3

Children's Acquisition of Syntax: Simple Models are Too Simple*

XUAN-NGA CAO KAM AND JANET DEAN FODOR

3.1 Introduction

3.1.1 *Studying early syntax acquisition*

There has been a renewal of interest in statistical analysis as a foundation for syntax acquisition by children. At issue is how much syntactic structure children could induce from the word sequences they hear. This factors into three more specific questions: How much structure-relevant information do word strings contain? What kinds of computation could extract that information? Are pre-school children capable of those kinds of computation? These points are currently being addressed from complementary perspectives in psycholinguistics and computational linguistics.

Experimental studies present a learning device with a sample of sentences from a target language, and assess what aspects of the target syntax are acquired. The learning device may be a child, an adult, or a computer program. The language may be artificial or (part of) a real natural language. Each of these combinations of learner and language is responsive to one of the methodological challenges in research on early syntax acquisition. In the research reported here, the language was natural but the learner was artificial. We explain below why we regard this combination as especially fruitful.

Testing infants has the undeniable advantage that the psychological resources (attention, memory, computational capacity) of the subjects match the resources available for real-life primary language acquisition. However, the input language in child studies is typically artificial, because it is improper to tamper with the acquisition of the subjects' native language, and also to control across subjects exactly what input

* This work originated in a joint project with Iglika Stoyneshka, Lidiya Tornyova, and William Sakas, published as Kam et al. (2008). The project was continued in Kam's (2009) dissertation. The present paper has benefited from the technical expertise of Martin Chodorow and William Sakas.

they receive. In order for an infant to acquire properties of an artificial language in the span of an experimental session, the language must also be very simple. See Gómez and Gerken (1999) for a classic example.

Adult subjects can undergo more extensive and rigorous testing than infants, providing more data in less time. But again, the input language must be artificial and fairly simple for purposes of experimental control and uniformity across subjects (e.g., Thompson and Newport 2007). With adult subjects, moreover, it is not possible to exclude from the experimental context any expectations or biases they may have due to their existing knowledge of a natural language. For example, Takahashi and Lidz (2008) found that the adult subjects in their study respected a constituency constraint on movement in the test phase, even when the training sample contained no movement constructions. Although of considerable interest, this is prey to uncertainties similar to studies of 'normal' adult L2 acquisition: was the sensitivity of movement to constituency due to an innate bias, or to analogy or transfer from the subject's L1?

Artificial language studies, with children or adults, provide no insight into what could be learned from word strings in the absence of any innate biases or prior linguistic experience. But this is the issue that has animated many recent computational studies of language acquisition, motivated in large part by a conjecture that language acquisition may not, after all, require any innate substrate, despite long-standing assumptions to the contrary by many linguists and psycholinguists. The focus of these computational studies is on pure distributional learning, relying solely on the information that is carried by regularities in the sequences of words.[1] For investigating this, only an artificial learner will do. If the learning system is an algorithm implemented in a computer program, there is complete certainty as to whether, before exposure to the target input, it is innocent of linguistic knowledge of any kind (as in the model we discuss below), or whether it is equipped with certain biases concerning what language structure is like, such as the 'priors' of Bayesian learning models (Perfors et al., 2006) or some version of Universal Grammar as espoused by many linguists.

Another advantage of artificial learners is that the target language can be a real natural language, or a substantial part thereof. Since the learning algorithm has no L1, there are no concerns about transfer. More complex phenomena can be examined because there is little constraint on the extent of the training corpus or how many repetitions of it the learner is exposed to. Moreover, not only is the presence/absence of prior knowledge under the control of the experimenter, but so too are the computational sophistication and resources of the learning device. So this approach can provide systematic information concerning what types of computational system can extract what types of information from input data. We illustrate this in section 3.3 below.

[1] While children make use of prosodic, morphological and semantic properties of their input (Morgan 1986), these sources of information are set aside in many computational studies in order to isolate effects of co-occurrence and distribution of words.

The artificial learner approach has its disadvantages too, especially uncertainty as to which experimental outcomes have bearing on human native language acquisition. In compensation, however, a wide range of different algorithms can be fairly effortlessly tested, and informative comparisons can be drawn between them. The hope is that it may one day be possible to locate children's learning resources and achievements somewhere within that terrain, which could then provide guidance concerning the types of mental computation infants must be engaging in as they pick up the facts of their language from what they hear.

3.1.2 *Transitional probabilities as a basis for syntax acquisition*

The specific learning models we discuss here are founded on transitional probabilities. It has been demonstrated that infants are sensitive to transitional probabilities between syllabic units in an artificial language, and can use them to segment a speech stream into word-like units (Saffran et al. 1996). For syntax acquisition, what is relevant is transitional probabilities between one word and the next. Infant studies have documented sensitivity to between-word transitional probabilities which afford information about word order patterns and sentence structure (Gómez and Gerken 1999; Saffran and Wilson 2003). The type of learning model discussed below puts the word-level transitional probabilities to work by integrating them into probability scores for complete word strings, and on that basis predicts which strings are well-formed sentences of the target language (details in section 3.2.2). We assess the model's accuracy under various circumstances, and where it falls short we ask what additional resources would be needed to achieve a significant improvement in task performance.

The original stimulus for our series of experiments was a dramatic report by Reali and Christiansen (2003, 2005) (see also Berwick, Chomsky, and Piattelli-Palmarini in this volume). They found that an extremely simple model using transitional probabilities between words, trained on extremely simple input (speech directed to one-year-olds), was able to ace what is often regarded as the ultimate test of syntax acquisition: which auxiliary in a complex sentence moves to the front in an English question? If that finding could be substantiated, there would appear to be no need to develop more powerful acquisition models. Distributional learning of a complex syntactic construction would have been proved to be trivially easy.

We checked the finding and replicated it (results below). However, as we will explain, we found it to be fragile: almost any shift in the specific properties of the test sentences resulted in chance performance or worse. Thus, two questions presented themselves. (i) What distinguishes the circumstances of the original success from those of the subsequent failures? (ii) Does an understanding of that give grounds for anticipating that broader success is within easy reach, needing perhaps only slight enrichment of the original model or the information it has access to?

To address these points, the first author conducted a series of eighteen computer experiments, reported in full in Kam (2009). The earlier experiments, summarized here as background, showed that the model's use of transitional probabilities at the level of words, or even with part-of-speech categories, does not suffice for reliable discrimination between grammatical and ungrammatical auxiliary fronting (Kam 2007). In this paper we report our most recent experiments in the series, which were directed to the role of phrase structure in the acquisition of auxiliary movement. To anticipate: we found that if, but *only* if, the learning model had access to certain specific phrase structure information, it succeeded spectacularly well on the auxiliary-fronting construction. The implication is that transitional probabilities could be the basis for natural language syntax acquisition only if they can be deployed at several levels, building up from observable word-level transitions to relations between more abstract phrasal units.

3.2 The Original *N*-Gram Experiments

3.2.1 *Linguistic preliminaries*

The sentences tested in these experiments were instances of what we call the PIRC construction (Polar Interrogatives containing a Relative Clause), in which question formation requires fronting of the auxiliary in the main clause, not the auxiliary in the RC (N. Chomsky 1968 and since). Grammatical and ungrammatical forms were compared. Examples are shown in (1), with the trace of the moved auxiliary indicated here (though of course not in the experiments).[2]

(1) a. Is$_i$ the little boy who is crying t_i hurt?
b. *Is$_i$ the little boy who t_i crying is hurt?

Reali and Christiansen (henceforth R&C) tested *n*-gram models: a bigram model and a trigram model. A bigram is a sequence of two adjacent words; a trigram is a sequence of three adjacent words. These *n*-gram models did not differ radically in their performance, so for brevity here we focus on the bigram model. It gathers bigram data from a corpus of sentences, and feeds it into a calculation of the probability that any given sequence of bigrams would also occur in the corpus. The bigrams in sentences (1a,b) are shown, in angle brackets, in (2a,b) respectively.

(2) a. <is the> <the little> <little boy> <boy who> <who is> <is crying> <crying hurt>
b. <is the> <the little> <little boy> <boy who> <who crying> <crying is> <is hurt>

[2] Following standard practice we refer to the inverting verbs as auxiliaries, though the examples often contain a copula (as in the main clause of (1) above). Below we also discuss *do*-support and inversion of main verbs.

Bigram statistics could be employed in many different ways within a learning model (see for example Chang et al. 2006; also section 3.4 below). The bigram model as defined by R&C puts bigrams to work in a direct and simple manner. It does not represent syntactic structure. It does not compose grammar rules. Its knowledge of the language consists solely of the set of all the bigrams in the training corpus, each assigned an estimated transitional probability (see below). R&C's experimental project thus raises the linguistic-theoretical question: Is it possible *in principle* to discriminate grammatical and ungrammatical forms of auxiliary inversion by reference solely to pairs of adjacent words?

We think most linguists would judge that it is not, for several reasons. One consideration is Chomsky's original point: that the generalization about the right auxiliary to move is not that it is in any particular position in the word string, but that it is in a particular position in the syntactic tree; auxiliary inversion is 'structure-dependent'. There are non-transformational analyses of the inversion facts, but they also crucially presuppose phrase structure concepts (see section 3.4.2 below). Also, auxiliary movement creates a long-distance dependency between the initial auxiliary and its trace if defined over the word string (six words intervene in (1a), clearly beyond the scope of a bigram) whereas the dependency spans just one element, an NP in every case, if defined over syntactic structure, bringing it within reach at least of a trigram model. So a purely linear analysis in terms of word pairs would seem unlikely to be able to capture the relevant differences that render (1a) grammatical and (1b) ungrammatical. However, R&C's noteworthy finding of successful discrimination by the bigram model suggests that we should pause and reconsider. Perhaps, after all, there are properties of the word pairs in the two sentence versions which, in some fashion, permit the grammatical one to be identified.

For instance, the bigram model might judge (1b) ungrammatical on the basis of its bigram <*who crying*>, which presumably is absent or vanishingly rare in a typical corpus. This may sound like a sensible strategy: judge a sentence ungrammatical if it contains an 'ungrammatical' (i.e., unattested) bigram. Against such a strategy the objection is often raised that a linguistic form may be unattested in a corpus for many reasons other than its being ungrammatical (cf. *Colorless green ideas sleep furiously*; Chomsky 1957). But in the case of auxiliary inversion, there is another and quite specific problem with this approach: the grammatical version (1a) also contains a vanishingly rare bigram <*crying hurt*>. By parity of reasoning, that should indicate to the model that (1a) is also ungrammatical, leaving no obvious basis for preferring one version of the sentence over the other. Thus, a decision strategy based on weighing one low-frequency bigram against another is delicately balanced: it might sometimes succeed, but not reliably so unless there were a systematic bias in the corpus against bigrams like <*who crying*> and in favor of bigrams like <*crying hurt*>. It is not clear why there would be; but that is just the sort of thing that corpus studies can usefully establish.

An alternative strategy might focus instead on the higher-frequency bigrams in the test sentences. The learner might judge a sentence grammatical if it contains one or more strongly attested bigrams. A good candidate would be the bigram <*who is*> in (1a), which can be expected to have a relatively high corpus probability. Since the ungrammatical version has no comparably strong bigram in its favor, there is an asymmetry here that the learner might profit from. This generates an experimental prediction: If the grammatical version in all or most test pairs contains at least one strong bigram, a high percentage of correct sentence choices is likely;[3] if not, the model's choices will not systematically favor the grammatical version. In the latter case, exactly how well the model performs will depend on details of the corpus, the test sentences, how bigram probabilities are calculated, and the sentence-level computations they are entered into. These we now turn to.

3.2.2 *Procedure*

For maximum comparability, all our experiments followed the method established by R&C except in the specific respects, indicated below, that we modified over the course of our multi-experiment investigation. The training corpus consisted of approximately 10,000 child-directed English utterances (drawn from the Bernstein-Ratner corpus in CHILDES; MacWhinney 2000). The test sentences were all instances of the PIRC construction. In a forced-choice task, grammatical versions were pitted against their ungrammatical counterparts (fronting of the RC auxiliary), as illustrated by (1) above.

For our Experiment 1, a replication of R&C's, we created 100 such sentence pairs from words ('unigrams') in the corpus, according to R&C's templates in (3), where variables A and B were instantiated by an adjective phrase, an adverbial phrase, a prepositional phrase, a nominal predicate, or a progressive participle with appropriate complements.

(3) Grammatical: Is NP $\left\{\begin{matrix} \textit{who} \\ \textit{that} \end{matrix}\right\}$ is A B?

Ungrammatical: Is NP $\left\{\begin{matrix} \textit{who} \\ \textit{that} \end{matrix}\right\}$ A is B?

The corpus contained monoclausal questions with auxiliary inversion (e.g., *Are you sleepy?*), and non-inverted sentences with RCs (e.g., *That's the cow that jumped over the moon*), but no PIRCs.

R&C computed the estimated probability of a sentence as the product of the estimated probabilities of the bigrams in the sentence.[4] The sentence probability was

[3] Of course it is possible that the bigrams in the ungrammatical version collectively outweigh the advantage of the strong bigram(s) in the grammatical version, so this strategy is not guaranteed to always lead to the correct choice. See results below.

[4] The probability of a bigram not in the corpus must be estimated. We followed R&C in applying an *interpolation smoothing technique*. In what follows, we use the term 'bigram probability' to denote the *smoothed* bigram probability.

entered into a standard formula for establishing the cross-entropy of the sentence (see details in R&C 2005 and Kam et al. 2008). The cross-entropy of a sentence is a measure of its unlikelihood relative to a corpus; a lower cross-entropy corresponds to a higher sentence probability. In the forced-choice task the model was deemed to select as grammatical whichever member of the test sentence pair had the lower cross-entropy relative to the training corpus. To simplify discussion in what follows, we refer to sentence probabilities rather than cross-entropies; this does not alter the overall shape of the results.

It is important to note that a bigram probability in this model is not the probability that a sequence of two adjacent words (e.g., *boy* and *is*) will occur in the corpus. It is the probability of the second word occurring in the corpus, given an occurrence of the first: the bigram probability of <*boy is*> is the probability that the word *is* will immediately follow an occurrence of the word *boy*. So defined, a bigram probability is equivalent to a transitional probability, as manipulated in the stimuli for the infant learning experiments noted above.

3.2.3 *Initial results and their implications*

In R&C's Experiment 1, the bigram model selected the grammatical version in 96 of the 100 test sentence pairs. In our Experiment 1 the model also performed well, predicting 87 percent of the test sentences correctly. Now we were in a position to be able to explore the basis of the model's correct predictions.

Some bigrams in the test sentences could not have contributed, because they were identical in the grammatical and ungrammatical versions. For the sentence pair (1), the bigrams <*is the*>, <*the little*>, <*little boy*>, and <*boy who*> are in both versions. The bigrams that differ are shown in Table 3.1; we refer to these as *distinguishing bigrams*. The model's selection of one sentence version over the other can depend only on the distinguishing bigrams.

The results showed, as anticipated in our speculations above, that the majority of correct choices were due to the contribution of the distinguishing bigram containing the relative pronoun in the grammatical version: either <*who is*> or <*that is*>. (Henceforth, we abbreviate these as <*who|that is*>.) This bigram had the opportunity to influence all judgments in the experiment because it appeared in every grammatical test sentence, and not in any ungrammatical versions. Note that this was by design: it was prescribed by the templates in (3) that defined the test items. The <*who|that*

Table 3.1. Distinguishing bigrams for the test sentence pair (1a)/(1b)

(1a) grammatical	<*who is*>	<*is crying*>	<*crying hurt*>
(1b) ungrammatical	<*who crying*>	<*crying is*>	<*is hurt*>

is> bigram boosted selection of the grammatical version in many cases because it had a higher corpus frequency than most other bigrams in the test sentences, in part because its elements are both closed-class 'functional' items, which recur more often than typical open-class lexical items.[5] In the ungrammatical version, by comparison, the word *who* or *that* was followed by a lexical predicate, differing across the sentence pairs and mostly with low corpus frequency (e.g., <*who crying*> in (1b)).

In short: the <*who*|*that is*> bigram is the means by which the model was able to select the correct form of auxiliary inversion. Its performance rested on a strictly local word-level cue, without any need to recognize the auxiliary movement dependency per se or to learn anything at all about the structural properties of PIRCs. Thus, one part of our mission was accomplished. Discovering the decisive role of the <*who*|*that is*> bigram explains the model's strong performance in R&C's original experiment, and in our replication of it. But this discovery raises a doubt about whether the model could select the grammatical version of PIRCs that lack a helpful 'marker' bigram such as <*who*|*that is*>. Our next task, therefore, was to find out whether other varieties of PIRC contain bigrams that can play a similar role.

3.3 Limits of *N*-Gram-based Learning

3.3.1 *Extending the challenge*

The templates in (3) are very specific. They pick out just a subset of PIRC constructions, those with *is* as the auxiliary in both clauses, and an RC with a subject gap (i.e., the relative pronoun fills the subject role in the RC). But there are many other variants of the PIRC construction: the auxiliaries may differ, the RC could have a relativized object, the matrix clause might have a lexical main verb that requires *do*-support in the question form, or in some languages the main verb may itself invert. The rule is the same in all cases, but the bigrams it creates vary greatly. Table 3.2 shows some examples.

In our subsequent group of experiments, aimed at assessing how generally the bigram model could pick out grammatical versions, we tested PIRCs with *is* in both clauses but an object gap RC, and PIRCs with a main verb and *do*-support. We also tested Dutch examples in which the main verb inverts.

The bigram model did very poorly on these PIRC varieties not constrained by R&C's templates; see Table 3.3.

These weak results suggest that the model did not find any reliable local cues to the grammatical version. Inspection of the distinguishing bigrams confirmed that these other PIRC varieties do not contain any useful 'marker' bigrams. These results thus support the diagnosis that when the bigram model does succeed, it does so on the basis

[5] Other factors bestowing a powerful role on the <*who*|*that is*> bigram were the specific nature of R&C's smoothing formula, and the fact that many other bigrams in the test sentences were not in the corpus; for details see Kam et al. (2008: section 3.2).

Table 3.2. More varied examples of auxiliary (or main verb) inversion

Sub-type of PIRC	Example
Is-is subject gap	(1a) Is the lion boy who is crying hurt?
Other auxiliaries	Can the lion that must sleep be fed carrots?
Is-is object gap	Is the wagon that your sister is pushing red?
Main verbs with *do*-support	Does the boy who plays the drum want a cookie?
Main verb inversion in Dutch	Wil de baby [die op de stoel zit] een koekje? 'Does the baby that is sitting on the chair want a cookie?'

Table 3.3. Bigram model performance for four varieties of PIRC

Subtype of PIRC	% correct	% incorrect	% undecided
Is-is subject gap RC (as above)	87	13	0
Is-is object gap RC	35	15	50
Main verbs with *do*-support	49	51	0
Main verb inversion in Dutch[6]	32.5	55	12.5

of information that is neither general nor inherently related to the structurally relevant properties of PIRCs. It is no more than a lucky chance if some specific instantiation of the PIRC construction—such as the one originally tested—happens to offer a high-probability word sequence that correlates with grammaticality.

A tempting conclusion at this point is therefore that this simple learning model is too simple to match the achievements of human learners. The original result was impressive, but subsequent tests appear to bear out the hunch that a word-level learner is not equipped to recognize the essential difference between correct and incorrect auxiliary-inversion. Neither the early success on *is-is* subject gap PIRCs nor the nature of the subsequent failures encourages the view that broader success could be attained by minor adjustments of the model or its input. So perhaps one might rest the case here. However, we really hoped to be able to settle the matter once and for all, so that later generations of researchers would not need to revisit it.

Also, to be fair, it should be noted that no child acquisition study to date has investigated the age (and hence the level of input sophistication) at which learners of English or any language achieve mastery of object gap PIRCs and *do*-support PIRCs.[7] This lacuna in the empirical record includes the much-cited early study by Crain and Nakayama (1987), which focused on the *is-is* subject gap variety. One step in the

[6] For Dutch, only forty sentence pairs were tested. All other experiments reported here had 100 test pairs for each subtype of PIRC.

[7] It has been maintained (Ambridge et al. 2006) that children before five years do not have a fully productive rule for auxiliary inversion even in single-clause questions.

right direction is taken in a recent study by Ambridge et al. (2008), which extends the domain of inquiry from *is-is* to *can-can* PIRCs.

One last reason for not rejecting *n*-gram models out of hand for auxiliary-inversion is that it is not at all an uncommon occurrence in current research to find that, as computational techniques have become ever more refined and powerful, they can achieve results which would once have been deemed impossible (Pereira 2000). Thus, given our goal of establishing an unchallengeable lower bound on learning mechanisms that could acquire a natural language, it was important to assess whether or not the failures we had documented stemmed from the inherent nature of the *n*-gram approach. Thus we entered the next phase of our project. We conducted additional experiments in which we provided the *n*-gram model with better opportunities to succeed if it could.

3.3.2 *Increasing the resources*

In Experiments 7–12, keeping the basic mechanism constant, we provided it with enriched training corpora:

- a longitudinal corpus of speech to a child (Adam) up to age 5;2;
- a corpus approximately ten times larger than the original, of adult speech to older children, up to age eight years, containing more sophisticated syntax;
- a corpus into which we inserted PIRC examples (fifty object gap; fifty *do*-support), providing direct positive information for the model to learn from if it were capable of doing so;
- the original corpus but with sentences coded into part-of-speech tags, as a bridge between specific words and syntactic structure.

In Experiments 13–15, we moved from the bigram model to a trigram model, gathering statistical data on three-word combinations, thus expanding the model's window on the word string. The trigram model was trained on the original corpus and the larger corpus with and without part-of-speech tags. (See Kam 2009: ch. 3 for detailed results.)[8] In all these studies we used the object gap and *do*-support PIRCs as test cases for whether an *n*-gram model could go beyond reliance on an 'accidentally' supportive surface word sequence such as <*who*|*that is*> in the subject gap examples.

These resource enhancements did improve the *n*-gram models' success rate to some extent, but performance on object gap and *do*-support PIRCs was still lackluster. Performance did not rise over 70 percent correct, except in one case (out of twenty-one results) which could be attributed to the presence of a 'marker' trigram.[9] Moreover, the *n*-gram models never did well across all PIRC varieties under the same conditions:

[8] The chapter by Berwick, Chomsky, and Piattelli-Palmarini in this volume, which includes a critique of R&C's approach to auxiliary inversion, presents data for the trigram model trained on an additional corpus: one created by Kam et al. (2008) in which the relative pronouns *who* and *that* were distinguished from interrogative *who* and demonstrative and complementizer *that*.

[9] The trigram was <n v:aux&3S part-PROG> (e.g., *sister is pushing*). It appeared only in grammatical versions, and in most of them due to materials construction: object gap RCs needed transitive verbs rather

sometimes performance on object gap PIRCs improved but *do*-support PIRCs did less well, and vice versa. Even the *is-is* subject gap type was less successful in many cases than in the original experiment. (See Kam 2009: ch. 3 for detailed results.)

Thus this series of experiments provided little support for the view that *n*-gram models are on basically the right track and need only a little more assistance from the environment to begin performing at a consistently high level. Two conclusions seem to be warranted. One is that either there wasn't rich information in the corpus or the *n*-gram models were too weak to extract it. Either way, the experimental findings offer no demonstration of 'the richness of the stimulus', which is the conclusion that R&C drew from their results: 'the general assumptions of the poverty of stimulus argument may need to be reappraised in the light of the statistical richness of language input to children' (R&C 2005: 1024). The second conclusion is that the *n*-gram models were unable to extend a pattern learned for one subvariety of PIRC onto other instantiations of the same linguistic phenomenon. The object gap and *do*-support forms were not mastered on their own terms, based on their own particular distributional properties; but equally clearly, the *n*-gram models did not form a general rule of auxiliary inversion which could be projected from the subject gap type to other varieties.

All of this points to a deep inability of a localistic word-oriented learning model to detect or deploy the true linguistic generalization at the heart of auxiliary inversion phenomena. Therefore a more radical shift seems called for: a qualitative rather than a merely quantitative augmentation of the learning model or its resources. Very different ideas are possible concerning what more is needed. Linguists may regard UG as the essential addition; computer scientists might call instead for stronger statistics, perhaps as embodied in neural networks;[10] psychologists might argue that negative data (direct or indirect) plays an essential role in child syntax acquisition. These possibilities are worth pursuing. But we chose, in our most recent set of experiments, to examine the role of phrase structure as a basis for the acquisition of transformational operations such as auxiliary inversion.

This third phase of our project thus moves toward a more positive investigation of the computational resources needed for the acquisition of natural language syntax: How could the previous learning failures be rescued? Here we address the specific question: In acquiring the auxiliary inversion construction, could an *n*-gram model benefit from access to phrase structure information? Chomsky's observation concerning the structure dependence of auxiliary inversion suggests that it might. In

than other predicate types such as adjectives. Apart from this, the only other success occurred when we ran the bigram model on the *Wall Street Journal* corpus (Marcus et al. 1999), which is presumably of little relevance to child language acquisition.

[10] Neural network models are at the opposite end of the scale from *n*-gram models in respect of computing power. Simple Recurrent Networks (SRNs) have been applied to the PIRC construction in work by Lewis and Elman (2001) and R&C (2005) and have performed well. But so far they have been tested only on the *is-is* subject gap variety which even the bigram model mastered, so the results are uninformative. More telling will be how they perform with other PIRC varieties on which the bigram model failed. (See also Berwick, Chomsky, and Piattelli-Palmarini, this volume)

non-transformational analyses, as in Head-driven Phrase Structure Grammar (HPSG; Sag et al. 2003), there is also crucial reference to phrase structure. Linguists disagree about many things, but on this point they are in full accord: there is no viable linguistic analysis that characterizes the auxiliary inversion construction in terms of unstructured word sequences.

3.4 Providing Phrase Structure Information

The aim of our phrase structure (PS) experiments was to integrate hierarchical structural representations into otherwise simple statistical learning models like those above, which rely solely on transitional probabilities between adjacent items. This project raises novel questions. How would such a learning system obtain PS information? How could it represent or use it?

On these matters we can only speculate at present. We suppose it might be possible to implement a sequence of *n*-gram analyses, at increasingly abstract levels, each feeding into the next: from words to lexical categories (parts of speech) to phrases and then larger phrases and ultimately clauses and sentences. The phrase structure information thus acquired would then enter into the PIRC discrimination task to assist in selecting the grammatical sentence. We emphasize that this is an experiment in imagination only at present. There do exist algorithms that compute phrase structure from word sequences,[11] but it remains to be established whether they can do so without exceeding the computational resources plausibly attributable to a two-year-old child (however approximate any such estimate must be). Multi-level tracking of transitional probabilities has been proposed as a means for human syntax acquisition. Some of the data are from adult learning experiments (Takahashi and Lidz 2008). But Gómez and Gerken (1999: 132) speculated for children: 'A statistical learning mechanism that processes transitional probabilities among linguistic cues may also play a role in segmenting linguistic units larger than words (e.g. clauses and phrases)'. Of interest in this context are the findings of an infant acquisition study by Saffran and Wilson (2003), which suggest that one-year-olds can perform a multilevel analysis, simultaneously identifying word boundaries and learning the word order rules of a finite-state grammar.

The approach we are now envisaging is sketched in (4):

(4) Multilevel *n*-gram analysis → phrase structure → PIRC discrimination

We decided to tackle the second step first, temporarily imagining successful accomplishment of the first one via some sort of cascade of transitional probability analyses at higher and higher levels of structure. We thus made a gift of PS information to the bigram learning model, and then tested it again on the auxiliary inversion

[11] We cannot review this literature. Some points of interest include Brill (1993); Ramshaw and Marcus (1995); Bod (2009); Wang and Mintz (2010).

forced-choice discrimination to see whether it would now succeed more broadly. Whether it would do so was not a foregone conclusion. But if phrase structure knowledge did prove to be the key, that would represent a welcome convergence between theoretical and computational linguistics.

3.4.1 *Method*

To run these experiments we had to devise ways by which PS information could be injected into the learning situation. We did so by assuming that the PS building process produced as output a labeled bracketing of the word string. Thus we added labeled phrase brackets into all word strings in the training corpus and test sentences.[12]

We inserted only NP brackets in the present experiments, for two reasons. We were concerned that a full bracketing would overwhelm the system. Within the constraint of a limited bracketing, the fact that the word sequence following the initial auxiliary is an NP seemed likely to be of most benefit to the learner (see discussion below). In future work we can explore the consequences of supplying a full phrase structure bracketing.

NP brackets were manually inserted surrounding all noun phrases in the original corpus and in the test sentences used in our earlier experiments (subject gap, object gap, and *do*-support PIRCs). Left and right brackets were distinguished; see example (5).

(5) Let $_{NP}$[the boy]$_{NP}$ talk on $_{NP}$[the phone]$_{NP}$.

For purposes of the bigram analysis, each bracket was treated on a par with words in the string. Thus a bigram now consisted of two adjacent items which might be words and/or labeled brackets. For example, one bigram in (5) is <*the boy*> and another is <*boy]*$_{NP}$>. Bigram and sentence probabilities (and cross-entropies) were then computed as before, and employed in the forced-choice discrimination task to select one sentence version as the grammatical one.

Two experiments were conducted. They differed with respect to the labels on the brackets in the test sentences. In PS-experiment 1 the labeled bracketing was as illustrated in (6). It does not distinguish well-formed NPs such as *the boy who is crying* in (6a) from ungrammatical NPs such as *the boy who crying* in (6b).

(6) a. Gramm: Is $_{NP}$[$_{NP}$[the little boy]$_{NP}$ $_{NP}$[who]$_{NP}$ is crying]$_{NP}$ hurt?
b. Ungramm: Is $_{NP}$[$_{NP}$[the little boy]$_{NP}$ $_{NP}$[who]$_{NP}$ crying]$_{NP}$ is hurt?

This labeling would allow us to see whether the model could identify the grammatical version based solely on the locus of a sequence of an NP followed by a non-finite predicate, which is acceptable in the main clause of (6a) but not in the RC in (6b).

[12] In other experiments we substituted the symbol *NP* for word sequences constituting noun phrases. (See Kam 2009: ch. 4 for details.)

In PS-experiment 2 we used the label *NP on the brackets around the ill-formed complex NP in the ungrammatical sentence version, as in (7b).

(7) a. Gramm: Is ${}_{NP}[\ {}_{NP}[$the little boy$]_{NP}$ ${}_{NP}[$who$]_{NP}$ is crying $]_{NP}$ hurt?
b. Ungramm: Is ${}_{*NP}[{}_{NP}[$the little boy$]_{NP}$ ${}_{NP}[$who$]_{NP}$ crying $]_{*NP}$ is hurt?

This avoids giving the learning model misleading information about the grammatical status of the word sequence *the little boy who crying*; it is not in an equivalence class with strings like *the little boy* or *Jim*. Note, though, that employing this labeling presupposes that in the prior PS-assignment stage, the learning model would have been able to recognize the deviance of *who crying* and percolate that up from the RC to the NP. We return to this point in discussion below. In any case, explicit indication that a word sequence such as *the little boy who crying* is not a well-formed constituent could be expected to provide the strongest support for rejection of ungrammatical PIRCs in the discrimination task.

3.4.1.1 *PS-experiment 1: Results and discussion* The percentages of correct choices for the object gap and *do*-support PIRCs were essentially unchanged compared with the original experiment without brackets; see Table 3.4. For the subject gap PIRCs, on which the model had previously succeeded without bracketing, there was a highly significant drop in performance.

This may appear paradoxical: provided with richer relevant information, the model performed less well. A positive outcome might have been anticipated due to the coding of the whole complex subject as an NP. Yet the data suggest that this hindered rather than helped. To understand this, let us consider the *is-is* subject gap examples in (6), with distinguishing bigrams as in (8).

(8) a. <is crying> < $]_{NP}$ hurt>
b. < $]_{NP}$ crying> <is hurt>

The unlikely bigrams <*crying hurt*> and <*who crying*> in (1a) and (1b) respectively (section 3.2.1 above) have now been transformed by the bracketing into better-supported bigrams: <$]_{NP}$ hurt> in (8a) and <$]_{NP}$ crying> in (8b). These might well occur in the corpus, instantiated in sentences like *Are ${}_{NP}[$you$]_{NP}$ hurt?* and *Is ${}_{NP}[$Baby$]_{NP}$ crying?* (also in small-clause constructions such as *I like ${}_{NP}[$my*

Table 3.4. Begram model performance in PS-experiment 1

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

porridge]$_{NP}$ hot). But since these bigrams with NP-brackets benefit both sentence versions, they provide no net gain for the grammatical one. For the object-gap and do-support PIRCs, comparable considerations apply, but we will not track through the details here.

Outcomes thus remain much as for the original unbracketed corpus—with the one exception of the *is-is* subject gap PIRCs which have plummeted from 87 percent to 31 percent correct. The reason is clear: the bracketing has broken up the previously influential <who|that is> bigram into <who|that]$_{NP}$> and <]$_{NP}$ is>. The former is in both test sentence versions, and so is the latter although at different sentence positions, so they are not distinguishing bigrams and cannot affect the outcome. The original striking success without brackets is thus reduced to the general rough-and-tumble of which particular item sequences happen to be better represented in the corpus.

Thus there is no indication here that NP brackets can solve the discrimination problem for the bigram learner. Although the NP brackets carry relevant information, a bigram model is unable to make good use of that information because it has too local a view of the sentence patterns.[13] Its problem is the same as before: there is a local oddity in *both* the grammatical and the ungrammatical word string, consisting of a non-finite predicate not immediately preceded by the sort of auxiliary that selects for it. The NP-bracketing adds only that what does precede the non-finite predicate is an NP. From a linguistic perspective, however, the relevant difference is that in the ungrammatical version what precedes the main predicate is a defective NP, while in the grammatical version it is a well-formed NP. These are distinguished in the next experiment.

3.4.1.2 *PS-experiment 2: Results and discussion* In PS-experiment 2 we supplied the model with the information it evidently could not compute for itself in the previous experiment: that an NP followed by a non-finite predicate is damaging to the sentence as a whole if it occurs in an RC inside an NP, but not if it is in the main clause. NPs containing an ill-formed RC were labeled with the *NP notation. The results in Table 3.5 show that there were now virtually no errors. The model overwhelmingly favored the grammatical sentence versions.

What caused rejection of the ungrammatical sentences in this experiment was not the * symbol itself (which has no meaning for the learning model), but the fact that, unlike all other unigrams in the test sentences, including $_{NP}$[and]$_{NP}$, the unigrams $_{*NP}$[and]$_{*NP}$ are not present in the corpus. (No utterances in the Bernstein-Ratner corpus were found to contain ungrammatical NPs.)[14] Standard treatment in cases where a unigram is unknown in the corpus is to assign it an estimated probability; we

[13] With trigrams, which have a wider compass than bigrams, results improved but were still unsatisfactory: 58% correct for subject gap; 52% for object gap; 47% for *do*-support. (See Kam 2009: ch. 4 for details.)

[14] We re-ran the experiment after inserting sixty ungrammatical NPs into the corpus, so that the unigrams $_{*NP}$[and]$_{*NP}$ had a positive probability without invoking the Witten-Bell formula. This made little difference: all three PIRC varieties showed 100% correct.

Table 3.5. Bigram model performance in PS-experiment 2.

Word string with NP and *NP-labeled brackets	% correct	% incorrect	% undecided
Is-is subject gap PIRCs	100	0	0
Is-is object gap PIRCs	100	0	0
Do-support PIRCs	99	1	0

did so using the Witten-Bell discounting technique (Witten and Bell 1991). However, the estimated probability is low relative to that of actually occurring unigrams, so its presence in the ungrammatical sentence can drag down the sentence probability, leading to preference for the grammatical version.

Together, these two experiments show that an *n*-gram-based learner could discriminate grammatical from ungrammatical PIRCs only if it could distinguish NPs from *NPs. Earlier, we postponed the question of whether and how it could do so. Now we must consider that.

3.4.2 *How to recognize *NP?*

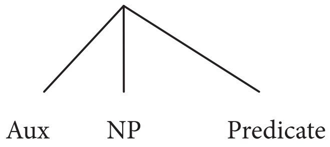
Presumably, the recognition that '*the boy who crying*' in (1b) is an ungrammatical noun phrase would have to occur during the process of assigning phrase structure to the sentence, based on recognition of '*who crying*' as an ungrammatical RC, missing an auxiliary. However, in the grammatical version (1a) there is also a missing auxiliary in the bigram <*]NP hurt*>. The absence of the needed auxiliary has a very different impact in the two cases: in (1b) it contaminates every larger phrase that contains it, while in (1a) it is amnestied by presence of the auxiliary at the start of the sentence. In general: since natural languages allow movement, absence of an obligatory item (a 'gap') in one location can be licensed by its presence elsewhere in the sentence. But there are constraints on where it can be. RCs are 'extraction islands', i.e., a gap inside an RC cannot be rescued by an item outside it (cf. the Complex NP Constraint of Ross 1967). By contrast, the main clause predicate is not an extraction island, so the lack of a needed auxiliary there *can* be rescued by association with the 'extra' auxiliary at the beginning of the sentence.

The notion of extraction islands has been refined and generalized as syntactic theory has progressed. In current theory, the contrast between legitimate and illegitimate movement is most often portrayed not in terms of specific constructions such as main clauses versus RCs but in terms of structural locality: local movement dependencies are favored over more distant ones by very general principles of economy governing syntactic computations. Deeper discussion of these matters within the framework of the Minimalist theory can be found in the chapters by Berwick, Chomsky, and

Piattelli-Palmarini and Chomsky in the present volume; see also the chapter by Rizzi and Belletti which shows locality/economy principles at work in child language.

By contrast with the transformational approach, recent discussions by Ambridge et al. (2008), Clark and Eyraud (2006), and Sag et al. (2003) suggest that as long as phrase structure is in place, the correct choice between grammatical and ungrammatical PIRCs follows even more naturally in a non-transformational theoretical framework, and hence might be even more readily accessible to a modest learning algorithm. In particular, a ternary structure for auxiliary inversion constructions, as in (9), is very simple, and would be frequently attested in the input in sentences such as *Is Jim hurt?*.

(9)



Once acquired, this analysis would automatically extend from *Is Jim hurt?* to *Is the little boy who is crying hurt?*. Without a transformational operation that moves the auxiliary from one site to another, there would be no question of moving it from the wrong location. Ungrammatical PIRC examples like (1b) would be simply ungeneratable. It might even be argued, contrary to stimulus poverty reasoning, that it is actually beneficial for learners that they would hear many simple questions like *Is Jim hurt?* before ever encountering a PIRC.

However, the grammar must not allow a sequence of Aux, NP, and a non-finite predicate to be freely generated. There is a selectional dependency which must be captured between the sentence-initial aux and the non-adjacent main clause predicate, as Sag et al. note. The predicate must be of a type that is selected for by the auxiliary; see (10).

(10) Is Jim running? *Is Jim run?
Jim is running. *Jim is run.

*Can Jim running? Can Jim run?
*Jim can running. Jim can run.

In a transformational framework this selectional dependency across the subject NP is captured by the assumption that the auxiliary originates adjacently to the predicate. In HPSG, without movement operations, a lexical rule manipulates the argument structure of the auxiliary. In declaratives its first argument (the subject) is realized preceding the auxiliary while its other argument (the non-finite predicate) follows the auxiliary. The lexical rule modifies this pattern so that in interrogatives both of the auxiliary's arguments follow it.

A lexical rule is inherently local since it manipulates the argument structure of one lexical head. Therefore an error such as (1b), spanning two clauses, can never arise.

Note, however, that this nice solution to the auxiliary inversion learnability problem only holds if it is *necessary* for auxiliary inversion to be captured by a lexical rule. If not, there is still a risk of a learning mis-step, even in the HPSG framework. Long-distance phenomena such as wh-'movement' or topicalization cannot be handled by lexical rules. HPSG treats them by means of a different formal device: GAP features are passed through the tree, from one node to another, between the 'gap' position and the surface position of the item. While there are some constraints on the inheritance of GAP features, there is no bound on how far a GAP feature can be passed.

Therefore, an HPSG-based learner that encountered questions in the input, even simple questions like *Is Jim running?*, would have to choose between formulating a lexical rule, which is local, or establishing GAP feature-passing for auxiliaries. If preference for a lexical rule were innate, then indeed a learner's grammar could not license displacement of the 'wrong' auxiliary as in (1b). But if a learner could opt for a GAP-feature analysis of simple questions, then errors like (1b) could ensue on PIRCs. To prevent this, an innate constraint would be needed on GAP feature passing, comparable to the locality constraint needed in a transformational system: despite formal differences, both theories must make the RC an extraction island. (For discussion of complex NP islands in HPSG, see Pollard and Sag 1994: ch. 5.)

3.5 Conclusions

This study of the prospects for *n*-gram-based learning of natural language syntax leads to the following conclusions:

(I) Low-level statistics over word strings might contribute to syntax learning but cannot substitute for syntactic knowledge.

(II) Specifically: such statistics cannot capture the generalization about auxiliary inversion.

(III) Theoretical differences aside, the only route to the correct generalization requires a bias toward local syntactic dependencies, defined over a phrase structure analysis of the sentence.

(IV) Hence, a learner that makes use of word-level statistics as the basis for auxiliary inversion must, at a minimum, also have an innate propensity to project phrase structure onto word strings—just as Noam Chomsky observed four decades ago.