each letter representing a specific form class which contained 3 words.

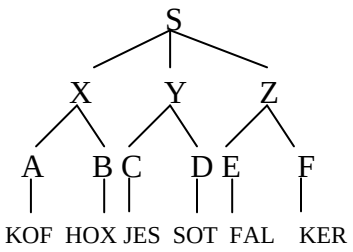


Figure 1: One example sentence

The form classes were grouped in “phrases”: AB, CD and EF which must be in that order. There were 4 variants of the basic language, which differed with respect to the surface cues to the phrasal structure of the sentences. In the “optional phrases” language, a grammatical sentence could consist of all 3 phrases or only 2 of them (preserving phrase order), resulting in 4 different sentence types: ABCDEF, ABCD, ABEF and

CDEF. In the "repeated phrases" language, a grammatical sentence could be just ABCDEF or could be ABCDEF to which one of the 3 phrases was added at the end: ABCDEFAB, ABCDEFCD, and ABCDEFEF. In the "moved phrases" language, a grammatical sentence was a combination of the 3 phrases in any order. The permutation generated 6 different sentence types: ABCDEF, ABEFCD, CDABEF, CDEFAB, EFABCD, and EFCDAB. In the "all-combined" language, a grammatical sentence belonged to one of the languages defined above, i.e., the "optional phrases" language, the "repeated phrases" language or the "moved phrases" language.[1]

For each of these variants, there was also a corresponding "control" language, in which the phenomena (repetition, movement) were not constrained by the phrase-structure. E.g., in the "optional control" language, a grammatical sentence could be BCDE; in the "repeated control" language, a sentence type could be ABCDEFBC and finally, in the "moved control" language, one sentence type was BCAFDE.

For each of these languages, the subjects listened to recordings of the sentences (which contained no relevant prosodic cues) for 20 minutes repeated over 5 days. In the test phase (on day 1 and day 5), they were tested on 2 forced-choice tests. In the "sentence test", they heard a pair of sentences in which one was grammatical (i.e., was compatible with the rules of grammar that generated that language) though it did not appear in the training sample, and the other was ungrammatical; they had to select the one they believed was grammatical. In the "phrase test", the choice was between 2 pairs of words, one of which was a grammatical phrase but the other was not. Participants were

[1] This language also differed from the others with respect to the number of words in each form class.

explicitly instructed to choose the pair of words that sounded “more like a group or unit from the language”. (p. 13)

The subjects performed above chance on day 1 and extremely well (approximately 78% to 89%) on day 5 for all of the language variants and for both the sentence and phrase tests. The authors attribute this to the fact that the phrasal manipulations created systematic peaks (within phrases) and dips (between phrases) in the transitional probabilities of successive words, which subjects could detect. While it might be debated whether the outcome of this learning was a representation of transitional probabilities or sentence patterns or of grammar rules, it is clear that some data-based learning was occurring in this adult population. The authors raise the question “whether the ability of adults to use transitional probability patterns to form groupings in the laboratory mirrors processes that subserve the acquisition of syntactic structure in infants.” (p. 40). Whether children could acquire syntax via transitional probability patterns is precisely the question of interest in the present dissertation.

Since it is difficult morally and technically to experimentally engineer the language input to a child, a useful contribution can be made by simulation studies in which the learner is a computational system, acquiring language facts on the basis of input selected by the experimenter for purposes of testing alternative models of how language learning might proceed. These computational systems vary greatly in their power, ranging from the simplest devices that gather co-occurrence data, to connectionist

models (neural networks) which can have up to hundreds of interacting units that track sequential patterns in input sentences.

An example at the lowest end is a bigram model such as studied by Reali and Christiansen (2005). A bigram model predicts which word sequences are in the target language on the basis of transitional probabilities, more specifically on the probability of one word given the preceding one. This will be the main learning model tested in the experiments reported in the present dissertation. The goal of these experiments is to establish the capabilities of this simple device in order to determine whether it can – as has been claimed – master the complexities of human language or whether richer and more sophisticated models are necessary. This is an important aspect of the current debate between data-driven and UG-based approaches to language acquisition.

The Reali and Christiansen study will be reported in section 1.2 below. As background for this, we first touch briefly here on results that have been reported for more powerful models. The most modest enrichment of the bigram model is perhaps the "lexical producer" learner or "Lexstat" developed by Chang et al. (2005, 2006). Lexstat had access to a training corpus of sentences of either child speech or child-directed speech. In the test phase, Lexstat was given the words of a sentence drawn from the corpus, but unordered. Its task was to reconstitute the sentence, predicting the next word at each point. In order to select the next word, Lexstat calculated a *choice score* for each unused word left in the set. This score was the sum of a *context score* and an *access score*. The context score was simply a bigram probability in the training corpus. The

access score of a word was the probability, based on the training corpus, of that word preceding any other word in the sentence (with any number of words between them). The latter measure was innovative; it may be useful to track long-distance dependencies between words. The authors view the access score as an indicator of the lexical activation level of the word: “When the access information for all the candidate words is combined together, it simulates the competition between the words that are activated by the message in the message pathway”. (2005, p. 3)

The corpora (child speech and child-directed speech) were drawn from five typologically different languages to establish whether Lexstat’s ability to learn can be generalized across all language types. For each language, one of the tests was to predict the child’s utterances based on the adult sample. The success rate was approximately a range of 43% to 70% correct across the five languages, on average 36% greater than chance performance. Based on these results, the authors maintained that “lexical access knowledge should play a part in theories of syntax acquisition” (2005, p. 5). More generally, they concluded that “the input provides much of what the learner needs, to predict syntactic orderings, and that would militate against the view that the input is impoverished.” (2005, pp. 1-2). Nevertheless, our experiments in the following chapters did not include Chang et al.’s access statistic but were restricted to unsupplemented bigram statistics in line with Reali and Christiansen’s study. Besides, informal inspection of the results in Chang et al. (2006, see figure 2) shows that the access statistic only added a few percentage points to the successful performance rate compared with the “Prevword” measure which was the most closely related to a simple bigram statistic).

A major advance in computational power is provided by neural networks. A neural network is a computational model intended to simulate some of the properties of interactions between the neurons in the human brain. Network theory (‘connectionism’) has had applications in many scientific fields but here we will restrict discussion to language learning. A neural network is defined by a set of nodes (akin to the neurons in the brain) that are interconnected with each other and they serve different functions depending on where they are in the network architecture. There are input nodes (the ‘input layer’) which receive the input or data. They are linked through weighted connections. Learning is attained by adjusting the weights to get the desired output read off from the ‘output nodes’. Usually, in application to language tasks, there is also a ‘hidden layer’ which contains nodes that are intermediate between the input nodes and the output nodes. The hidden nodes make it possible for the network to solve classification problems where the output is an arbitrary function of the input.

In particular, SRNs (Simple Recurrent Networks) have had some noteworthy success in language learning, including the learning of syntax. The term ‘simple’ means that there is only one hidden layer and one output layer. The term ‘recurrent’ refers to a network in which weights from previous activations are fed into the input nodes together with the new input information to provide information about previous events (words) in the sentence. Because there is in principle no limit on how many previous words a network can track, it has the potential for more sophisticated computation than an n-gram model in which there is a strict limit on the size of *n* (for bigram models, n = 2).

One of the most famous studies involving an SRN and language learning was that of Elman (1993). He tested whether the neural network could learn basic language facts such as part-of-speech and number agreement, and also more complex grammatical constructions like embedded relative clauses. The task of the network was to predict the next word. The test sentences had subject-verb agreement and verb argument structure patterns which spanned the embedded relative clauses, requiring the system to keep track of syntactic dependencies at multiple levels. A property which has been much emphasized was that learning was successful but only when the complexity of the input data was increased over time. Elman argued that the network needed to acquire basic structural facts about number, argument structure and relative clauses before being able to integrate them together, and speculated that the same was true for children (Newport, 1990). More generally, Elman concluded that "networks are able to extrapolate beyond their training data in ways which obviate the need (for example) to see all possible combinations of words in sentences. In other words, networks generalize" (p. 87).

Networks such as Elman studied will be touched on only briefly in the course of this dissertation. They are far more sophisticated learning models than n-gram models, but some studies as noted above have claimed that a simple learning model involving only bigrams might be sufficient for syntax acquisition. Clearly, it is appropriate to verify such claims before going on to investigate more complex learning models. Perhaps, it would be found that the latter are not needed.

1.2. PIRCs: the construction of interest

Returning to the poverty of the stimulus argument which claims that children can acquire a construction despite the absence of evidence in the data that they receive, one construction has often served as an illustration of the principle. The construction involves a complex question containing a relative clause: e.g., *Is the boy who is crying hurt?*. It is referred to in this study as a *Polar Interrogative with Relative Clause* (henceforth PIRC). A polar interrogative is formed in English by the inversion of an auxiliary verb. The auxiliary-inversion rule requires the fronting of the highest auxiliary. As Chomsky observed (Chomsky, 1971), a learner might alternatively form a hypothesis that it is the first auxiliary that is inverted. Note that both rules predict the correct form for a simple example such as *Is the boy hurt?*. However, only the inversion of the highest auxiliary is correct for a wider range of examples including multi-clause constructions such as: *Is the boy who is crying hurt?*. Children's knowledge of which auxiliary to front (as demonstrated by Crain and Nakayama, 1981) must be explained. They would have relevant evidence if they encountered PIRC examples like: *Is the boy who is crying hurt?*. However, Chomsky argued that children's linguistic environment does not provide many such exemplars (or maybe none), hence the knowledge must be innate. On the other hand, supporters of data-driven learning have attempted to demonstrate that a computational model tracking statistical probabilities can detect the correct form of auxiliary-inversion in PIRCs.

In particular, the study by Reali & Christiansen (2005) was the starting point of our present research. They trained a bigram model on a corpus of child-directed speech,

which contained simple examples of relative clauses and of auxiliary inversion but no PIRCs in which the two co-occurred. The model was then presented with sentence pairs consisting of grammatical and ungrammatical PIRC constructions (the grammatical version with the highest auxiliary fronted and the ungrammatical with the first auxiliary fronted). The model had to choose the correct version (see chapter 2 below for more details on the methodology). The results showed an overwhelming success: 96% correct selection. These results are impressive for three reasons: a) the complexity of the test construction, b) the simplicity of the statistical model, and c) the fact that the corpus of child-directed speech was limited because addressed to very young children. However, this study exhibited some flaws that we address in the next chapter, including the fact that the model was able to succeed without having gained any genuine knowledge of the syntactic rule underlying the PIRC construction.

1.3. Outline of the dissertation

The overall goal of the dissertation is to make a contribution to delineating the learning resources and capacities of a normal child. The approach taken here is innovative because it thoroughly explores the lower bounds of what is required in order to be able to acquire the correct rule for PIRCs. It hopes to serve as groundwork for future research into the upper bounds on a psychologically plausible language learning system.

This work focuses on the bigram model, which tracks only transitional probabilities between pairs of words, because it is the simplest type of statistical model for language. As noted above, it is known that children can track transitional

probabilities; therefore if this simple model can master a complex syntactic construction, then it could be concluded that children need nothing more than this low-powered statistical mechanism. The PIRC construction will serve as a probe in the experiments, as it has become an established target in the poverty of the stimulus debate.

Chapter 2 will delve into the success of Reali & Christiansen's learning model and analyze its source. It will also test whether the bigram model's learning ability can be extended to the full range of variants of the PIRC construction. These experiments have mostly negative results which expose the weakness of the bigram model and imply that some scaling up is needed. This will be the work of chapter 3.

In chapter 3, the strategy is to add resources to the learning situation in increments, to find out at what point in this escalation the learning model begins to exhibit a general grasp of auxiliary inversion. The results show that these manipulations bring surprisingly little benefit, indicating that yet more sophisticated learners and/or input information must be explored.

The nature of the problems encountered in chapter 3 suggest the diagnosis that the PIRC learning failure was due to the bigram model's inability to compute basic phrase-structure. In chapter 4, it was decided therefore to make some phrase-structure information available to the model by injecting it into the corpus. One of several different implementations of this did finally succeed in producing essentially perfect performance

on PIRCs. What remains to be established in further research, therefore, is whether the model itself can build the phrase-structure on which this kind of success is based.

Chapter 2: Bigram-based learning and the richness of the stimulus for language acquisition[2]

2.1. 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

[2] This chapter was published in its entirety as a co-authored paper in *LIBA: CUNY/NYU Working Papers in Linguistics* (Kam et al., 2007) and parts of it were published in *Cognitive Science* (Kam et al., 2008).

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.[3]

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

[3] 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).

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.[4] In this respect it breaks new ground, 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).

[4] 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 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.

(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 inversion[5] 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.

[5] 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.

2.2. 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 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)[6].

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

Ungrammatical Is NP {who|that} A is B ?

where A is instantiated by VP and B by VP, PARTICIPLE, NP, PP, ADJP, etc[7].

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.

[6] Note that these templates imposed tight limits on the form of the PIRC sentences: in both the main and relative clause, there could be only one auxiliary verb and it had to be *is*; the noun phrase containing the relative clause had to be singular; the relative clause always had its 'gap' (trace) in subject position. Also, though it is not clear in number (3) above, the predicate in the relative clause had to be progressive (e.g. *running*). Of course there are PIRCs with plural noun phrases and/or with more than one auxiliary verb or none at all, in either the main or the relative clause (e.g. *Could the little boys who cried have been hurt?*). Removing some of the limitations imposed by Reali and Christiansen might well affect the performance of the bigram model; see discussion in 2.5. of PIRCs with auxiliary *do*, and relative clauses with object gaps. It could be of interest in future research to investigate more such variations. (See Ambridge et al., 2008 for data on children's performance on PIRCs with different auxiliaries.)

[7] For reasons not clear to us, all of R&C's test sentences had A instantiated as a progressive participle (*crying*), and for comparability of outcomes we matched our materials to theirs.

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 $w_{i-1}w_i$: $P_{ML}(w_i \mid w_{i-1}) = c(w_{i-1}w_i) / c(w_{i-1})$

Interpolated (smoothed) probability of bigram $w_{i-1}w_i$: $P_{interp}(w_i \mid w_{i-1}) = \lambda P_{ML}(w_i \mid w_{i-1}) + (1-\lambda)P_{ML}(w_i)$

Interpolated (smoothed) probability of trigram $w_{i-2}w_{i-1}w_i$: $P_{interp}(w_i \mid w_{i-1}w_{i-2}) = \lambda P_{ML}(w_i|w_{i-1}w_{i-2}) + (1-\lambda)(\lambda P_{ML}(w_i|w_{i-1}) + (1-\lambda)P_{ML}(w_i))$

Cross-entropy of a test sentence s_T: $$H(s_T) = -\frac{1}{N_T}\log_2 \prod_{i=2}^{N_T} P_{interp}(w_i \mid w_{i-1})$$

Equation 1: Formula of unigram, bigram and trigram probabilities and of cross-entropies

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.

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

Table 1: Selection by the bigram model in R&C's Experiment 1

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.[8] On the basis of this strong 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

[8] 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.

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.[9] 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

[9] 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).

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 *is-is* PIRC construction.[10] 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.[11] Our data show that it does not extend

[10] 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).

[11] 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 (2008) 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.

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.

2.3. Understanding the bigram model's success

2.3.1. 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".[12] 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.

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 2 portrays the data in Table 2 graphically. (All test sentences are available in the appendix.)

	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 - *Do*-support	100	49	51	0
Expt6 - Verb inversion in Dutch	40	32.5	55	12.5

Table 2: Selection by the bigram model in R&C's Experiment 1 and in Experiments 1–6

[12] 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.

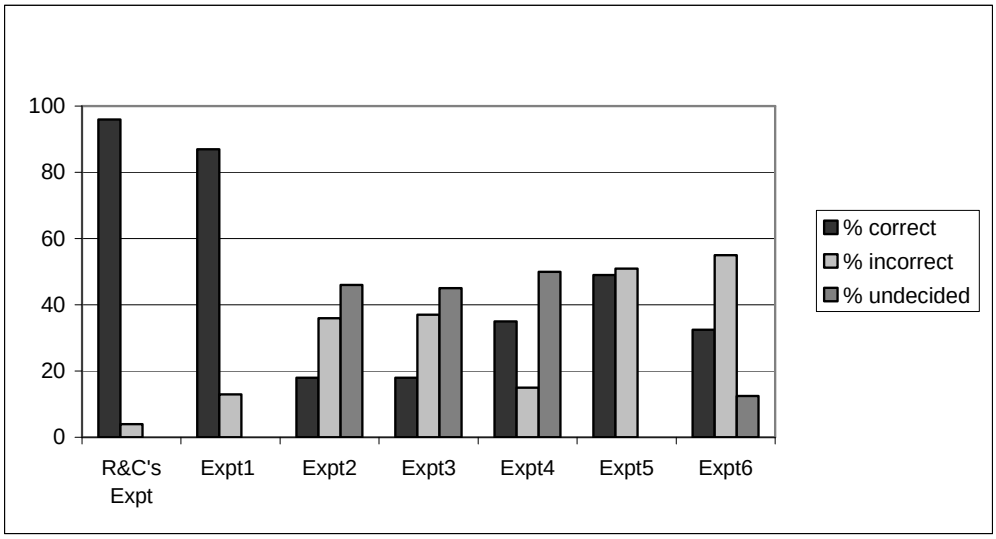


Figure 2: Selection by the bigram model in R&C's Experiment 1 and in Experiments 1–6

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

2.3.2. 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.[13] 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?

[13] 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.

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*>.

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

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

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 2.2). 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.[14] 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.

[14] All statements in sections 2.3, 2.4 and 2.5 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.

	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

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

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.[15] Hence, when the probabilities of the bigrams in each sentence version are multiplied together to give the estimated sentence probability,[16] 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.)

[15] 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".)

[16] 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.

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.[17] Aggregated over the set of test sentences, this can be expected 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: <*crying 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).[18] 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

[17] 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*).

[18] 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 <*crying hurt*> in sentence (4).

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*>.[19] 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

[19] 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.

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*>).[20] If its probability exceeds that of the <*who*|*that is*> bigram, the model could choose the ungrammatical 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

[20] 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.

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.[21] 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.

2.3.3. 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

[21] 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.

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; 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.*[22] 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

[22] 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.

(e.g., *That is a rose*).[23] Thus the "who" or "that" of a <*who*|*that is*> bigram was in every case merely a homograph of a relative pronoun.[24] 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

[23] 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 2.2). 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.

[24] 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.

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 preceding "is".[25] 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.[26] 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.

2.4. Without the 'wrong' bigrams

It is predicted that without the facilitating effect of the 'wrong' <*who*|*that is*> bigram in

[25] 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.

[26] The linguistics literature provides no definitive count of such languages, but an informal poll conducted through Linguist List also yielded Avestan, Haida, Hungarian, Kambera, 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.)

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.

2.4.1. 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 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.[27] Thus, some testable empirical predictions flow from the suggestion that

[27] 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

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.

2.4.2. Experiment 3: Homography with a determiner

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

---

tagging (over 100 categories) to avoid these potential problems. (An SRN study by Lewis & Elman, 2001, used non-tagged input; see section 2.6 below.)

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 2.6 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.

2.5. 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., <*who the*>, <*who Jim*>, <*who every*>) 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 <*who*|*that is*> 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 <*who*|*that is*>

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.

2.5.1. 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).[28] The trace of the object is shown as $t_{\mathbf{j}}$ in the examples below, coindexed with the phonologically null relative pronoun $Ø_{\mathbf{j}}$, 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 $t_{\mathbf{i}}$, coindexed with the moved auxiliary $is_{\boldsymbol{i}}$.

(6) $\text{Is}_{\mathbf{i}}$ the $\text{wagon}_{\mathbf{j}}$ [$Ø_{\mathbf{j}}$ your sister is pushing $t_{\mathbf{j}}$] $t_{\mathbf{i}}$ red?

(7) * $\text{Is}_{\mathbf{i}}$ the $\text{wagon}_{\mathbf{j}}$ [$Ø_{\mathbf{j}}$ your sister $t_{\mathbf{i}}$ pushing $t_{\mathbf{j}}$] 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

[28] 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.

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 contained an overt relative pronoun at the position of $Ø_j$. 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.

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

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

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.

2.5.2. 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$_{\mathbf{i}}$ the boy [who plays the drum] $t_{\mathbf{i}}$ want a cookie?

(9) * Does$_{\mathbf{i}}$ the boy [who $t_{\mathbf{i}}$ 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 $t_{\mathbf{i}}$.[29] 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.)

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

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

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

[29] 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.

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. 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.[30]

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

[30] 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.

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.

2.5.3. 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 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.[31] 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 2.6).

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

[31] 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.

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.[32] 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?
‘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

[32] 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’).

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").

	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

Table 7: Smoothed probabilities for the distinguishing bigrams in (10) and (11)

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.[33] 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

[33] 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.

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

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.

2.6. 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 2.1 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 arguably[34]) 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.

[34] 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.

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.[35] Even the neural network 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. 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.

[35] The realistic psychological resources condition excludes interesting work by Clark and Eyraud (2006) and Chater and Vitanyi (2007). It may or may not exclude Perfors et al. (2006) and neural network results such as those of Elman (1993), Lewis and Elman (2001) and Frank et al. (submitted); but the latter did not employ real-life corpora of child-directed speech.

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. 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 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*>. It would be interesting to quantify the advantage this confers (if any) on the troublesome subtypes of English PIRCs with object-gaps and *do*-support.

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 2.4.1).

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.

Chapter 3: Augmenting the resources for learning[36]

The bigram model in the Kam et al. experiments failed to generalize the PIRC construction, but it would be unfounded to claim that it fails as a learning model. Some minor adjustment of the learning mechanism or the linguistic input might tip it over into successful performance. It should be noted that R&C's test situation was especially challenging, in that the training corpus was limited in size and in the age of the children being addressed. Possibly a larger or more varied corpus would provide more cues that the bigram model could make use of. Alternatively, the corpus might have been adequate but the bigram statistics too limited in scope to benefit from it.

To investigate these issues I ran new experiments with increasingly larger corpora and with a trigram learning model, aimed at improving the n-gram model's chances of successful learning. As probes to test for improved performance, I used the English object-gap and do-support PIRCs on which the bigram model had previously failed.

3.1. Experiments 7, 8, and 9: Enriching the corpus

3.1.1. Methodology

Experiment 7 employed what I will call an "older" corpus, containing adult utterances directed to a child older than any of the children in the Bernstein-Ratner corpus. It was the corpus of Adam from age 2;3 to 5;2, containing 25,732 utterances of child-directed speech. Experiment 8 used an even larger corpus of 110,629 adult utterances (about 10

[36] Some sections of chapter 3 were published in the Proceedings of the 31st Boston University Conference on Language Development (Kam, 2007).

times larger than the original corpus of Bernstein-Ratner), created by merging corpora for 71 different children into the Bernstein-Ratner corpus, with a total age range from 0;7 to 8;0. These additional corpora were from Brown (1973), MacWhinney (2000), Menn & Gleason (1986), Suppes (1974), Valian (1991) and Warren-Leubecker (1984). In experiment 9, the bigram model was trained on the Wall Street Journal corpus (henceforth WSJ). We used the unparsed corpus that contains 2,499 stories selected from a three-year (1987 to 1989) WSJ collection of 98,732 stories. 2,235,937 utterances were extracted. Note that unlike experiments 7 and 8, the corpus was not a child-directed speech text but a collection of articles written by journalists. In the three experiments, the learning model and test procedure were identical to those of the R&C and Kam et al. studies. The object-gap and do-support test sentences were identical to those of Kam et al. except for some word replacements so that all unigrams that appear in the test sentences occurred in the training corpus.

3.1.2. Results

Results are shown in Table 8. The model's discrimination of grammatical PIRCs did improve under these novel conditions, but the errors on object-gap and do-support sentences still range from 23% to 53%[37], well below the performance on the original is-is subject-gap PIRCs (paired t-test comparisons of object-gap PIRCs between the original corpus and the older corpus: $t(99) = 1.031$, ns; of object-gap PIRCs between the original corpus and the larger corpus: $t(99) = 0.694$, ns; of object-gap PIRCs between the original corpus and the WSJ corpus: $t(99) = 7.700$, $p < 0.0001$; of do-support PIRCs between the

[37] The error rates include the incorrect and the undecided percent selection.

original corpus and the older corpus: $t(99) = 2.902$, $p < 0.01$; of do-support PIRCs between the original corpus and the larger corpus: $t(99) = 4.060$, $p < 0.0001$; of do-support PIRCs between the original corpus and the WSJ corpus: $t(99) = 3.844$, $p < 0.0001$). Note that only in the case of the Wall Street Journal, a linguistically more sophisticated as well as much larger corpus than the others, did performance on object-gap and do-support PIRCs exceed 70%. The WSJ data are included here for comparison purposes. If Crain & Nakayama are right that children have mastered the essentials of PIRCs in their pre-school years, then the WSJ corpus results are strictly irrelevant to the issue of the poverty of stimulus for language acquisition.

IS-IS SUBJECT PIRC	**% correct**	**% incorrect**	**% undecided**
Original corpus	87	13	0
Adam (older) corpus	82	18	0
Larger and older corpus	79	21	0
WSJ corpus	65	35	0
OBJECT-GAP PIRC	**% correct**	**% incorrect**	**% undecided**
Original corpus	35	15	50
Adam (older) corpus	47	45	8
Larger and older corpus	63	30	7
WSJ corpus	77	23	0
DO-SUPPORT PIRC	**% correct**	**% incorrect**	**% undecided**
Original corpus	49	51	0
Adam (older) corpus	55	45	0
Larger and older corpus	70	30	0
WSJ corpus	71	29	0

Table 8: Selection by the bigram model based on different training corpora

3.1.3. Analysis

For the “is-is” subject PIRCs, performance decreases steadily as the corpus grows bigger (paired *t*-test comparisons of subject-gap PIRCs between the original corpus and the older corpus: $t(99) = 8.660$, $p < 0.0001$; of subject-gap PIRCs between the original corpus and the larger corpus: $t(99) = 2.711$, $p < 0.01$; of subject-gap PIRCs between the original corpus and the WSJ corpus: $t(99) = 5.680$, $p < 0.0001$). This is attributable to the fact that more and more bigrams are attested with the increase of the size of the corpus, so the <*who*|*that is*> bigram probability faces greater competition from other bigrams and slowly but surely loses its dominant role in selecting the grammatical version of the sentence. The model’s choice becomes more variable than before since both the ungrammatical as well as the grammatical version has more attested bigrams than before. Table 9 below provides some details. When trained on the WSJ journal, which is by far the largest corpus, the third distinguishing bigram looks like the main reason for why the bigram model is shifting its choices more towards ungrammatical sentences. This is plausible because, as noted in section 2.3.2, b3-ungrammatical (<*is hurt*>) is a totally valid sequence in English while the corresponding b3-grammatical (<*crying hurt*>) is not, except in PIRCs and other relatively infrequent constructions (e.g. a complex sentence such as *the boys who are crying hurt their knees*). This difference is reflected in the percentage of attested bigrams for b3-ungrammatical: 36.65% and percentage of attested bigrams for b3-grammatical: 15.09%. At least where the smoothing factor is the same, as in the present case, attested bigrams always contribute more to the probability of the sentence than unattested bigrams.

	100 grammatical test sentences	100 ungrammatical test sentences
Total of distinguishing bigrams	300	300
Total of attested Distinguishing bigrams	232 (77.3%)	251 (83.66%)
Total of attested 1st distinguishing bigrams	100 (43.1%) average probability: 0.021	73 (29.09%) average probability: 0.0022
Total of attested 2nd distinguishing bigrams	97 (41.81%) average probability: 0.0035	86 (34.26%) average probability: 0.0144
Total of attested 3rd distinguishing bigrams	35 (15.09%) average probability: 0.0064	92 (36.65%) average probability: 0.0096

Table 9: Distribution of the attested bigrams in is-is subject PIRCs

The object-gap PIRCs benefited the most from increasing the size of the corpus. The older/larger corpora greatly reduced the number of undecided cases. The WSJ completely eliminated the undecided cases. Since an older and/or larger corpus is likely to contain more bigram types as well as tokens, undecided cases arise less often. Furthermore, unlike the "is-is" experiment, increasing the size of the training corpus favors the grammatical sentence. When trained on the WSJ journal, 226 bigrams out of 300 in the grammatical sentences (i.e., 75.33%) are attested in the corpus, while 187 out of 300 bigrams in the ungrammatical sentences (62.33%) are attested. In this case, the first distinguishing bigram in the grammatical sentence (see table 10 below for the distinguishing bigrams in object-gap PIRCs) is the one most responsible for influencing the selection towards the grammatical sentence.

	100 grammatical test sentences	100 ungrammatical test sentences
Total of distinguishing bigrams	300	300
Total of attested distinguishing bigrams	226 (75.33%)	187 (62.33%)
Total of attested 1st distinguishing bigrams	88 (38.89%) average probability: 0.0168	10 (5.35%) average probability: 0.00022
Total of attested 2nd distinguishing bigrams	95 (42.0%) average probability: 0.0005	82 (43.85%) average probability: 0.0074
Total of attested 3rd distinguishing bigrams	43 (19.11%) average probability: 0.0027	95 (50.80%) average probability: 0.0027

Table 10: Distribution of the attested bigrams in object-gap PIRCs

In an object-gap PIRC the bigram b1-grammatical always consists of the last word of an NP followed by *is*; obviously, this is a very common pattern in English. In the original small corpus, many of the b1-grammatical bigrams in the test sentences were not attested in the corpus, but with the much larger WSJ corpus, 88% of them were attested, and the average probability was relatively high compared with the other distinguishing bigrams because of the systematic position of *is* (a very frequent token) in the 2nd position in the bigram. In the formula for calculating bigram probabilities, the count of the second word in the bigram is used: in the numerator of the smoothing factor i.e., the 2nd term of the sum. The smoothing factor involves dividing the counts of the 2nd word by the total number of words in the training corpus. As a consequence the higher the number in the numerator is, the higher the smoothing factor is, which in turn increases the bigram probability.

For do-support PIRCs, the richer corpora brought some benefit: the maximum success rate was 70% for the larger child-directed corpus and 71% for the WSJ corpus. However, the percentage of attested bigrams was very close in the grammatical and

ungrammatical test sentences (92.67% versus 93%). Even their distribution was very similar. None of the distinguishing bigrams in these sentences recurred in many of the test pairs, since they all contained a content word; therefore the bigram model had no particular bigram(s) that it could count on to favor the grammatical version. As it happens, the combined probability of the bigrams in the grammatical version often outweighed the bigrams in the ungrammatical version, leading to 70% correct selections.

	100 grammatical test sentences	100 ungrammatical test sentences
Total of distinguishing bigrams	300	300
Total of attested distinguishing bigrams	278 (92.67%)	279 (93%)
Total of attested 1st distinguishing bigrams	94 (33.81%) average probability: 0.001	94 (33.69%) average probability: 0.001
Total of attested 2nd distinguishing bigrams	74 (26.62%) average probability: 0.0356	85 (30.47%) average probability: 0.0341
Total of attested 3rd distinguishing bigrams	27 (9.71%) average probability: 0.0029	23 (8.24%) average probability: 0.0036
Total of attested 4th distinguishing bigrams	83 (29.86%) average probability: 0.0291	77 (27.60%) average probability: 0.0277

Table 11: Distribution of the attested bigrams in do-support PIRCs

The lack of improvement between the larger corpus of child-directed speech, and the much larger WSJ corpus suggests that learning of do-support PIRCs has reached a plateau and that providing yet more input won't increase the learning rate. While only a speculation, this is an important possibility deserving further investigation. To obtain an additional indicator of whether performance does reach an inherent limit regardless of the amount of corpus data provided to the model, I split the larger child-directed speech corpus into five bins of roughly equal size (22,126 utterances each) and tested how they contribute to the learning of do-support PIRCs. Those bins were employed as the training

corpus for the discrimination task, cumulatively and in three different orders as shown in the tables below. The question was whether performance would improve systematically as the training corpus increased in size; and also whether there would be an effect of the presumed increase in richness of the corpus from bin 1 to bin 5. As can be seen in Table 12 below, bin 1 alone (the earliest sentences of the corpus) doesn't contribute much; it yielded only a 55% success rate. Interestingly, however, results are similar for bin 5 alone, as seen in the second and third tables. Thus, the cause is not whether the sentences were addressed to very young children, and hence are likely to be impoverished; rather, it appears that what depresses the success rate is just the amount of data received by the model.

In line with this conclusion, performance did improve cumulatively -- though only slightly -- as more data were added by merging bins together in larger and larger combinations. However, across all three such trials, the maximum performance still remained at only 71%. Note that some but not all of the modest improvement with corpus size was due to reduction of the 'undecided' responses. It will be worth repeating this study in future with an even larger corpus, such as the WSJ corpus; possibly the differences between the success rates as a function of corpus size will be greater than here, so that improvement trends can show up more sharply.

	% Grammatical	%Ungrammatical	% undecided
Bin1	55%	30%	15%
Bin1+2	59%	32%	9%
Bin1+2+3	59%	32%	9%
Bin1+2+3+4	60%	32%	8%
Bin1+2+3+4+5	68%	25%	7%

	% Grammatical	% Ungrammatical	% undecided
Bin5	56%	29%	15%
Bin5+4	65%	21%	14%
Bin5+4+3	66%	23%	11%
Bin5+4+3+2	67%	25%	8%
Bin5+4+3+2+1	71%	23%	6%

Table 12: Selection by the bigram model based on cumulative corpora[38]

3.1.4. Enriching the corpus: summary

Experiments 7, 8, 9 agree that even when the bigram model is trained on a larger data set, the outcome for object-gap and do-support PIRCs is not very impressive. This sluggish responsiveness to corpus size is explicable. A larger corpus tends to yield higher counts of the bigrams in the test sentences, but as noted above, this is true of bigrams in *both* sentence versions so it could favor the ungrammatical version just as much as the grammatical version. For example, the results of the "is-is" PIRCs and the object-gap PIRCs with the WSJ corpus in table 8 show that the additional resources favored the bigrams in the ungrammatical versions for the former (from 87% correct selection with the original corpus down to 65% correct selection with the WSJ corpus) but the bigrams in the grammatical sentences for the latter (from 35% correct selection with the original corpus up to 77% of correct selection with the WSJ corpus).

Another relevant fact is that there is no distinguishing bigram that appears systematically in these object-gap and do-support test sentences (comparable to the

[38] The results for Bin1+2+3+4+5 were different from Bin5+4+3+2+1's results because in the former, bin 1 was the first run, hence the test sentences were composed using the words of the corpus in bin 1. In the latter bin order, a slightly different set of test sentences had to be created using the vocabulary of bin 5 (the first bin to be tested). These minor adjustments in the vocabulary of the test sentences explain the slight difference in performance in the two cases.

<*who*|*that is*> bigram in the is-is subject-gap test sentences), which would serve as a good marker for the grammatical version if only it were strongly represented in the corpus. As noted in section 3.1.3 above, object-gap PIRCs lack a good 'marker' bigram for the grammatical version because the relative pronoun does not appear in any of the distinguishing bigrams, and the "is" in the relative clause co-occurs in bigrams with a different lexical item in every test pair. In do-support PIRCs the relative pronoun is followed by a different lexical verb in each grammatical test item. Moreover, a higher bigram count does not necessarily entail a higher smoothed probability for that bigram as established by equation 1 above. A higher count of word w_1 can decrease the bigram probability (because it is in the denominator), while a higher count for word w_2 contributes to a higher smoothed bigram probability (because it is in the numerator). Moreover, a word may appear as w_1 in one bigram and as w_2 in another (e.g., "plays" in <*plays the*> versus <*who plays*>), thus further complicating its contribution to the overall probability of the sentence. These rather unruly factors explain why the gains from a richer corpus are more meager than might have been expected.

3.2. Experiment 10: Providing exemplars of PIRCs in the corpus

3.2.1. Motivation and methodology

Another way to enrich the corpus (and thereby possibly improve discrimination performance) would be to present some grammatical PIRC exemplars to the bigram model. R&C were very clear that their goal was to demonstrate bigram-based learning of PIRCs even in the absence of PIRCs from the training corpus. If successful, this would be the strongest way to undercut Chomsky's poverty of stimulus arguments. However, that

goal may have been too ambitious. The results presented in this dissertation do not strongly support the initial success reported by R&C (for the is-is subject-gap PIRC type). It is at least worthwhile, therefore, to find out whether what has been holding success rates down is that the bigram model does really need exemplars of the target construction.

To do so, I manually constructed 50 object-gap and 50 do-support PIRCs and added them to the original Bernstein-Ratner corpus. These exemplars were newly constructed and were not identical to any of the test sentences used in this experiment (See appendix for the complete set of sentences). The procedure and the test sentences were identical to the previous experiments.

One object-gap PIRC example that was added to the training corpus was:

(13) Is the toy that the baby is holding in his hands a teddy bear?

One do-support PIRC example that was included to the training corpus was:

(14) Does the rabbit that eats grass prefer carrots instead?

3.2.2. Results

The results show that is-is subject gap PIRCs are still well-predicted with 80% correct (paired *t*-test comparisons of subject-gap PIRCs between the original corpus and the enriched corpus with PIRCs: $t(99) = 1.580$, ns). Note that for these test sentences there was no gain (even a slight loss). Performance on object-gap and do-support test sentences does show some improvement (paired *t*-test comparisons of object-gap PIRCs between the original corpus and the enriched corpus with PIRCs: $t(99) = 3.897$, $p < 0.0001$; of do-

support PIRCs between the original corpus and the enriched corpus with PIRCs: $t(99) = 3.880$, $p < 0.0001$), but remains weak: only 64% correct for object-gap type and 57% correct for do-support type (see table 13 below).

IS-IS SUBJECT PIRC	**% correct**	**% incorrect**	**% undecided**
Original corpus	87	13	0
PIRCs added in corpus	80	20	0
OBJECT-GAP PIRC	**% correct**	**% incorrect**	**% undecided**
Original corpus	35	15	50
PIRCs added in corpus	64	12	24
DO-SUPPORT PIRC	**% correct**	**% incorrect**	**% undecided**
Original corpus	49	51	0
PIRCs added in corpus	57	43	0

Table 13: Selection by the bigram model when the corpus was provided with PIRCs

The fact that providing explicit exemplars of correct PIRCs does not result in mastery of the construction is an important finding, because it suggests that what is holding back learning is not the poverty of the stimulus; rather it is the weakness of the learning model. Evidently, the bigram model does not have the ability even to detect or represent the pertinent properties of a well-formed PIRC. As could have been suspected from the outset, these pertinent properties of PIRCs cannot be formulated as facts about bigrams. Hence they are not detectable or usable by a learning system that can make use of information only if it is expressible in bigram format. This is not possible for the syntactic dependency that is the essence of the PIRC construction; it is a dependency between non-adjacent positions (the surface position of the moved auxiliary, and position of a ‘gap’ where that auxiliary would appear in a declarative sentence).

Stating this conclusion more cautiously: Coding the auxiliary-movement dependency in terms of bigrams is impossible if sentences are treated as mere sequences of words, as in all the experiments reported so far. If there is to be any breakthrough into bigram-based learnability of PIRCs, it is therefore most likely if the word sequences can first be translated into, or supplemented with, syntactically relevant categories. This will be the research strategy for all remaining experiments in the dissertation, starting with part of speech categories and then moving onto phrasal categories.

3.3. Experiments 11 and 12: Providing syntactic category information

3.3.1. Motivation and methodology

Quite apart from this plan of moving in the direction of a more linguistically appropriate abstract representation of sentences, there are additional reasons for thinking that a statistical model would benefit from part of speech (PoS[39] or syntactic category) designations for the words in sentences. The previous experiments have indicated that part of the trouble for a bigram model may lie in the unstable frequencies of specific words and word combinations in the training corpus due to the sparse data available for individual items. Representing sentences in terms of the syntactic categories of their words, rather than the words themselves, helps to solve this problem of the sparseness of data (which was not greatly improved by the shift to a considerably larger corpus in Experiments 7, 8 and 9). When specific words are aggregated into more general syntactic categories, the bigram counts are increased and become more reliable, and unattested bigrams are reduced. The greatest beneficiaries are bigrams containing open-class items.

[39] Note that this abbreviation PoS for part of speech is used here to differentiate it from POS which was used as the abbreviation of Poverty of the Stimulus in previous chapters.

For example, <*sister is*> would be represented as consisting of a singular noun followed by a 3rd person singular auxiliary. The same would be true of <*day is*> and <*hat has*> and many more word pairs. Aggregating these in terms of PoS categories results in the bigram <*n v:aux&3S*>, a much more frequent bigram than <*sister is*> alone. Because of this, bigrams containing content words can play a greater role in selecting between sentence versions, which otherwise tends to be dominated by bigrams consisting of frequent closed-class items (function words) such as <*that is*> as noted in chapter 2. This stabilization of the bigram analysis of the training corpus is in addition to the potential benefit of syntactic category information to the learning model by increasing its ability to capture a general structural pattern for auxiliary inversion.

In Experiment 11, all words were replaced with their PoS-tags. In Experiment 12, function words were left unchanged. The motive for this derives from the finding by Mintz (2003) that function words are the most useful items for identifying the grammatical roles of other words. Mintz's 'frequent frames' algorithm groups words into classes if they share a frame (consisting of the preceding word and the following word in the corpus). The frames that occurred most frequently yielded accurate and comprehensive part of speech classifications, and Mintz observed that the framing items in these cases were predominantly closed-class words (e.g., *put__on*, *what__you*). Thus it seemed that the mixed-level representation in Experiment 12 could be especially helpful since it grouped lexical items into broader categories without losing the specific information carried by individual function words.

In Experiments 11 and 12 the learning model and procedure were exactly the same as in the Kam et al. study though the sentence representations differed. Starting with the same corpus and test sentences as Kam et al.'s experiments 1, 4 and 5, all the words in them were replaced in Experiment 11 by their part-of-speech tags, using the MOR program (available in the Childes database; MacWhinney, 2000). 117 distinct part-of-speech tags were used (see appendix for the list of part-of-speech tags). For instance, sentence (15) from the corpus was converted into the string (16). In experiment 12, only lexical words in the corpus and test sentences were replaced by their part-of-speech tags; the function words were left in their original form (function words included prepositions, auxiliaries and determiners) Thus in this experiment, sentence 15 was replaced in the corpus by sentence (17).

(15) you want to see the book

(16) pro v inf v det n

(17) you v to v the n

3.3.2. Results

A part-of-speech tagger solves the problem of the sparseness of data by collapsing words with the same syntactic categories under the same type. Because many of the words in the test sentences shared the same categories, it was expected that some test sentences would have identical part-of-speech patterns, thus reducing the number of distinct test sentences. In fact, because the part-of-speech tagger that was used was very fine in the degrees of tagging, many test sentences did not collapse with others into a single pattern, resulting in a good number of different test sentence patterns. For the object-gap PIRCs, the original

100 test sentences were reduced to 76 different patterns for the PoS-tags only experiment and to 83 for the PoS-tags + function words experiment. For the do-support PIRCs, the 100 test sentences were collapsed into 98 different patterns and 99 for the PoS-tags only experiment and the PoS-tags + function words experiment respectively. Results are shown in Table 14.

OBJECT-GAP PIRC	**% correct**	**% incorrect**	**% undecided**
PoS-tags only (100 tagged test sentences)	54.0	44.0	2.0
PoS-tags only (76 uniquely tagged test sentences)	51.3	46.1	2.6
PoS-tags+ function words (100 tagged test sentences)	45.0	52.0	3.0
PoS-tags + function words (83 uniquely tagged test sentences)	41	55.4	3.6
DO-SUPPORT PIRC	**% correct**	**% incorrect**	**% undecided**
PoS-tags only (100 tagged test sentences)	70.0	30.0	0
PoS-tags only (98 uniquely tagged test sentences)	70.4	29.6	0
PoS-tags + function words (100 tagged test sentences)	64.0	36.0	0
PoS-tags + function words (99 uniquely tagged test sentences)	64.6	35.4	0

Table 14: Selection by the bigram model based on different types of syntactic category information[40]

Contrary to expectations, the mixed representation resulted in lower performance than the full part-of-speech tagging for both object-gap and do-support PIRCs. This is of interest since it suggests that gains due to formation of larger groupings with increased

[40] For experiments 11 and 12, we only tested object-gap and do-support PIRCs which were troublesome in chapter 2.

bigram counts tend to outweigh gains due to the specific signposts to syntactic structure provided by the individual function words in the mixed representation.

Success rates ranged only from 41% to 70.4%, with a miss rate too high to be explained away as due to occasional performance errors by a learner that has mastered the basic rule. The object-gap PIRCs benefited by losing most of their 'undecided' cases, but where the model did make a choice, the proportion of cases that were decided correctly actually declined, from 70% (35 correct answers out of 50 choices made) in the original experiment, to 55% (54 correct answers out of 98 choices) for the PoS-tags only experiment and 46% (45 correct out of 97 choices) in the PoS-tags + function words experiment. (Paired *t*-test comparisons of object-gap PIRCs between the original corpus and the PoS-tags only corpus: $t(99) = 46.790$, $p < 0.0001$ and *t*-test comparisons between the original corpus and the PoS + function words corpus: $t(99) = 38.539$, $p < 0.0001$.) For the do-support PIRCs, both PoS representation schemes fared better (paired *t*-test comparisons between the original corpus and the PoS-tags only corpus: $t(99) = 40.865$, $p < 0.0001$; between the original corpus and the PoS-tags + function words corpus: $t(99) = 30.529$, $p < 0.0001$) than the purely word level representation of the original experiment, but performance still barely overtopped the 70% level that we have seen with some regularity in the previous experiments.

3.3.3. Analysis

The fact that part of speech information was not more strongly helpful is presumably because of the varied effects on bigram probabilities as corpus frequencies increase (from

the result of syntactic categories collapsing together), as noted above in discussion of Experiments 7 and 8 (see 3.1.3). This is documented in Table 15. Though this table is for only one illustrative example, it is representative of the set of sentences used. The only variation might come from the third (grammatical and ungrammatical) bigrams as their second unigram could be of different grammatical categories (adjective, preposition, determiner).

	Bigram1	Bigram2	Bigram3
(16) Grammatical	< *n v:aux&3S* > 88.3 + 86.7 = 175	< *v:aux&3S part-prog* > 25.1 + 49.5 = 74.6	< *part-prog adj* > 33.1 + 82.9 = 116
(16) Ungrammatical	< *n part-prog* > 74.8 + 49.5 = 124.3	< *part-prog v:aux&3S* > 0 + 86.7 = 86.7	< *v:aux&3S adj* > 31.4 + 82.9 = 114.3

Table 15: Distinguishing bigrams for a test sentence in the PoS-tags only experiment[41]

	Bigram1	Bigram2	Bigram3
(17) Grammatical	< *n is* > 84.3 + 76.9 = 161.2	< *is part-prog* > 28.3 + 48.4 = 76.7	< part-prog *adj* > 22.5 + 80.8 = 103.3
(17) Ungrammatical	< *n part-prog* > 73.2 + 48.4 = 121.5	< *part-prog is* > 0 + 76.9 = 76.9	< *is adj* > 35.4 + 80.8 = 116.2

Table 16: Distinguishing bigrams for a test sentence in the PoS-tags + function word experiment

As suggested by the tables, no 'marker' bigram is available for helping the selection and because corpus frequencies have increased (compared to the original experiment 1), it is more difficult to determine which bigram (grammatical or

[41] For tables 15 and 16, the probabilities were multiplied by 10,000 for better visualization.

ungrammatical) can benefit from the part-of-speech information. Therefore exact outcomes may vary with the particular set of categories employed.

The general conclusion to be drawn from these PoS experiments is that while the auxiliary inversion rule cannot be captured in terms of a sequence of specific lexical items, it also cannot be captured in terms of a sequence of lexical/syntactic categories. The results are compatible with the hypothesis that the latter format is more productive for capturing a syntactic dependency, but they do not strongly support the idea that part of speech information is sufficient for this purpose. Of course, better success rates might be obtained in future by use of a different set of PoS-tags than was employed here, either a more refined one with more categories or possibly a less refined one with fewer categories. But the general indication seems to be that some richer type of linguistic representation system would be needed. After all, the two loci that are related by the aux-inversion dependency are still not adjacent even when represented in part of speech notation, so they are presumably still beyond the scope of any bigram analysis. In the next chapter some success is achieved by means of a phrase-structural representation, which does have an effect on what counts as adjacent to what.

Chang et al. (2006) found a similar outcome when comparing different algorithms for representing syntactic categories: a more abstract representation of words is not necessarily more helpful for learning syntactic constraints. The task of Chang et al.'s model (*Lexstat*) was to reconstitute a sentence by predicting the next word from the pool of words that constitute that sentence (see 1.1 for more details). In order to select the next

word, Lexstat collected a “context statistic”, which was “akin to bigrams in computational linguistics” (2006, p. 2) and an “access statistic”, which approximates the activation level of a word in the lexicon. The algorithm was compared with four other algorithms involving progressively more abstract category representation. The “prevword learner” categorized the word based on the most frequent previous word. The “freqframe learner” followed Mintz’s frequent frames (2003) and categorized the word based on the most frequent frame surrounding it. The last variations were the “Token/Type learner” and “Type/Token learner” which categorized the word with the frame that had respectively the highest lexical diversity and the lowest one. The results showed that Lexstat performed best and therefore that the model was “better able to characterize the order of words in child and adult speech using more specific categories rather than broad categories” (2006, p. 4).

3.4. Experiment 13: Enriching the learning model with trigrams

3.4.1. Motivation and hypothesis

Augmenting the original corpus as in the previous experiments yielded some improvement but did not result in high levels of discrimination. Before turning in the next chapter to additional enhancements of the input that might boost performance further, the possibility needs to be checked that it is not the input itself that is holding back performance levels, but the use that the bigram model makes of the information that the input contains. Therefore, another route to try in the search for simple statistical learning of PIRCs is to upgrade the learning model. Once again, it makes sense to do this in small increments, in order to see whether there is a clear turning point. Also, a primary point of

interest in R&C's original finding was that correct discrimination was achieved by one of the most simple data-driven devices imaginable. To preserve this aspect of their project, we upgraded their bigram model, in the first instance, just as far as a trigram model. (See sections 1.1 and 2.6 for a brief discussion of the considerably more powerful neural network models.)

Experiment 13 reverted to the original untagged corpus but the learner was trained on trigrams instead of bigrams. Trigrams span three words rather than two. Since they have a broader scope than bigrams, it could be hoped that trigram statistics might pick up more information from the corpus for the difficult object-gap and do-support varieties of PIRC. R&C had run a trigram model over the same test sentences and corpus that they used for the bigram model. The results showed the same successful performance level as with the bigram model (96% of correct predictions). Many other enhancements of the learning algorithm could be tried as well (e.g., different smoothing techniques, or more elaborate computations over bigrams (see Chang's Lexstat, 2005)). This modest upgrade from bigrams to trigrams should be seen as a first step forward, in keeping with the research strategy of incremental supplementation of the learning situation.

3.4.2. Results and Analysis

In fact, trigram-based discrimination proved to be no more successful than bigram-based performance for the more resistant varieties of PIRC. Table 17 below shows that for object-gap and do-support PIRCs the trigram and bigram results were very similar, with respect to both percent correct and percent undecided.

SUBJECT-GAP PIRC	**% correct**	**% incorrect**	**% undecided**
Experiment 1: bigrams	87	13	0
Experiment 13: trigrams	80	20	0
OBJECT-GAP PIRC	**% correct**	**% incorrect**	**% undecided**
Experiment 1: bigrams	35	15	50
Experiment 13: trigrams	36	16	48
DO-SUPPORT PIRC	**% correct**	**% incorrect**	**% undecided**
Experiment 1: bigrams	49	51	0
Experiment 13: trigrams	48	52	3

Table 17: Selection by the trigram model for different varieties of PIRCs

The explanation appears to be that the broader scope of the trigram statistics was counterbalanced by the lower corpus frequency of trigrams in the test sentences. Because the Bernstein-Ratner corpus is quite small, it was more difficult to match a sequence of three words; many of the test sentence trigrams were unattested and their probability therefore had to be estimated by smoothing. For instance, only 1.88% of the trigrams in the object-gap test sentences occurred in the corpus, compared with 11.83% of the bigrams. For example, the bigram <*he is*> occurs 30 times while the trigram <*cookie he is*> (in the grammatical test sentence *is the cookie he is making in the bowl?*) doesn't occur; the likelihood of a sequence of noun, pronoun and auxiliary is obviously lower than the likelihood of a sequence of just a pronoun and an auxiliary. When a trigram does not occur, the smoothing factor in equation 1 substitutes half of the smoothed bigram probability for the second bigram in the trigram. Because of this, the results of the trigram analysis differ very little from those of the corresponding bigram studies.

This suggests that the move from bigram to trigram statistics would be more beneficial in combination with a very much larger corpus that contains a higher

proportion of the test sentence trigrams. There is no guarantee that this could work. Most trigrams contain at least one open-class word (triples of closed class words such as "*is in the*" do occur but less often), so a typical trigram is unlikely to be very frequent even in a large corpus. Whether it could nevertheless succeed was investigated in the next pair of experiments.

3.5. Enriching the corpus and the learning model: Experiments 14 & 15

3.5.1. Methodology and results

We have seen that increasing separately the corpus size and the scope of the learning model did not substantially improve the bigram model's performance on the more challenging PIRC varieties. A logical follow-up, to provide just one more chance for success, would be to combine all the new resources that have been tried in previous experiments. Therefore, we ran two further experiments, testing a trigram model on the larger/older corpus of child-directed speech, untagged, and then on that corpus with PoS tagging. The methodology was the same as previously. The results are in Table 18.

SUBJECT-GAP PIRC	**% correct**	**% incorrect**	**% undecided**
Large corpus + trigrams	71	29	0
OBJECT-GAP PIRC	**% correct**	**% incorrect**	**% undecided**
Large corpus + trigrams	54	39	7
Large tagged corpus + trigrams	90	10	0
DO-SUPPORT PIRC	**% correct**	**% incorrect**	**% undecided**
Large corpus + trigrams	70	30	0
Large tagged corpus + trigrams	68	32	0

Table 18: Selection by the trigram model on different types of corpora[42]

[42] The tagged corpus was used previously to test only object-gap and do-support PIRCs, so it was not run against subject-gap PIRCs in these experiments either.

The combined escalation in corpus information and statistical power resulted in an impressive – though isolated – increase in success rate for the object-gap PIRCs only: 90% correct (paired *t*-test comparisons of object-gap PIRCs between the original corpus and the corpus combining all enrichments: $t(99) = 51.790$; $p < 0.0001$). By contrast, the do-support PIRCs stopped short at the by now familiar plateau of 70% correct. In line with our fundamental research policy of not only reporting but also explaining the results obtained under different conditions, it is important to understand why this sizeable improvement for object-gap PIRCs occurred, in order to assess whether it is merely fortuitous or whether it hints at a potentially more broad-based improvement if we were to continue along this path of enriching the input and the learning model simultaneously.

3.5.2. Analysis of object-gap PIRCs

An analysis of the distinguishing trigrams between the 2 test versions of objectgap PIRCs in the tagged corpus study showed that an overwhelming proportion of distinguishing trigrams in the grammatical test sentences occurred in the training corpus: 98% versus only 48% for the ungrammatical sentences (see Table 19), resulting in many correct selections of the grammatical version.

	100 grammatical test sentences	100 ungrammatical test sentences
Total of distinguishing trigrams	452	452
Total of attested Distinguishing trigrams	443 (98%) average probability: 0.1675	220 (48.67%) average probability: 0.1041

Table 19: Distribution of distinguishing trigrams for object-gap PIRCs

The reason for this overwhelming success is that among the attested trigrams for the grammatical version of the object-gap PIRCs there was a 'marker' trigram in the grammatical version: <*n v:aux&3S part-PROG*>. This trigram could be instantiated, for example, as the sequence of words <*sister is pushing*>, found in "Is the wagon your sister is pushing red?". Many of the object-gap PIRC test sentences shared this structure: *NP is* followed by a progressive participle, for practical reasons of sentence construction. The form of an object-gap PIRC is inherently constrained in some ways that do not apply to the subject-gap PIRC. The range of possible predicates in the relative clause is less rich in object-gap sentences because it is difficult to create a gap in object position if the predicate is a nominal or an adjective phrase. For instance: a subject-gap PIRC could have a copula *is* followed by a nominal predicate such as *a nurse*, as in: *Is your sister who is a nurse waving at you?* whereas this is not possible in an object-gap PIRC except with a prepositional phrase to provide a place for the gap: *Is the school that your mother is the principal of near the park?* Similarly, an adjectival predicate is fine in the relative clause of a subject-gap PIRC (*Is the dancer who is tall sitting down?*) but only rarely in an object-gap PIRC (*Is the kitty the dancer is mean to very fluffy?*). Because of this, the grammatical object-gap test sentences mostly had the *is* auxiliary followed by a progressive participle (e.g., *Is the kitty the dancer is brushing very fluffy?*), and in every case that sequence was preceded by the subject of the relative clause, usually a common noun. This is part of what created the recurrent 'marker' trigram which favored the grammatical selection for object-gap sentences. The other relevant factor is that the PoS tagger collapsed words with this sequence of syntactic categories into exactly the same trigram, yielding many instances of <*n v:aux&3S part-PROG*> in Experiment 15. This

contrasts with Experiment 14 where the sentences were not converted to PoS tags, so the relative clause varied from one test sentence to another (e.g., *sister is pushing*, *cat is watching*, etc.).

Together, these factors resulted in 65 instances of the <*n v:aux&3S part-PROG*> trigram in the grammatical test sentences (and none in the ungrammatical version where the auxiliary was moved out from the relative clause). Another relevant characteristic of this distinguishing trigram is that it appeared frequently in the training corpus. Indeed, it corresponds to the familiar sequence of a subject followed by a progressive verb complex. This pattern is quite common in English and because the size of the training corpus was increased in this experiment, the trigram received a boost in its probability: 0.3812 (it was ranked 12th in terms of probability out of 172 trigram types found among the 904 distinguishing trigrams in the test sentences).

Thus, although the <*n v:aux&3S part-PROG*> trigram was not found in every grammatical test sentence (as the <*who|that is*> bigram was in earlier experiments), it did have the characteristics of a 'marker' trigram: it appeared with some consistency in the grammatical test sentences and never in the ungrammatical ones, and it was associated with a high corpus probability. These are the optimal conditions for successful discrimination (as outlined in 2.4.2) and this explains why the trigram model made correct selections for the great majority of object-gap PIRCs. It is also clear why this success was limited to the object-gap sentences and did not extend to the do-support sentences. The relative clause in a do-support PIRC has a characteristic trigram <*n*

pro:wh v&3S> (e.g. from *boy who plays*) under PoS coding. However, the ungrammatical sentence has a similar trigram *<n pro:wh v>* (corresponding to *boy who play*). In fact neither of these is attested in the corpus. Unlike the marker trigram in the object gap sentences, these trigrams contain a relative pronoun, making the sequence improbable. For an unattested trigram, the smoothing factor which substitutes for the trigram probability is half of the bigram probability of the second bigram in the trigram, which in this case would be *<pro:wh v&3S>* for the grammatical test sentence and *<pro:wh v>* for the ungrammatical test sentence. As discussed in relation to Experiment 11, the latter is slightly favored by the greater morphological freedom of the plural verb, and so there is a fair number of incorrect selections.

3.6. Conclusion for chapter 3

To summarize the collective results reported in this chapter: To achieve optimal performance, each PIRC type displays a preference for some particular testing conditions, but there is no one set of conditions in which PIRC learning across the board yields results that are comparable to what human learners apparently achieve. There is, unfortunately, a dearth of psycholinguistic data concerning success rates on this discrimination task even by adults, and certainly no data (apart from the limited Crain & Nakayama studies) which tracks the developmental course of this ability, for each variant of the PIRC construction. A question of particular interest is whether all PIRC varieties develop at a similar pace, which would suggest that learners have detected the general pattern which applies in them all, rather than relying on particular cues that happen to be available in one variety or another. Despite this regrettable lack of psychological data at

present, informal observation suggests that by the time learning is complete, a normal native speaker of English can discriminate between grammatical and ungrammatical PIRCs of all varieties with near 100% accuracy. If there is variation among them, it seems most likely to be due to a child's familiarity or lack of familiarity with a particular auxiliary verb and its selection properties, which dictate the required form of the non-finite matrix predicate which follows the complex noun + RC subject phrase. (See Richards (1990) and Ambridge et al. (2008) for some data on knowledge of auxiliaries in child language.)

The n-gram studies, across the many conditions experimented with here, show obvious signs of dependence on "accidental" local cues for particular constructions and in particular learning circumstances. As a result, success rates varied considerably; they were at their highest in three different circumstances:

(a) For is-is subject gap PIRC: rather small corpus and bigrams

(b) For is-is object gap PIRC: large tagged corpus and trigrams

(c) For do-support PIRC: WSJ corpus/large corpus and bigrams

In almost all cases, the success rate appears to reach a plateau at approximately 70%, suggesting that this is the maximum level of performance that can be expected from a simple n-gram statistical learning device for any corpus size and any level of corpus sophistication/age of child addressed – except when there happens to be some serendipitous cue in a particular case, such as the <*who*|*that is*> bigram of the initial experiments. This all adds up to a powerful body of evidence that using a statistical learning model which tracks only adjacent words or word categories (even with large

amounts of data) is not a realistic learning mechanism for natural languages, at least with respect to the PIRC construction, which has been widely regarded in the literature as a critical indicator of the poverty versus richness of the stimulus.

The next and final chapter pays closer attention to the consensus in linguistic analyses, which is that a reliable rule for auxiliary inversion in English questions cannot be stated without some reference to hierarchical phrase structure. If this is correct, then it seems that an n-gram model must fail, since it emphasizes linear adjacencies between sentence elements; it could succeed in capturing the linguistic facts only if it were to apply at several levels, building up – somehow – a hierarchical sentence structure in which the true linguistic dependency in a PIRC construction could be captured.

Chapter 4: Structure Dependence

4.1. Motivation

The previous experiments have shown that a simple probabilistic learning model trained on sentences represented with only a flat (linear) structure cannot reliably capture complex dependencies such as the filler-gap relationship in PIRCs. Providing more or richer input can improve performance, but the improvements are not systematic; they are more evident for some varieties of the construction than for others. There were no circumstances in which all three sub-types of the auxiliary-inversion construction were well-discriminated. Together, these data were interpreted as indicating that the n-gram models were not capturing the linguistic generalization that unites all instances of auxiliary-inversion. In this chapter we consider the possibility that the weak results of the previous experiments are not unexpected, given that none of the learning contexts provided any information about phrase structure.

Chomsky's conclusion (1980 and since) from the facts about PIRCs was that innate knowledge of the "structure-dependence" of syntactic transformations is essential for solving the puzzle of which auxiliary should be fronted by the auxiliary-inversion transformation. What he meant by "structure-dependence" is simply a sensitivity to the hierarchical phrase structure of sentences. The linear order of words cannot predict what moves to where in natural language sentences; it is phrases that move, and they move from one structural context to a particular structurally defined position. Interestingly, this is echoed in much more recent work in a non-transformational framework. Clark and Eyraud (2006) argue that the generalization about auxiliary-inversion is easily addressed

in a context free phrase structure grammar which characterizes polar interrogatives as having the (surface) structure *Is NP predicate*. (Generalization to other auxiliaries besides *is* would obviously be needed.) All that is required is that the learner can establish that a simple sequence like *the boy* and a complex one like *the boy who is crying* are inter-substitutable in English sentences. Distributional analysis as advocated by Zelig Harris (1954) would show that these two word sequences are substitutable in some contexts, e.g., the context -----*is hurt* in declarative examples. By extrapolation, this type of learner will assume inter-substitutability of *the boy who is crying* and *the boy* in other contexts, such as *Is----hurt?* in interrogatives. Therefore, an encounter with the question *Is the boy hurt?* will lead the learner to infer that *Is the boy who is crying hurt?* is also grammatical. This substitution-test approach to syntax acquisition will not lead to the inference that *Is the boy who crying is hurt?* is grammatical, because *the boy who crying* is not in the input and hence is not in a substitution class with any other word sequence.

In a domain of phrase structure grammars, therefore, the choice of the correct rule for 'auxiliary-inversion' (which of course involves no actual movement operation) is now simplified to the choice between S → Is NP B? (with B being instantiated by a verb phrase, participle, noun phrase, prepositional phrase, adjective phrase), versus S → Is X is B?, where X indicates a word sequence that is not (otherwise) observed in the language. The preferred choice therefore becomes trivial. Any data-based learner capable of tracking the ungrammaticality of X would favor the former. The ambiguity created for a transformational grammar learner simply disappears; the correct auxiliary is 'moved' because the preferred phrase structure rule does not even acknowledge the presence of

the embedded auxiliary[43]. However, this distributional approach also has its own limitations in terms of its applicability to psychologically-realistic models of language acquisition. Although Clark and Eyraud showed that PIRCs could be learned, they also conceded that "our demonstration is primarily mathematical/computational: we present a simple experiment that demonstrates the applicability of this approach to this particular problem neatly, but the data we use is not intended to be a realistic representation of the primary linguistic data, nor is the particular algorithm we use suitable for large scale grammar induction" (p. 1127)[44].

A more extensive attempt at combining structured representation with statistical inference was presented by Perfors et al. (2006). Their Bayesian[45] model had to select a grammar (between 3) that best fit a corpus of child-directed speech. The corpus was tagged and split into smaller corpora with increasingly complex syntactic structures. The 3 grammars were hand-crafted so that they could parse all the sentences in the corpus.

(a) A "flat grammar" containing a list of the sentences from the corpus (word strings)

(b) A "regular grammar" which is similar to a Markov model (a set of states along with their transitional probabilities)

(c) A Probabilistic Context-Free Grammar (PCFG)

---

[43] It should be noted that a compact and explanatory phrase structure grammar would need to restrict the category B in the rule *Is NP B?* to a category that can appear in the declarative context *NP is B*, as current extensions of phrase structure grammars do (see Pollard & Sag, 1994). Similarly for other fronted auxiliaries; for example, in *Must NP Y?* the category of Y must be one that is selected by the modal *must* in declaratives.

[44] There exist other models proposed in the literature that assign phrase-structure to sentences (e.g. Chater & Manning, 2006) but these make too heavy use of computational resources to be considered psychologically realistic models of language acquisition.

[45] A Bayesian model combines the likelihood of the data (language input) with the prior probability of the grammar to select among different grammar hypotheses.

Perfors et al. tested how well each grammar accounts for the set of sentences in the corpus, modulated by a simplicity measure: grammars with fewer productions and symbols were assigned higher prior probability. They found that although the context-free grammar was descriptively richer, it was not favored for the smallest corpus on simplicity grounds (it required 17 productions and 7 non-terminals to parse only eight sentences). Therefore, the Bayesian model picked the flat grammar for the simpler corpora and the PCFG for the more complex ones. Perfors et al. observed that "the context-free grammar always shows the highest level of generalizability" (p. 5). Though it was not the original aim of their research, they also noted that "PCFGs also generalize more appropriately in the case of auxiliary fronting" (p. 5).

Thus, whether the syntactic framework is transformational as when Chomsky originally formulated his poverty of stimulus argument based on PIRCs, or is overtly phrase structural as in current challenges to Chomsky's argument, the consensus is that sensitivity to phrase structure is essential for correct formulation or application of a rule for auxiliary inversion in complex sentences. It could very well have been anticipated, therefore, that no learning system attuned merely to word sequences could possibly succeed on PIRCs. However, Reali and Christiansen claimed to have evidence to the contrary. The issue they raised – acquisition of sophisticated linguistic knowledge by simple statistical mechanisms from low-level input data – is a very important one, with major consequences for our understanding of what makes human language possible. So it needed to be checked out, as it has been in the experiments reported so far in this dissertation. Clearly, the simplest form of this claim has now been disconfirmed. But

more elaborate versions might nevertheless succeed, so the next step in our pursuit of how minimal a syntax-learning model could be must be to investigate the prowess of simple statistical learners that are able to benefit from structural information of the kind that is agreed on all sides to be relevant to solving the auxiliary-inversion problem.

Therefore, we now explored the bigram model's ability to discriminate between grammatical and ungrammatical PIRCs when phrase-structure information was injected into the training corpus. We deliberately retained the low-powered bigram computational system used in the previous experiments, in order that our results should isolate the benefits of phrase structure information per se. We report below two experiments which combined a bigram learning model with hierarchical representations of sentences. Also in line with our escalation methodology, we began at the simplest level that might yield positive results. Rather than represent a full-fledged phrase structure, which could overwhelm a simple n-gram model, we encoded only NP constituency information. As explained below, we did this by introducing NP brackets into the word string, and alternatively by replacing all noun phrases by the designation NP.

Before examining the results of these experiments, we should consider the psychological plausibility of the underlying assumptions. How could an n-gram learner have access to the sorts of phrase structure information that we anticipate are essential for solving the auxiliary-inversion problem? If we suppose that language learning is a multi-stage process, we could imagine that the learning mechanism goes through a sequence of n-gram analyses, each feeding into the next and at increasingly abstract levels. Statistical

analyses of relations between words might yield part of speech categories. Then a further round of analysis of strings in the form of part of speech tags might identify phrasal constituents such as NP. It should be noted here that no such layered n-gram model has been reported in the literature, to the best of our knowledge. But its potential is nevertheless worth considering, as the only promising way in which an n-gram computational device could exhibit structure dependence.

For the following experiments (Experiments 16 and 17), we employed the original Bernstein-Ratner corpus of the earlier experiments, in order to be able to make comparisons between the new results and those from the previous experiments, isolating the role of phrase structure (PS) information. Furthermore, the NP bracketing/labeling had to be done manually so 10,000 sentences was a realistic corpus size.

4.2. Experiment 16: Adding NP brackets in PIRCs

4.2.1. Methodology

In the corpus, NP brackets were inserted surrounding all noun phrases. Left and right brackets were distinguished. An example of a tagged sentence in the corpus is shown:

(18) Can $_{NP}$[you]$_{NP}$ feed $_{NP}$[it]$_{NP}$ to $_{NP}$[the doggie]$_{NP}$?

The test sentences (identical to the ones used in the original experiment) were tagged as shown:

(19) a. Is $_{NP}$[$_{NP}$[the little boy]$_{NP}$ $_{NP}$[who]$_{NP}$ is crying]$_{NP}$ hurt?

b. * Is $_{NP}$[$_{NP}$[the little boy]$_{NP}$ $_{NP}$[who]$_{NP}$ crying]$_{NP}$ is hurt?

For purposes of the bigram analysis, each bracket was treated exactly like a word in the string. Apart from the introduction of NP brackets into the corpus and test sentences, the remaining steps of the experiment were identical to those reported above. The statistical learner used was the bigram model, but a bigram now consisted of two adjacent items which might be words and/or labeled brackets. For example, one bigram would be <*is hurt*> while another would be <$]_{NP}$ *hurt*>. It is also important to note that the tagging system in this experiment did not distinguish between well-formed NPs such as *the boy who is crying*, and ungrammatical NPs such as *the boy who crying.* This was based on the assumption that whatever lower-level procedure was responsible for identifying NPs would not be capable of making these fine distinctions based on the detailed internal structure of constituents. Alternative assumptions could be made (see experiment 18 below) but, in this experiment, both of these word sequences were simply tagged as NP. The consequence of this is that discrimination of the grammatical form could not rest simply on the absence of any exemplars in the corpus for the ungrammatical subject NP but must involve the relation between that NP and the rest of the matrix clause.

4.2.2. Results

The results differ strikingly from the original experiments with only linear sequences of words. With NP brackets, the bigram learner favored the ungrammatical sentences over the grammatical ones.

Word string with NP brackets added	Sentences tested	% correct	% incorrect	% undecided
Is-is subject-gap PIRCs	100	31	62	7
Is-is object-gap PIRCs	100	37	43	20
Do-support PIRCs	100	45	55	0

Table 20: Selection by the bigram model in for Experiment 16

As Table 19 shows, the patterns of discrimination were not quite the same across the sub-varieties of PIRC, but in all cases the ungrammatical form was chosen more often than the grammatical form (paired *t*-test comparisons of subject-gap PIRCs between the original corpus and the bracketed corpus: $t(99) = 3.286$, $p < 0.0001$; of object-gap PIRCs between the original corpus and the bracketed corpus: $t(99) = 6.334$, $p < 0.0001$; of do-support PIRCs between the original corpus and the bracketed corpus: $t(99) = 5.118$, $p < 0.0001$). This could be seen as a surprising result, since the learning model was provided with richer phrase structure information than in the previous experiments where the bigram model was sometimes more successful. Why did the phrase structure information provided not facilitate learning?

4.2.3. Analysis

4.2.3.1. Is-is PIRCs

There are four distinguishing bigrams for this construction. For example, for the pair of test sentences in (20), the distinguishing bigrams are as shown in table 20.

(20) a. Is $_{NP}$[$_{NP}$[the little boy]$_{NP}$ $_{NP}$[who]$_{NP}$ is crying]$_{NP}$ hurt?

b. * Is $_{NP}$[$_{NP}$[the little boy]$_{NP}$ $_{NP}$[who]$_{NP}$ crying]$_{NP}$ is hurt?

Test sentences	Bigram1	Bigram2
Grammatical	*<is crying>*	*<]$_{NP}$ hurt>*
Ungrammatical	*<]$_{NP}$ crying>*	*<is hurt>*

Table 21: Distinguishing bigrams for the test sentence pair (20).

	Bigram1	Bigram2
Mean probability for grammatical sentences	*<is A>* = 0.0064	*<]$_{NP}$ B>* = 0.0410
Mean probability for ungrammatical sentences	*<]$_{NP}$ A>* = 0.0032	*<is B>* = **0.1384**

Table 22: Mean probabilities of the distinguishing bigrams in subject-gap PIRCs

The *<is B>* bigram generally had the highest probability and it favored the ungrammatical sentences (hence 62% incorrect selection). Because of the way the test sentences were constructed, A was never an NP[46] while B was sometimes instantiated by an NP (and as well by VP, PARTICIPLE, PP, ADJP, etc.). For that reason, the *<is B>* bigram was realized as *<is $_{NP}$[>* in some ungrammatical test sentences. Because the *<is $_{NP}$[>* bigram is composed of two adjacent unigrams which both rank among the highest probabilities, its bigram probability was high. This particular "strong" bigram boosted the overall probability of the second distinguishing bigram *<is B>* in the ungrammatical sentences (0.1384) and therefore favored the selection of the latter over the grammatical ones.

[46] For reasons not clear to us, all of R&C's test sentences had A instantiated as a progressive participle, and for comparability of outcomes we matched our materials to theirs. Note that this was the case in all the experiments reported in the dissertation. In future work this arbitrary restriction could be lifted, along with the R&C template's restrictions on the choice of auxiliaries and other aspects of the test sentences (section 2.2).

4.2.3.2. Object-gap PIRCs

An example of a pair of NP-tagged test sentences is shown in (21) and its distinguishing bigrams are shown in Table 22.

(21) a. Is $_{NP}$[$_{NP}$[the wagon]$_{NP}$ $_{NP}$[your sister]$_{NP}$ is pushing]$_{NP}$ red?

b. * Is $_{NP}$[$_{NP}$[the wagon]$_{NP}$ $_{NP}$[your sister]$_{NP}$ pushing]$_{NP}$ is red?

Test sentences	Bigram1	Bigram2
Grammatical	<*is pushing*>	<*]$_{NP}$ red*>
Ungrammatical	<*]$_{NP}$ pushing*>	< *is red*>

Table 23: Distinguishing bigrams for the test sentence pair (21)

	Bigram1	Bigram2
Mean probability for grammatical sentences	<*is A*> = 0.0002	<*]$_{NP}$ B*> = 0.004
Mean probability for ungrammatical sentences	<*]$_{NP}$ A*> = 0.0002	<*is B*> = **0.0105**

Table 24: Mean probabilities of the distinguishing bigrams in object-gap PIRCs

In this experiment, the patterns of the distinguishing bigrams were identical to the ones observed in the subject-gap PIRCs (the data differ because of the variety of lexical items used to construct the object-gap versus the subject-gap test sentences, thus resulting in different bigram probabilities). Because A was only instantiated as a progressive participle, the bigrams containing A were relatively unlikely to occur and therefore their probabilities were low. On the other hand, because B was sometimes instantiated by an NP, as mentioned for the previous experiment, it raised the mean probability of the <*is B*> bigram (0.0105) hence discrimination towards the ungrammatical sentence was again favored.

4.2.3.3. Do-support PIRCs:

The example that was used in the original do-support experiment (section 2.5.2) is now tagged as:

(21) a. Does $_{NP}$[$_{NP}$[the boy]$_{NP}$ $_{NP}$[who]$_{NP}$ plays $_{NP}$[the drum]$_{NP}$]$_{NP}$ want $_{NP}$[a cookie]$_{NP}$?

b. *Does $_{NP}$[$_{NP}$[the boy]$_{NP}$ $_{NP}$[who]$_{NP}$ play $_{NP}$[the drum]$_{NP}$]$_{NP}$ wants $_{NP}$[a cookie]$_{NP}$?

Test sentences	Bigram1	Bigram2	Bigram3	Bigram4
Grammatical	<]$_{NP}$ *plays*>	< *plays* $_{NP}$[>	<]$_{NP}$ *want* >	< *want* $_{NP}$[>
Ungrammatical	<]$_{NP}$ *play* >	< *play* $_{NP}$[>	<]$_{NP}$ *wants* >	< *wants* $_{NP}$[>

Table 25: Distinguishing bigrams for the test sentence pair (21)

The analysis is similar to the original do-support experiment. There are 8 distinguishing bigrams but none that systematically occurs in either version of the test sentences. There are no undecided cases because the smoothing factor is non-identical for two pairs. Only the pairs *bigram2-grammatical* and *bigram2-ungrammatical*, and *bigram4-grammatical* and *bigram4-ungrammatical* could cancel out. Hence the two sentence versions will always have different probabilities. The bigram model has no particular bigram(s) that it can count on to favor either the grammatical or ungrammatical version so discrimination can be expected to be roughly at chance, influenced by the particular frequencies of the bigrams in individual test sentences. In the experiment with NP brackets, there are 45% correct selections versus 55% incorrect.

Overall, adding NP brackets to noun phrases turned out to be ineffective, despite the linguistic reasons for expecting that phrase structure information should be relevant to identifying the PIRC rule. A large part of the explanation probably lies in the fact that paired brackets never fell into the same bigram, so the phrase structure information they

provided was fractured. Also, they destroyed adjacencies between the original lexical items, so that any lexical co-occurrences that were helpful previously were now lost.

4.3. Experiment 17: Replacing noun phrases by NP tags

4.3.1. Methodology

In this experiment, noun phrases in the corpus and the test sentences were replaced by the symbol NP. For example, this sentence in the corpus: *Can you feed it to the doggie?* became: *Can NP feed NP to NP?*. The test sentences: *Is the little boy who is crying hurt?* and **Is the little boy who crying is hurt?* were respectively recast as follows:

(22) a. is NP hurt?

b. *is NP is hurt?

Note that the whole string *the little boy who is crying* was replaced by the NP symbol. This is because when one noun phrase appeared inside another it was impossible to replace both of them by the NP symbol, and we applied a general criterion of resolving the conflict by replacing the larger phrase (providing the highest level phrasal information).

This experiment differed from all the preceding ones in that the ungrammatical sentence had one extra word than the grammatical sentence. The extra word was only found in is-is subject-gap and object-gap sentences. The do-support construction was not affected since the difference between the grammatical and ungrammatical versions lay in the ending attached to the lexical verb. For instance, a pair of do-support test sentences was:

(23) a. Does NP want NP? (stands for *does the boy who plays the drum want a cookie*?)

b.* Does NP wants NP? (stands for **does the boy who play the drum wants a cookie*?)

Where there was a disparity in the number of words, an adjustment for sentence length was needed in the computation of cross-entropies, in order to permit a meaningful comparison across the sentences. We opted for the two following methods to control for sentence length. All remaining steps were identical to all the previous experiments.

(1) Like R&C: the cross-entropy was calculated by dividing the log probability of a sentence by the number of words in the sentence;

(2) Like Chen and Goodman (1996): the cross-entropy was calculated by dividing the log probability of a sentence by the number of bigrams in the sentence.

4.3.2. Results

The results were very positive, regardless of which adjustment method was used.

NP-replacement: sentence-length adjustment method (1)	**sentences tested**	**% correct**	**% incorrect**
Is-is subject-gap PIRCs	100	91	9
Is-is object-gap PIRCs	100	96	4
Do-support PIRCs	100	86	14
NP-replacement: sentence-length adjustment method (2)	**sentences tested**	**% correct**	**% incorrect**
Is-is subject-gap PIRCs	100	97	3
Is-is object-gap PIRCs	100	98	2
Do-support PIRCs	100	86	14

Table 26: Selection by the bigram model for Experiment 17

4.3.3. Analysis

4.3.3.1. Is-is subject-gap and object-gap PIRCs

Because the subject of the main clause, including the relative clause inside it, has been replaced by NP, both subject-gap and object-gap is-is forms present the following pattern for the grammatical version: Is NP B, where B can be instantiated by a verb phrase, participle, noun phrase, prepositional phrase, adjective phrase. For instance, a subject-gap PIRC "is the baby who is already asleep hungry?" and an object-gap PIRC such as "is the bird Annie is feeding hungry?" became "is NP hungry" in this experiment. Therefore, the analysis that follows applies to both constructions.

As stated earlier, the ungrammatical sentence contained one more word than the grammatical version and in consequence it also contained one more distinguishing bigram. The grammatical sentence always had the form *Is NP is B*, compared with the grammatical form of *Is NP B.*

Test sentences	Bigram1	Bigram2
Grammatical	< *NP hurt* >	
Ungrammatical	< *NP is* >	< *is hurt* >

Table 27: Distinguishing bigrams for the test sentence pair (22)

	Bigram1	Bigram2
Mean probability for grammatical sentences	<*NP B*> = 0.0524	
Mean probability for ungrammatical sentences	<*NP is*> = 0.0264	<*is B*> = 0.147

Table 28: Mean probabilities of the distinguishing bigrams in subject-gap PIRCs

	Bigram1	Bigram2
Mean probability for grammatical sentences	<*NP B*> = 0.0034	
Mean probability for ungrammatical sentences	<*NP is*> = 0.0264	<*is B*> = 0.0066

Table 29: Mean probabilities of the distinguishing bigrams in object-gap PIRCs[47]

Note in Tables 28 and 29 that there is a 'marker' bigram <*NP is*> that contains no specific lexical item and that appeared in every ungrammatical test sentence. This would suggest that the ungrammatical version would be selected most often. In fact the opposite is true. Thus, a question that needs to be addressed is the role of this bigram <*NP is*> in the cross-entropy of the ungrammatical sentence. It is surely a very common bigram in the corpus, yet it decreased rather than increased selection of the ungrammatical sentences. Of the 4629 bigrams of the corpus, the <*NP is*> bigram was only ranked 1909th in terms of its probability. Though the <*NP is*> bigram was composed of 2 frequent 'words', its probability was not as high as might be expected (rather low at 0.0264) because the probability was computed by dividing the count of bigrams (420) by the count of its first unigram "NP", which was the most frequent unigram in the corpus with 11,421 occurrences[48]. The bigram probability may be thought of as the probability of the second unigram, given the first unigram. So in the end, the <*NP is*> bigram which was found in every ungrammatical test sentence, did not play a key role in the selection task because its probability was not high enough to influence the discrimination in one way or another.

[47] As noted above, the mean probabilities varied between object-gap and subject-gap PIRCs because of the choice of the lexical items used in the test sentences.

[48] Note that the situation is different from the <*is* $_{NP}$*[*> bigram above. Both bigrams occur often in the corpus: 467 times for <*is* $_{NP}$*[*> and 420 times for <*NP is*>; however, the former had a higher probability than the latter because it was computed dividing the count of bigrams by the count of the first unigram *is* (641), which is significantly lower than the count of *NP*, the first unigram of <*NP is*>.

Another major reason why the grammatical sentence was favored over the ungrammatical one lies in the fact that the latter has one extra word than the former, resulting in one extra bigram in the ungrammatical version. A sentence probability is computed by multiplying the probabilities of the bigrams that constitute that sentence. Because each bigram probability is lower than 1, the higher the number of bigram probabilities that are multiplied, the lower the sentence probability is. For instance, imagine a sentence (a) with 3 bigrams of equivalent probability: 0.01 each. The sentence probability for (a) is therefore 0.1*0.1*0.1 or 0.001. This can be compared with a sentence (b) which has 4 bigrams with identical probability as the ones in sentence (a). The sentence probability for (b) is 0.1*0.1*0.1*0.1 or 0.0001, lower than the sentence probability for (a).

A look at the distinguishing bigrams for the subject-gap and object-gap PIRCs in Experiment 17 (Tables 28 and 29) confirms this mathematical truth. The difference in sentence probability between the grammatical and ungrammatical versions depends strictly on the distinguishing bigrams. Grammatical subject-gap PIRCs have only one distinguishing bigram with a mean probability of 0.0524 whereas the corresponding ungrammatical PIRCs have two distinguishing bigrams and their product 0.0039 (from 0.0264*0.147) yields a lower mean probability than 0.0524. A similar observation applies to object-gap PIRCs, where it is of interest that the distinguishing bigram in the grammatical version has a lower probability than either of the distinguishing bigrams in the ungrammatical version, but has a higher probability than their product.

However, though the bias towards the grammatical version is strong with NP-replacement in these is-is varieties, the basis for it is once again not well meshed with the relevant linguistic properties. In general it is not a good criterion, in discriminating grammatical from ungrammatical sentences, to select the one that is shorter. Moreover, the reason why the grammatical one was shorter in this experiment was the fact that only NP information was provided to the learning system (on the expectation that it should be most relevant to a linguistic discrimination). The grammatical sentence was shorter because the 'gap' where an aux would be in a declarative is in the VP, while in the ungrammatical sentence it is in the complex NP. Thus if we had provided VP information also, the two would have balanced out and sentence length would not have affected the learning model's decision.

4.3.3.2. Do-support PIRCs

For the do-support PIRCs the grammatical and ungrammatical versions were equal in length in the number of words, so the explanation above, based on just the number of words or bigrams involved, cannot explain the good outcome in this case. Some other factor(s) must be responsible.

A sentence like "*Does the boy who plays the drum want a cookie?*" became "*does NP want NP?*" and the corresponding ungrammatical version "**Does the boy who play the drum wants a cookie*?" became "*does *NP wants NP?*"

Test sentences	Bigram1	Bigram2
Grammatical	< *NP play*>	< *play NP*>
Ungrammatical	< *NP plays* >	< *plays NP* >

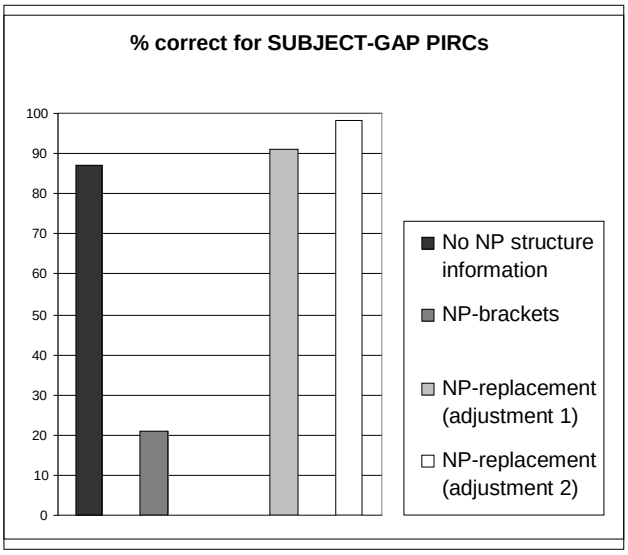
Table 30: Distinguishing bigrams for one pair of test sentences

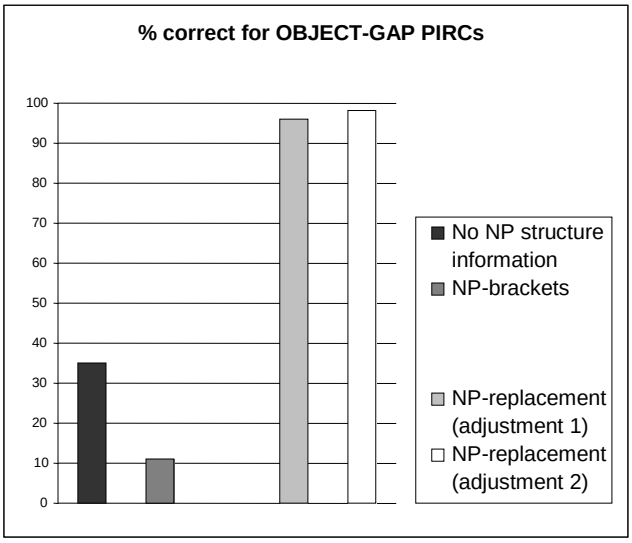
The second distinguishing bigram in the grammatical and ungrammatical versions might cancel out if they didn't occur in the training corpus because they share the same second unigram. They were relatively unlikely to occur because they contained lexical verbs. The discrimination thus depended mainly on the first distinguishing bigram. Here the relevant considerations mirror the discussion of Experiment 5 in Chapter 2[49]. Both the grammatical and ungrammatical bigrams had their second unigram instantiated as a lexical verb. However, the lexical verb in the grammatical sentence (for instance, here, *play*) could be finite or non-finite in the corpus, and if finite it could be first or second person singular or plural, or third person plural. By contrast, the verb in the ungrammatical version was only finite and only 3rd person singular. This morphological restriction reduced its chance to appear in the corpus. Because of this, the average probability of the first distinguishing bigram was several times higher in the grammatical sentence (0.00286) than in the ungrammatical sentence (0.00052). This explains why the bigram model quite strongly preferred the grammatical sentence. Note that this difference in frequency of the morphological form of the verb is not inherently related to the phrase structural properties of the sentences, so this is a somewhat fragile phenomenon even though it led to a successful outcome in this case.

[49] Though the reasoning is similar, it has more impact in Experiment 17 because the sparseness of the data which existed in the original small corpus is now reduced thanks to the syntactic clustering of all the noun phrases by the NP tag. The results of Experiment 11 where the corpus was enriched with part-of-speech information reflected some of that benefit: successful selection of do-support PIRCs reached 70% compared with 49% in the first do-support experiment.

4.4. What would solve the PIRC problem?

Let us now take stock of the outcomes of all these experiments involving encoding phrase-structure information. The graphs below compile the results of the phrase-structure experiments for purposes of comparison. The NP-bracketing experiment showed the worst performance of all, with no gain over the original experiment without phrase structure information.





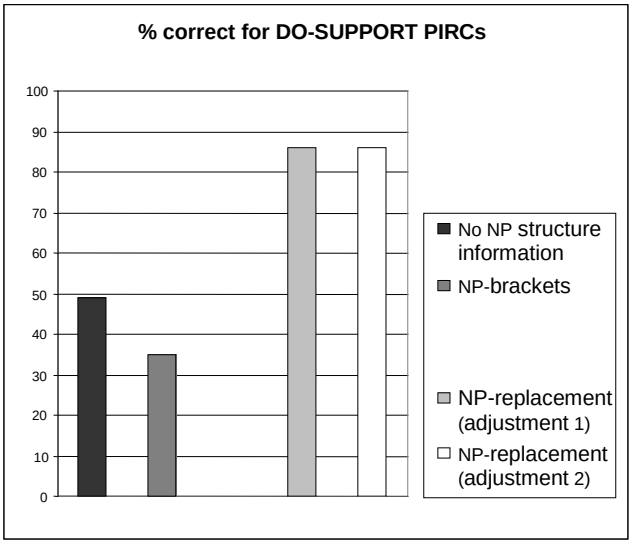


Figure 3: Correct choices by the bigram model using phrase-structure information

For the NP-replacement experiment, all three subvarieties of PIRCs were in the neighborhood of 90% correct, for the first time in the whole series of experiments reported in this dissertation. Thus, we have discovered one set of circumstances in which PIRC discrimination by a bigram model is generally good, which is of considerable interest.

It is important to compare the bigram model's reason for success with NP-replacement, with the reasons that would be anticipated on the basis of a linguistic analysis. One linguistically appropriate reason for rejecting the ungrammatical version is the presence of the extra auxiliary in the matrix VP, e.g. the second *is* in **Is NP is hurt*?. Since the scope of bigrams is limited to only two adjacent words, a bigram analysis cannot in principle capture the sequence: auxiliary NP auxiliary. A more powerful n-gram model (such as one that records trigrams) could do so, but it would be unlikely to improve performance because the trigram <*is NP is*> is not a valid sequence in English.

Due to its absence from the corpus, its probability would be computed by smoothing based on the final bigram within it (in the methodology of our project, derived from R&C's). That bigram could not capture the ungrammaticality, for the same reason as above, hence nothing would have been gained by the shift to trigrams. Some more radical revision of the bigram model would be needed to detect the incorrect duplication of the auxiliary.

The second linguistically authentic reason for rejecting the ungrammatical PIRC is the absence of an auxiliary in the relative clause E.g.: **who crying* in **Is the boy who crying is hurt?*. The PS-experiments 16 and 17 did not address this issue because they did not distinguish between the correct NP (*the boy who is crying*) and the ungrammatical *NP, i.e., the NP missing an auxiliary (**the boy who crying*). Yet this seems to be an ideal cue for a learner to rely on. So now we must consider: Can it be incorporated into a bigram model?

So far, in all our experiments, the test sentences contained only unigrams that occurred in the corpus. But **NP* is a unigram missing from the corpus. In order to incorporate it into an experiment, the **NP* unigram was coded as UNK (for "unknown") and the bigram model had to choose between a pair of sentence versions like:

(23) a. is NP hurt? (replacing a sentence such as *is the boy who is crying hurt?*)

b. *is UNK is hurt? (replacing a sentence such as *is the boy who crying is hurt?*)

Now a probability must be established for the unigram UNK. Several smoothing techniques exist to compute the probability of an unknown unigram. We opted for the Witten-Bell discounting technique (Witten & Bell, 1991) which computes the probability of the UNK unigram by dividing the number of word types found in the corpus by the number of word tokens added to the number of types. The motivation for the formula is the following: it uses the probability of words seen for the first time to model the probability of new/unknown words. Let us imagine a set of words (representing the corpus) and one word from that set is picked each time. There is a function whose role is to label a word as "new" if it has been picked for the first time. If a word had already been seen, the function would not be called for. In other words, the function will only be activated for every new word type encountered. The rest of the methodology and computations identically followed the previous experiments.

The results showed perfect performance, with 100% correct for all three types of PIRCs. This success is explained by the fact that the UNK unigram in the ungrammatical version has a very low probability and therefore lowers the sentence probability of the ungrammatical version, hence favoring the selection of the correct version by the bigram model. This is the first of our experiments to achieve 100% correct selection across the three types of PIRCs, using a rather small and early child-directed speech corpus but adding some partial phrase-structure information. So this is a direction that it may well be profitable to pursue in the search for a simple model that can discriminate complex constructions. However, if it is to be taken any further, the primary challenge is obviously to account for how the learning model could recognize the ungrammaticality of the

relative clause that lacks an auxiliary, and thereby could recognize the ungrammaticality of the NP containing it. In this last experiment, we simply provided the model with that information, but for a complete model it would be necessary to show that it could be achieved by an n-gram model, perhaps by building up phrase structure in successive steps as suggested above. We will now develop these issues in the concluding section.

4.5. General conclusion

This dissertation began by noting that recent studies with a data-driven learning stance have challenged the long-standing claim by Chomsky that the existence of innate linguistic knowledge (UG) is an essential foundation for acquiring language from the primary linguistic data available to children. This dissertation has focused specifically on the positive results obtained by Reali and Christiansen (2005) for bigram-based learning of a construction often cited as supporting Chomsky's position: PIRCs in English. R&C's results were originally construed as undermining Chomsky's poverty-of-the-stimulus claim. However, further data from experiments 1 – 6 presented here demonstrated that the striking success of the bigram model for English PIRCs is very limited; good performance was found only for subject-gap is-is PIRCs. When the model was tested on other instances of the PIRC construction (English object-gap and do-support PIRCs, lexical verb fronting in Dutch), the poor results revealed the limitations of the simple bigram model. It became clear that the model was choosing between sentence versions on a trivial basis, which did not tap into the essential linguistic properties of the PIRC construction. When it succeeded, it did so only thanks to an 'accidental' fact about is-is subject-gap PIRCs in English, a fact which <u>is</u> detectable on the basis of just two adjacent

words (*who* and *is*) as a bigram analysis can be sensitive to. This is why the successful discrimination was not – and could not be expected to be -- replicated for other variants of the PIRC construction in English or other natural languages. Not only did the bigram model not identify one discriminative cue that would work generally for all types of PIRC; it did not even identify a possibly different cue for each type.

However, rather than settling for this negative conclusion, it was deemed essential to verify that the failure on PIRCs other than the "is-is" PIRCs was truly due to the statistical simplicity of the bigram model and not due to a weakness of the learning resources. The original corpus used to train the model was very small, compared to what a child learner would be expected to be exposed to in the first two or three years of language learning. The corpus was therefore enriched in several different ways: by providing more sentences (experiments 7 – 9) or by substituting lexical items with syntactic categories (experiments 10 and 11). These maneuvers yielded little improvement, however. A modest upgrade of the learning model to trigrams also did not help at all; as before, the n-gram model achieved solid mastery of only one PIRC type.

Though augmenting the resources in these experiments did not achieve mastery of the auxiliary-inversion rule, they were nevertheless informative for our goal of circumscribing the minimal computational power needed to acquire a natural language. Previously, the model's weak performance led to only a disjunctive conclusion: either the stimulus (the child-directed speech) was not rich in relevant information or the model's statistical measures were too weak to extract it. The finding that enriching the

information source had relatively little impact strengthens the speculation that the computational resources of a bigram-based system are too limited to be able to take good advantage of the linguistic information the input does contain. On the basis of the bigram-based learning project, therefore, the challenge against UG as an essential foundation for language acquisition remains unsubstantiated.

The investigative strategy in this dissertation has been to power up the learning situation in small steps, since it could not be known in advance at what point there might be a break-through at which performance did reach satisfactory levels. An alternative research strategy would be to start at the other end of the scale, with the most powerful statistical induction procedures. If they fail, then of course the whole general approach of statistical learning of natural language would need to be abandoned; but if they succeeded, one might then strip off their resources in small steps to find out the minimum needs for success. As noted in the introductory chapter, neural networks constitute very powerful data-driven devices, representing a major advance in computational capacity. Studies by Lewis & Elman (2001) and Reali & Christiansen (2005) have shown that an SRN can learn is-is subject-gap PIRCs; other work with SRNs has shown success with some other complex syntactic constructions. Neural networks receive words of sentences in sequence. Their task is to predict after each word what the next word would be, based on the weights of the connections between nodes in the system. This is a more demanding task than the forced-choice whole-sentence task employed in the bigram model research. The accomplishment of neural networks on this task has been impressive and connectionists have argued that “if the network succeeds in making the correct

generalization, then it has succeeded by extracting the structure of NPs, and of *aux*-questions, from the statistical regularities in the data, and generalizing across NPs" (Lewis & Elman, p. 4). However, as demonstrated with the bigram model, the ability to deal with this one sub-variety of PIRC is very far from demonstrating knowledge of auxiliary-fronting in general. Thus the SRN results on PIRCs to date show little that is new, in view of the fact that SRNs have not been tested (so far) on any variety of PIRC beyond the is-is subject-gap form that the bigram model can handle. It would be interesting, therefore, to test whether SRNs can generalize their success on is-is subject-gap PIRCs, to similar structures like do-support and object-gap PIRCs. If learning was found to be successful in these two constructions, which have been shown to be more demanding for the bigram model, it would be important to understand how the process was accomplished, and the extent to which it reflects linguistic structure. Most importantly, does any aspect of the network's weights after learning encode some general property of auxiliary inversion? Or is it more a 'case-by-case' learning with the neural network picking up specific cues presented in each construction?

Infants have been shown to have remarkable abilities to detect transitional relationships (Saffran et al., 1996) but this has been demonstrated primarily at lower levels of language, such as within-word structure. Syntax is different: it is recursive and thus involves distributional relationships at multiple levels. The primary finding from the current work is that tracking simple dependencies between adjacent words is not sufficient by itself for syntax acquisition. What has been established is that more complex and powerful linguistic/learning models of some kind must be investigated.

One attempt to meet the greater challenge of syntax learning has been made here by adding phrase-structure information in experiments 16 and 17. Unless a statistical learner is capable of assigning phrase structure to word strings, then no increase in statistical power is likely to make a significant difference if the linguistic analyses developed for PIRCs are on the right track. In a much earlier version of transformational grammar, Chomsky's argument for a 'structure dependence' constraint on transformational rules entailed that the auxiliary to be fronted be defined in terms of height in the tree structure rather than in terms of linear order of words. Chomsky maintained that UG informs infants of this fact, since it could not itself be learned. In more recent developments of transformational theory, UG makes an even stronger contribution. It does not merely insist on some structural identification of the moving auxiliary, but specifically identifies it as a locality constraint: the auxiliary that moves is the one that is closest (in structural terms) to the landing site. Moreover, this is just one instance of a much more broadly applicable universal locality constraint ('heads must move locally'). On the innateness theory, then, there is essentially no learning needed at all for correct auxiliary-fronting, whereas in a data-driven approach all aspects of the auxiliary-inversion process must be learned from primary linguistic data, presumed to consist largely of linear word sequences (though some prosodic or semantic cues to syntactic grouping might be included in a learning model). To the extent that the earlier series of experiments failed to find impressive evidence of PIRC learning, it might be presumed that phrase structural analysis of word strings is a necessary condition, even if not a sufficient condition, for identifying the correct pattern for auxiliary inversion. Experiments 16 and 17 tested whether the bigram model was able to detect the auxiliary-

inversion rule after receiving some phrase-structure information in the training corpus. The outcome was very successful for experiment 17, suggesting that the solution to learning the auxiliary rule probably lies in the statistical model's ability to build some sentence structure. However, a number of problems remain to be resolved as this line of research is pursued.

One issue is how to represent phrase-structure information in a way that makes it usable by a simple learning model that tracks transitional probabilities, as the bigram model does. In keeping with our strategy of starting at the most modest level, in the goal of circumscribing the mental mechanisms by which learners acquire natural languages, our phrase structure experiments conducted to date did not involve a full phrase structure analysis, but encoded only noun phrases. We tried two different ways of representing the noun phrase structure: by bracketing or by replacement. These representation systems are formally equivalent, but they may well not be psychologically equivalent. This is especially so for a learning model that has a limited window of view, such as a bigram model. Bracketing requires the association of two symbols non-adjacent in the representation, which is not within the scope of a bigram model. As was not unexpected, the bracketing method proved to be inefficient. On the other hand, representing each syntactic phrase by means of a single symbol such as NP requires a multi-level representation, since one NP can appear inside another, without limit; phrasal structure is recursive (for instance, the NP "the pen in the drawer of the desk" contains several embedded NPs, "the pen", "the drawer", "the desk", "the pen in the drawer", "the drawer of the desk"). The bigram model accepts only strings as input, and a string representation

with NP-replacement is unable to cope with hierarchical structure at all: the embedding of one constituent inside another one. In Experiment 17, we opted for an NP-replacement representation that sacrificed low level phrasing information in favor of higher level: only the most inclusive NP was replaced with the NP symbol. But this did not give the learning model access to errors within that larger NP. Discrimination performance was very good, but was largely due to considerations of mere sentence length or structure-irrelevant properties of English morphology. If phrase-structure is the solution to the PIRC problem, the bigram model needs to be able to recognize at some point that "*the boy who crying*" is an ungrammatical NP, unlike its grammatical counterpart "*the boy who is crying*". As shown in 4.4, once the bigram model had access to this distinction between the correct NP and the incorrect NP (*NP), performance was 100% correct across all three sub-varieties of PIRCs. This finding represents a hopeful trend and suggests that future research might benefit from focusing on how this discrimination between the grammatical and ungrammatical NPs could be made by a simple statistical learner – presumably by the kind of multi-level analysis sketched above.

In one sense the series of experiments reported here has done no more than document what many linguists would have predicted with confidence all along. However, it has been worth generating the specific evidence on a detailed level which more precisely establishes what a simple n-gram model is capable of achieving. This is especially important in view of the strong current interest in statistical learning as a model of human language acquisition, and the successes it has had. The current project has made two primary contributions to that general line of research. First, it has

emphasized that it is not enough to show, for any learning model, just that it can make correct linguistic discriminations; it is as least as important to understand how it does so. This is relevant to assessing the likelihood that its prowess will extend more broadly to additional linguistic constructions, and also relevant to evaluating it as a psychologically plausible component of a human language learning mechanism. The second contribution towards future research is to propose the specific hypothesis that low-power statistics over word strings could be the basis for modeling syntax acquisition only by giving up the assumption that just one pass through each word string is sufficient to discover the kind of abstract linguistic patterning involved in constructions such as auxiliary-inversion. Rather, it would seem to be necessary to apply the statistics in a more intricate and multi-level fashion, building up layer by layer the hierarchical structure that is the basis of all syntactic patterns in human language.

Appendices

Appendix A: Test sentences[50]

1. "Is-is" subject-gap PIRCs

Is the boy who is generous feeding food to the squirrels?
Is the baby who is already asleep hungry?
Is the man who is tired of the game still playing?
Is the box of cereal that is open Cheerios?
Is the book that is so interesting about lions?
Is the girl who is smart wearing a jeans jacket?
Is the man who is sleepy drinking tea?
Is the kitten that is very cute Annie's?
Is an elephant who is old still kind?
Is the game that is harder more interesting?
Is the soup that is quite delicious made by grandma?
Is the bedroom that is neat his friend's?
Is the tree that is high hard to climb?
Is the boy who is afraid of dogs here today?
Is the train that is late coming at last?
Is the dog who is hungry licking the bowl?
Is the key that is different for the car?
Is her friend who is a doctor pretty?
Is the girl who is here hiding?
Is a butterfly that is yellow in the picture?
Is the horse that is outside dangerous?
Is the party that is put together for Alice at five o'clock?
Is the bedspread that is white and blue dirty?
Is the fish that is black his favorite one?
Is the telephone that is ringing orange?
Is the picture that is there a funny one?
Is the boy who is here your friend?
Is your grandpa who is away bringing you a cat?
Is the bunny that is there fluffy?
Is the man who is in the bedroom shaving?
Is the mummy that is in the museum very old?
Is the alligator that is in the lake dangerous?
Is the door that is on the right the kitchen?

[50] Only the set of grammatical test sentences are provided here. The corresponding ungrammatical sentences have the auxiliary removed from the relative clause and it is instead included after the relative clause. For instance, the ungrammatical version of "*Is the boy who is crying hurt?*" is "*Is the boy who crying is hurt?*". For the grammatical do-support PIRC, the finite verb is found in the relative clause and the non-finite one in the main clause whereas for the corresponding ungrammatical sentence, it is the other way around.

Is the crayon that is in the box the blue one?
Is the thing that is in your pocket the blue comb?
Is the animal that is in the next cage a tiger?
Is the house that is by the lake grandma's?
Is the little house that is behind the tree the doghouse?
Is the hair that is on his face a beard?
Is the stroller that is in the trunk for the baby?
Is the spoon that is on the bed the big one?
Is the man who is in the pool swimming?
Is the bird that is in our garden a goose?
Is the boy who is being good on the beach building a castle?
Is the song that is playing familiar?
Is your friend who is waving from the boat going fishing?
Is the picture that is hanging by the door done by Amelia?
Is the girl who is wearing a purple dress going to the party?
Is the man who is standing on his head Peter?
Is the muppet that is swinging the funny one?
Is the little boy who is crying hurt?
Is the man who is thinking of moving here grandpa?
Is the balloon that is flying over there Michael's?
Is the man who is talking to Jesse the doctor?
Is the Indian who is fishing in the lake catching anything?
Is the man who is holding Cindy's hand her daddy?
Is the kid who is patting the elephant the small one?
Is the monster who is kicking and throwing everything friendly?
Is the person who is leaving ready?
Is the lady who is kissing Jeffrey's brother their aunt?
Is the cutie who is sitting next to Ann Jeffrey?
Is the girl who is being dressed in white saying goodbye?
Is the monkey that is scratching his head thinking?
Is the dog that is lying under the table sleeping?
Is the breakfast that is made for Joey the same?
Is the mistake that is made in this book strange?
Is the dolphin who is hurt playing with the balls?
Is the hairbrush that is not grandma's cleaner?
Is juice that is made from carrots and bananas really delicious?
Is the car that is covered in dirt the new one?
Is the person who is interested in music listening to a tape?
Is the lion that is fed only carrots happy?
Is the girl who is dressed in pink trying to run away?
Is the box that is wrapped in blue paper for Paul's birthday?
Is the office that is closed today supposed to be open tomorrow?
Is the lady who is shut in the tower crying all the time?
Is the ring that is attached to the block for lifting it?
Is the quilt that is wrapped around his head made by grandma?
Is the boy who is bigger than Christopher called Pepe?

Is the doll that is beautiful mommy's?
Is the pig that is bigger making a mess?
Is the music that is so soft coming from that car?
Is the little boy who is tired sleeping?
Is the dog that is friendly in the house?
Is the umbrella that is strawberry red Gail's?
Is the jello that is on top of the marshmallows green?
Is the liquid that is in the jar cold?
Is the bedspread that is covered in red paint never getting clean again?
Is the engine that is turned off broken?
Is the water that is falling from the sky clean?
Is the boy who is drinking milk listening to her?
Is the girl who is wearing a jacket feeding the birds?
Is the phone that is on the table ringing?
Is everything that is in the kitchen food?
Is the key that is in her purse for the office?
Is the girl who is sitting on the chair lovely?
Is the dance that is so long tricky?
Is the man who is so gentle in the red car?
Is your aunt who is here old?
Is the machine that is making such a noise in the kitchen?

2. “Is-is” object-gap PIRCs

Is the cake the boy is eating delicious?
Is the little girl grammy is looking at brushing her teeth?
Is the airplane bert is not flying in the sky?
Is the favorite toy she is holding clean?
Is the cookie he is making in the bowl?
Is the picture christopher is getting for scott on the table?
Is the chair the kid is bringing rocking?
Is the bird Annie is feeding hungry?
Is this bell your friend is ringing old?
Is the dog Jeff is patting jumping?
Is the book kristin is reading interesting?
Is the black dog the boy is walking looking tired?
Is the truck the old man is testing going to work soon?
Is the beautiful horse the boy is waving to going fast?
Is the new word cindy is saying hard?
Is the fly her sister is catching alive?
Is the flower his mom is smelling lovely?
Is the dessert the kid is eating good?
Is the baby she is lifting licking her hands?
Is the music he is listening to nice?
Is the finger the nurse is sticking hurting?
Is every doll this girl is swinging falling?
Is the grandma she is calling napping?
Is the baby the mommy is lifting wicked?
Is the duck cindy is touching swimming?
Is the stroller the daddy is pushing blue?
Is any of the utensils he is hiding round?
Is the tower the dragon is building pretty?
Is any of the caves she is standing nearby beautiful?
Is the girl the lady is kissing combing her hair?
Is the song the kid is playing familiar?
Is the book her friend is reading funny?
Is the apple juice your brother is drinking good?
Is the bedroom he is cleaning neat?
Is the crayon our kid is trying to draw with in the car?
Is her sweater our friend is showing you new?
Is annie's cake he is having with his coffee hot?
Is the diaper he is throwing dirty?
Is the dress she is wearing short?
Is the white box he is opening heavy?
Is the pile of books Jessie is stacking up high?
Is the shoe the boy is moving cute?
Is the jeans jacket the girl is hanging in the kitchen?

Is the shirt jeffrey is putting on blue?
Is the job her brother is doing dangerous?
Is the zoo her sister is going to far from the park?
Is the new table gail is scratching in the office?
Is the beard your dad is shaving off long?
Is the puppy he is letting in smart?
Is the chair Paul is sitting on big enough?
Is the breakfast my mom is making for joey the same?
Is the clock Jake is playing with still going?
Is the box daddy is holding for paul's birthday?
Is the mirror mom is cleaning big?
Is the lion her sister is feeding carrots and an apple happy?
Is the girl the boy is talking to trying to run away?
Is the person your friend is listening to interested in music?
Is the shoe he is wearing making a scratchy sound?
Is the shirt the boy is brushing the dirt from peter's?
Is the pool your brother is swimming in close to your house?
Is the door the doctor is opening on the right?
Is the box of cereal the man is taking full?
Is the building the kid is hiding in hard to find?
Is the bed the kid is lying on soft?
Is the noise he is making just now coming from his tummy?
Is the daisy she is putting in water alive?
Is the dress Annie is zipping too small?
Is the party mimi is throwing going to be fun?
Is the steak his mom is bringing for the cat?
Is the movie she is looking for funny?
Is the kind of fish he is eating hard to catch?
Is the page he is reading interesting?
Is the horse the kid is patting still moving?
Is the wagon dan is cleaning attached to his car?
Is the museum the boy is standing at far away?
Is the castle mommy's friend is building for himself?
Is that tiger the girl is kissing in the stroller?
Is the street my dad is walking on quiet?
Is the toy he is feeling fluffy?
Is the bird Amelia is feeding lovely?
Is the new game the dragon is beginning dangerous?
Is the door the umbrella is hanging from broken?
Is the cookie his friend is eating almost gone?
Is the sock he is looking for on the floor?
Is the wagon your sister is pushing red?
Is the meat the dog is smelling bad?
Is the butterfly she is catching like a bird?
Is the animal daddy is waving to in the zoo?
Is the picture grandma is showing us neat?

Is the rug the kid is jumping on soft?
Is the mommy the girl is calling late?
Is the ball the boy is sitting on round?
Is the baby Grammy is rocking crying?
Is the lake the kitty is getting to gorgeous?
Is the face the baby is feeling scratchy?
Is the workout Marco is doing hard?
Is the child grandpa is not waking up sleepy?
Is the letter he is saying from the alphabet book?
Is the beach the boy is lying on along the river?
Is the chair he is lifting in his truck?

3. Do-support PIRCs

Does the man who likes animals work in an office?
Does the boy who plays the drum want a cookie?
Does the lady who kisses the baby every morning love him?
Does the girl who talks so much have time for anything else?
Does the monkey that looks funny bite?
Does the train that leaves at ten go there?
Does the doctor who lives here never sleep late?
Does the parrot that makes that funny noise need a home?
Does every kid who comes here open the door?
Does the guy who often calls Daisy get to see her at all?
Does a dog that smells a rabbit bark?
Does the dragon that bites its tail live in the castle?
Does the show that starts tomorrow have any new songs?
Does the boy who pats the bunny talk to it?
Does the old lady that phones the doctor so often look sick?
Does the girl that ties her shoes make nice bows?
Does the bird that lives in a cage feel happy?
Does the box that holds the key stay in the tower?
Does the person who says sorry feel sorry?
Does the toy that rings look like a phone?
Does the coffee that tastes so great wake you up?
Does the baby who wants to play peekaboo hold mommy's hand?
Does the doll that says nightnight mean a lot to Alice?
Does Gail's daddy who works far away leave at seven every morning?
Does the bird that wakes me up every morning live in that tree?
Does a dog that barks a lot bite hard?
Does the cat that goes to the park like flowers?
Does Kay's mommy who puts her to bed always kiss her good night?
Does a monkey that sees its face in a mirror pat its head?
Does Paul who stays home alone every evening have much fun?
Does a pig that smells so bad make a nice pet?
Does the turtle that sleeps in the bathtub get cold?
Does the dress that ties at the back match her eyes?
Does the duck that loves swimming play in the lake?
Does a girl who wants to be pretty put a ribbon in her hair?
Does the bell that looks broken ring at all?
Does the jacket that matches my shoes look nice enough for the office?
Does the clock that ticks on the wall need to be fixed?
Does your grandma who lets you run in the garden live nearby?
Does anybody who opens a book love reading?
Does your friend who gets all these letters call you back?
Does the cat that feels sorry for a mouse stay hungry?
Does your aunt who calls you every month say something interesting?

Does every guy who happens to see that play come again?
Does the rabbit that gets a carrot start to jump?
Does the house that means so much to him happen to be a castle?
Does a tiger that needs to be fed go to sleep?
Does every person who has a phone talk all night?
Does a flower that looks so lovely always smell nice?
Does the dog that barks in the bedroom smell good?
Does the boy who likes crayons color the dragon?
Does the soup that has bananas in it taste funny?
Does the girl who likes puppies pat the doggie?
Does the man who goes to the beach need sandals?
Does the funny guy who has a hat bow?
Does a parrot that says sorry mean it?
Does the zebra that bites live in the zoo?
Does a squirrel that holds a peanut put it down?
Does a man who feels sick stay at home?
Does a fish that has a big tail look bigger?
Does the boy who puts on his jacket tie his shoes?
Does a lion that wants to catch a rabbit stay quiet?
Does a puppy that sees a butterfly bark?
Does the old lady who likes coffee have trouble sleeping?
Does the clock that ticks so much wake you at night?
Does the bunny that holds a valentine have a yellow tail?
Does the child who colors a picture first get a new toy?
Does a cat that plays with fire have whiskers?
Does a grandma who loves you make you pancakes?
Does the girl who makes bags need buttons?
Does the tiger that lives in the zoo let anybody ride him?
Does a man who sleeps all night feel tired in the morning?
Does the clock that looks broken actually work?
Does a book that starts so strange let you sleep?
Does a child who holds a balloon feel happy?
Does the dragon that talks to her happen to be real?
Does your uncle who works in a kitchen like his job?
Does honey that tastes good mean you can make a cake with it?
Does the lion that bows get a big dinner?
Does a daddy who loves his children let them eat on the floor?
Does the girl who kisses her doll talk to it as well?
Does a horse that sleeps on its feet get any rest?
Does a puppy that opens the door mean trouble?
Does her brother who likes music play the violin?
Does his grandma who lives far away see him often?
Does the birdie that comes back stay with them every fall?
Does a lady who likes a necklace get it?
Does the child who plays alone have any friends?
Does the boy who sees a tiger get scared?

Does a dog that bites play with the children?
Does a cat that sleeps all the time let the mouse play?
Does his sister who wants a car have a job?
Does the jacket that matches her jeans have three buttons?
Does the guy who lives in his wagon talk to anybody?
Does the show that opens tomorrow start at ten?
Does her grandpa who lives in a castle phone her?
Does a monkey that goes for a walk put on a hat?
Does your sister who works in a zoo pat the animals?
Does a brother who makes fun of you love you?
Does a mother who works have time for her kid?

4. Dutch PIRCs with lexical verb fronting

Wil de baby die op de nieuwe stoel zit een koekje?
Vindt mijn hond die op de deken zit een tas?
Wil de olifant die daar rust water?
Drinkt mijn oma die op het bed ligt thee?
Maakt een muis die daar eet een grappig geluid?
Helpt mijn mama die hier werkt haar vrienden?
Zoekt het konijn dat op de grond springt iets?
Vangt een tijger die snel loopt zijn eten?
Ziet een vogel die boven de boom vliegt de kinderen?
Krijgt de kat die met het kind speelt meer melk?
Voelt de dame die in het bed slaapt pijn?
Zingt de postbode die de straat uit rijdt een lied?
Maakt de kok die moe lijkt een cake?
Wil mijn baas die overal heen vliegt een boot krijgen?
Zegt jouw bekende die hier komt hallo?
Heeft de clown die op zijn gezicht valt een rode neus?
Vertelt de oppas die soms laat blijft hem grappige verhalen?
Doet een paard dat te oud wordt weinig op de boerderij?
Kijkt de leeuw die op het gras ligt naar de vlinder?
Bereikt een schildpad die langzaam loopt het einde in ieder geval?
Kietelt de pen die haar raakt?
Wordt het kind dat vies water drinkt ziek?
Schrikt een klein meisje dat haar mama verliest?
Komt zijn tante die van appeltaart houdt hier vaak?
Kletst de oude man die auto rijdt aan de telefoon?
Moet de man die de vrouw bloemen geeft hier werken?
Lijkt mijn vriendin die boeken zoekt gelukkig in de bibliotheek?
Zingt haar papa die piano speelt ook?
Kletst de buur die haar soms helpt ook veel?
Stinkt de eend die alleen vis eet?
Rust de reus die een boek wil op de vloer?
Slaapt mijn oom die dit verhaal goed vertelt nu?
Draait mijn broer die de bal vangt in het rond?
Hangt het jasje dat drie knopen heeft daar?
Wordt een kraai die iets moois ziet gek?
Voelt mijn zusje die de plaat breekt zich slecht?
Valt de vlinder die de kaars raakt?
Spreekt haar vader die enkele woorden zegt onduidelijk?
Voelt de jongen die de brug in het donker vindt zich dapper?
Speelt de aap die naar haar kijkt in het bos?

Appendix B: Object-gap and do-support PIRCs added to the Bernstein-Ratner corpus

Is the toy that the baby is holding in his hands a teddy bear?
Is the sandwich that aunt Linda is making for the party hot or cold?
Is uncle David who the dog is barking at running away?
Is the notebook that Peter is putting in his bag the one for the English class?
Is the ring the lady is wearing missing one diamond?
Is the bee that Mike is trying to catch mocking him?
Is the drink that the prince is offering to the princess too sweet for her?
Is the policeman who this family is asking for help getting back in his car?
Is the boat that your brother is learning to sail big enough to go out on the ocean?
Is the tooth that the dentist is going to remove causing you pain?
Is the mountain that the hiker is climbing covered with snow?
Is the bookcase that my sister is painting missing a shelf?
Is the novel that Eric is reading captivating?
Is the dress that grandma is mending for her granddaughter's birthday?
Is the grammar rule that the teacher is teaching to her students difficult for them to understand?
Is the hammer that the carpenter is using going to break the nail?
Is the clown that this child is fond of crying for nothing?
Is the blue lamp that she is turning on on the desk?
Is the doll that the little girl is throwing on the floor her favorite one?
Is the letter that the soldier is writing to his family about the war?
Is the scarf that my cousin is knitting for Christmas blue?
Is the chair that the man is lifting heavy?
Is the telescope that the father is buying for his son the same as Peter's?
Is the dog that the gardener is feeding eating carrots?
Is the bed that Mark is fixing in his parents' bedroom?
Is the waterfall that Jenny is taking pictures of a famous one?
Is the blackboard that the teacher is writing on filled with math formulas?
Is the store that Julia is going to in the mall?
Is the frame that Adam is mending in the living-room?
Is the drawer that the secretary is closing hard to lock?
Is the wardrobe that mommy is throwing away emptied of everything?
Is the bottle of wine that Alice is bringing to the party from France?
Is Romeo who Juliet is chasing after not in love with her anymore?
Is the author who Susie is congratulating signing the book for her?
Is the elephant that the clown is playing with smiling at him?
Is the pencil that the child is sharpening a good one to draw with?
Is the student who she is giving a test to able to get a good grade?
Is the fruit that the cook is mixing with the ice-cream a good choice for the dessert?
Is the shirt that mom is ironing made of linen?
Is the table that the dog is jumping on sturdy enough to hold the weight?
Is the bread that mommy is cutting hot?
Is the car that you are pushing making a strange noise?
Is the house that your dad is building next to ours?

Is the coat that bob is buying the most expensive one?
Is the lady who you are talking to alicia's mom?
Is the milk that dad is preparing for the baby boiling?
Is the show that your sister is watching funny?
Is the screw that your brother is pulling out of the wall broken?
Is the number that mike is dialing the wrong phone number for the drugstore?
Is the cup of tea that he is heating full?

Does the rabbit that eats grass prefer carrots instead?
Does the student who plays tennis want to become a champion?
Does the doctor who takes care of me also treat my brother?
Does Eric who feels sorry for the child want to give him a present?
Does the duck that swims in the pond have any ducklings?
Does the tree that loses its leaves grow apples?
Does the horse that runs very fast dislike the other animals at the farm?
Does the car that runs on electricity need to be charged?
Does the man who sleeps heavily snore at night?
Does the dog that barks at the neighbor bite people hard?
Does the gardener who plants trees also grow tomatoes?
Does the baby who gives a kiss to his mom talk already?
Does the squirrel who likes nuts steal them from the bird?
Does the boy who hates spinach watch Popeye?
Does the cowboy who feeds his horse drink coffee?
Does William who sleeps in this small room pay his rent?
Does the kid who sells sodas profit from them?
Does rain that falls continuously stop the game?
Does the flower that smells so good bloom in the spring?
Does Amy who spends her summer in Rome speak Italian?
Does the light that goes on and off cause traffic jams?
Does the athlete who trains for the marathon follow a good diet?
Does the dragon that lives in the tower feel lonely?
Does the ghost that haunts the house scare people away?
Does the hairdresser who cuts my hair make much money?
Does the chair that folds up need to be painted?
Does the paper that requires so much revision tackle the right issues?
Does the scientist who works in his lab ever see the outside world?
Does the salesman who drives the BMW deceive people?
Does the bus driver who picks up the kids for school give them candies?
Does the chocolate that tastes so good contain much fat?
Does the coach who trains the football team motivate his players?
Does the movie that uses 3d technology sell many tickets?
Does the mother who sweeps the floor also clean the windows?
Does the man who wanders in the streets know his way home?
Does the surgeon who performs the operations give orders to the nurse?
Does the mad woman who screams in the night scare the neighbors?

Does the suitcase that weighs so much contain lots of books?
Does the mayor who favors new tax reforms support the president?
Does the tractor that expels toxic fumes affect the crops?
Does the architect who draws plans for skyscrapers live in a lake house?
Does a deadline that approaches quickly motivate me to work faster?
Does the chef who bakes wonderful cakes win lots of prizes?
Does the research that uses the newest technology reveal anything interesting?
Does Mark who travels all over the world plan to move to Paris?
Does the boy who collects stamps sort them in his album?
Does a driver who argues with a policeman receive a fine?
Does the show that attracts many spectators generate good revenue for the town?
Does the plane that lands in Calcutta have many passengers today?
Does the woman who keeps wiping her glasses see more clearly?

Appendix C: List of MOR tags found in the Bernstein-Ratner corpus and their counts

-sent-	9797
v	4113
n	4078
pro	3867
det	2837
co	2567
prep	2107
pro:dem	1171
v:aux	942
pro:wh	843
n-PL	807
v:aux&3S	796
adj	761
pro:wh~v:aux&3S	641
adv	600
n:prop	600
pro:poss:det	585
adv:loc	572
conj:coo	513
qn	500
part-PROG	454
pro:dem~v:aux&3S	405
pro~v:aux&3S	405
v&ZERO	351
inf	339
v:aux&PRES	327
pro:indef	313
n-DIM	289
v&3S	264
det:num	243
fil	237
adv:wh	209
v:aux~neg	206
int	188
v~inf	184
v&PAST	179
n:let	153
v:aux&PAST	137
part-PROG~inf	135
pro~v:aux&PRES	129
unk	127
adv:wh~v:aux&3S	110
neg	108
ptl	106
v-PAST	106
pro~v:aux&1S	97
v~pro	97

pro~v:aux 94
n~v:aux&3S 92
conj:subor 91
pro:exist~v:aux&3S 89
adv:int 76
n-POSS 76
adv:loc~v:aux&3S 73
part&PERF 64
v:aux&3S~neg 62
adv:adj-LY 59
adv:tem 43
n:prop-POSS 40
on 32
det:wh 26
part-PERF 21
v:aux&PAST~neg 21
v~prep 20
v:aux&PRES~neg 16
n:pt 16
pro:exist 15
pro:indef~v:aux&3S 14
pro:wh~v:aux&PRES 14
n&PL 14
n:prop~v:aux&3S 14
adv:aux 13
adv:wh~v:aux&PAST 12
pro~v:aux&COND 12
adj-CP 12
pro:poss 10
co:voc 10
v:aux&ZERO 9
n-DIM~v:aux&3S 9
pro:refl 8
adj&CP 8
n~v:aux 7
v:aux&PERF 7
v~neg 7
n-POSS~v:aux&3S 6
adv:loc~v:aux&PRES 6
n:adj-NESS 6
fam 6
n-FULL 5
pro:indef-PL 5
v:aux&1S 5
adv:wh~v:aux&PRES 5
adj:n-LY 4
pro:wh~v:aux&PAST 4
n:v-AGT 3
v&3S~neg 3
n:prop-POSS~v:aux&3S 3
v&PAST~neg 3

un#v 3
det:num-PL 2
v~v:aux&3S 2
n|dais-DIM-PL^n-PL 2
chi 2
v&PAST~inf 2
rel 2
v:aux~inf 1
0inf1
adj:n-Y 1
pro:indef~v 1
wplay 1
0aux 1
adj-SP 1
pro~v 1
adv:tem~v:aux&3S 1
pro:dem~v:aux 1
pro:wh~v:aux~pro 1
*v&PAST 1

Bibliography

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. (2008). Is structure dependence an innate constraint? New experimental evidence from children's complex-question production. *Cognitive Science*, *32*, 1, 222–255.

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.

Brown, R. (1973). *A First Language: The Early Stages*. Cambridge, MA: Harvard University Press.

Chang, F., Lieven, E., & Tomasello, M. (2005). Towards a quantitative corpus-based evaluation measure for syntactic theories. In *Proceedings of the Cognitive Science Society*, 418–423.

Chang, F., Lieven, E., & Tomasello, M. (2006). Using child utterances to evaluate syntax acquisition algorithms. In *Proceedings of the Cognitive Science Society*, 154–159.

Chater, N., & Vitanyi, P. (2007). 'Ideal learning' of natural language: positive results about learning from positive evidence. *Journal of mathematical psychology*, *51*, 135–163.

Chater, C., & Manning, D. C. (2006). Probabilistic models of language processing and acquisition. *TRENDS in Cognitive Sciences*, *10*, 335–344.

Chen, F. S., & Goodman, T. J. (1996). An empirical study of smoothing technique or language modeling. In *Proceedings of the 34th Annual Meeting of the ACL*, 310–318, Santa Cruz, California, June.

Chomsky, N. (1957). *Syntactic Structures*. The Hague: Mouton.

Chomsky, N. (1965). *Aspects of the Theory of Syntax*. Cambridge: MIT Press.

Chomsky, N. (1971). *Problems of Knowledge and Freedom*. London: Fontana.

Chomsky, N. (1980). *Rules and Representations*. Oxford: Basil Blackwell.

Chomsky, N. (1981). *Lectures on Government and Binding.* Mouton de Gruyter.

Chomsky, N. (1995). *The Minimalist Program*. Cambridge: MIT Press.

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: 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. (submitted). 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., Stoyneshka, I., Tornyova, L., Fodor, J. D., & Sakas, W. G. (2008). Bigrams and the richness of the stimulus. *Cognitive science, 32*, 771–787.

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

Kaplan, M. R., & Bresnan, J. (1982). Lexical-Functional Grammar: A formal system for grammatical representation. In Joan Bresnan (ed.) (1982). *The Mental Representation of Grammatical Relations*, 173 – 281. Cambridge: MIT Press.

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,* 65–73.

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.

Menn, L., & Gleason, J. B. (1986). Babytalk as a stereotype and register: Adult reports of children's speech patterns. In J. A. Fishman (Ed.), *The Fergusonian Impact, Volume I*. Berlin: Mouton de Gruyter.

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.

Pollard, C., & Sag. A. I. (1994). *Head-driven phrase structure grammar*. Chicago: University of Chicago Press.

Pullum, G. K., & Scholz, B. C. (2002). Empirical assessment of stimulus poverty arguments. *The linguistic review, 19,* 9–50.

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

Reali, 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.

Richards, B. J. (1990). *Language Development and Individual Differences: A Study of Auxiliary Verb Learning*. Cambridge: Cambridge University Press.

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.

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

Thompson, S. P., & Newport, E. L. (2007). Statistical learning of syntax: The role of transitional probability. *Language Learning and Development, 3,* 1–42.

Tomasello, M. (2003). *Constructing a language: A Usage-Based Theory of Language Acquisition*. Harvard University Press.

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.

Valian, V. (1991). Syntactic subjects in the early speech of American and Italian children. *Cognition, 40,* 21–81.

Vries, M. de (2002). *The syntax of relativization*. Ph.D dissertation. University of Amsterdam.

Warren-Leubecker, A., & Bohannon, J. N. (1984). Intonation patterns in child-directed speech: Mother-father speech. *Child Development, 55,* 1379–1385.

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.

Witten I., & Bell. T. (1991). The Zero-Frequency Problem: Estimating the Probabilities of Novel Events in Adaptive Text Compression. In *IEEE Transactions on Information Theory*, 37, (4).