Statistical Induction in the Acquisition of Auxiliary Inversion

Xuân-Nga Cao-Kam
The Graduate Center, City University of New York

1. Introduction

Important claims have recently been made concerning the role of statistical learning in child language acquisition. In parallel lines of research, young children have been shown to be sensitive to statistical regularities in strings of syllables (e.g., Saffran & Wilson, 2003), and computational linguists have begun to explore what aspects of a natural language can be acquired by a learning model that relies on statistical counts of syllable or word co-occurrences and transitional probabilities. If a simple model, with resources within the bounds that it is plausible to ascribe to young children, can attain mastery of realistically complex natural language constructions, that would undercut long-standing claims that innate knowledge is essential for language acquisition (Chomsky 1980).

A study by Reali & Christiansen (2005; henceforth R&C) illustrates this trend in computational linguistic research. They tested extremely simple learning models: a bigram language model whose sole source of information about the language is the incidence of pairs of adjacent words in sentences in the input sample; and a trigram model that uses similar information about word triples. By contrast, the target construction for acquisition was complex: auxiliary inversion in multi-clause sentences. The training corpus was relatively small, and consisted of a transcript of spontaneous speech by adults to children, segmented into words but not otherwise edited or annotated. The children were very young, all under two years, and therefore unlikely to be addressed in complex language. Indeed, there were no examples of the target construction in the training corpus, so successful learning by the model would have to be based on 'indirect evidence' obtained from other (probably simpler) constructions that did occur in the corpus. Thus, this was in every way an ambitious test of the power of statistical learning.

As detailed in the next section, R&C's learning models did show strong evidence of being able to identify grammatical forms of the target construction. However, a subsequent study by Kam et al. (2005; submitted) showed that this success was limited in scope. The grammatically correct form of auxiliary inversion was recognized only when the test items had certain very specific properties; there was no indication of knowledge of auxiliary inversion as a general process. It appears, then, that these simple n-gram based models are not, after all, sufficient for acquiring complex syntactic patterns in natural language, at least when applied without assistance of either innate guidance or more richly encoded input. This means that more work is needed in order to establish a

lower bound on the resources required for natural language syntax acquisition. The aim of Kam et al. was to understand how the simple n-gram models achieved the knowledge they did, and why that knowledge did not generalize. My research strategy in this paper will be 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. In the experiments reported in sections 4 and 5 below, I enriched the corpus in various ways while leaving the learning model constant, and I enriched the learning model while leaving the corpus constant. To anticipate: The results showed that these manipulations brought surprisingly little benefit, indicating that more sophisticated learners and/or input representations must be explored in future research.[1]

2. Auxiliary inversion as the acquisition target

Acquisition of auxiliary inversion is of interest in its own right, but it has gained particular significance in the acquisition literature as an illustration of Chomsky's (1980) observation that children's linguistic input does not uniquely determine the correct grammar rule that they should formulate. Chomsky noted that simple examples of auxiliary inversion as in (1) are compatible with two distinct rules, which would have quite different consequences for bi-clausal examples as in (2) and (3).

(1) Is$_i$ the little boy t_i hurt?
(2) Is$_i$ the little boy [who is crying] t_i hurt?
(3) *Is$_i$ the little boy [who t_i crying] is hurt?

I will refer to the two-clause construction in (2) as a *PIRC*, an abbreviation for a polar interrogative (*PI*) with a relative clause (*RC*) modifying its subject. (2) is the correct form, with fronting of the main-clause auxiliary; (3) with fronting of the auxiliary in the relative clause is incorrect. Chomsky's point (in somewhat updated terms here) was that a learner's natural tendency to minimize the distance moved by the auxiliary would yield (2) if distance is measured over hierarchical structure (since the main clause auxiliary is closer to the landing site at the top of the structure) but would yield (3) if distance is measured over the linear string of words (since the relative clause auxiliary is closer to the landing

[1] To anticipate later discussion (Section 6), I note that R&C also observed successful performance by a considerably more powerful Simple Recurrent Network (SRN). However, it was tested only on the same limited examples of auxiliary inversion as the n-gram models. This was also true of the SRN study by Lewis & Elman (2001). No data *at present* show that a network model achieves a more general grasp of auxiliary inversion than n-gram learners.

site at the beginning of the string).[2] Crain & Nakayama (1987) showed that though children (age 3;2 - 5;11) made a variety of errors in producing PIRCs, none were of a kind that implied that they had formulated a non-structure-dependent rule for auxiliary inversion. On the assumption that children are not exposed to well-formed examples of PIRCs such as (2), nor to evidence that forms like (3) are incorrect, Chomsky concluded that children must have an innate bias towards assigning hierarchical phrase structure to word strings they hear. However, the necessity of innate guidance would clearly be undermined if it were demonstrated that an algorithm which has been given no innate bias toward hierarchical structure could also reliably select grammatical forms like (2) rather than the ungrammatical (3). Such a finding would indicate, to the contrary, that the learner's *input* contains sufficient information to select the grammatical form. This was the hypothesis underlying R&C's study, and the successful performance of their n-gram models indicated not only that such information is present in children's input but also that a very simple statistical model can extract it.

3. Previous statistical modeling of acquisition of auxiliary inversion

In R&C's experiment, the learning model was required to choose between grammatical and ungrammatical PIRCs like (2) and (3). However, as in Chomsky's argument, the model was not exposed to any examples of the grammatical form, nor to negative evidence against the ungrammatical form. To succeed in the task, the model must have induced the correct form from *indirect* evidence in its language sample, which would most likely have to be gleaned from auxiliary inversion in simple polar interrogatives (e.g. *Is the little boy hurt?*), and relative clauses in declarative sentences (e.g. *The little boy who is crying is hurt*).

The design of this experiment will be important to the discussion below. 10,705 utterances of child-directed speech were extracted from the Bernstein-Ratner (1984) corpus. 100 well-formed PIRC sentences were constructed from words in the corpus in accord with the template (4a), each paired with a corresponding ungrammatical version as defined by template (4b). Examples (2) and (3) above fit this pattern.

(4) a. Grammatical Is NP {who|that} is A B ?

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

where A and B are instantiated by VP, PARTICIPLE, NP, PP, ADJP, etc.

---

[2] Not all linguistic theories posit movement operations as in Chomsky's transformational framework, but in other ways they still capture the fact that the sentence-initial auxiliary has to be one that is compatible with the main clause and 'missing' in some sense from its predicate, not one that is compatible with and missing from the relative clause.

Presented with a pair of test sentences, the learning model selected the more probable of the two as being the grammatical sentence. The computation of sentence probability was based on the probabilities of the n-grams it contains. A bigram is a sequence of two adjacent words, w_1 and w_2, in a sentence. The bigram probability with respect to a corpus is defined as the probability of w_2 given w_1. A trigram is a sequence of three consecutive words and its probability is the probability of w_3 given w_1 and w_2. All words (unigrams) in the test sentence pairs were in the corpus (since what was being tested was the model's combinatorial ability, not its vocabulary knowledge), but not all of the bigrams/trigrams in the test pairs were in the corpus. Because there were non-attested n-grams, R&C used an *interpolation smoothing* technique. The formulae in (5a,b) are for the smoothed probability (P_{interp}) for a bigram and trigram respectively, where "$count_{total}$" denotes the total number of unigram tokens in the corpus, and λ was set at 0.5 so that all terms are equally weighted.

(5) a. $P_{interp}(\text{bigram } w_1w_2) = \lambda(count_{w1w2}/count_{w1}) + (1-\lambda)(count_{w2}/count_{total})$

b. $P_{interp}(\text{trigram } w_1w_2w_3) = \lambda\,(count_{w1w2w3}/count_{w1w2}) + (1-\lambda)\,(\lambda\,(count_{w2w3}/count_{w2}) + (1-\lambda)(count_{w3}/count_{total}))$

Sentence probability was computed as the product of the smoothed probabilities of all the bigrams/trigrams in the sentence. (All test pair comparisons were in fact made in terms of sentence *cross-entropy*, a measure inversely related to sentence probability and more convenient to work with.) On this basis, both the bigram model and the trigram model consistently identified the correct form of PIRC: they selected the grammatical version in 96 of the 100 test pairs. In what follows I will first focus on the bigram model, and will return to the trigram model in section 5.

In Kam et al. (2005) we set ourselves the task of identifying what cues the bigram model was picking up from the input in order to achieve its success. First, we replicated R&C's experiment, using the same corpus and methodology as they did and constructing the test sentences using the same templates. The outcome was similarly successful, with the model predicting 87% of the test sentences correctly.

On looking more closely for the basis of the correct predictions, we found that successful discrimination rested almost entirely on one bigram in the grammatical version. Many of the bigrams in the test sentences did not contribute to the discrimination because they were identical in the grammatical and ungrammatical versions. In sentences (2) and (3), for example, the bigrams at the beginning of the sentences are identical: <*is the*>, <*the little*>, <*little boy*> and <*boy who*>. Only the subsequent bigrams differ. The grammatical version has <*who is*>, <*is crying*> and <*crying hurt*> while the ungrammatical version has <*who crying*>, <*crying is*> and <*is hurt*>. I will refer to these bigrams which differ between the two versions of a test pair as the *distinguishing bigrams*.

Table 1: Distinguishing bigrams for the test sentence pair (2)/(3).

Test sentences	bigram1	bigram2	bigram3
(2) grammatical	<*who is*>	<*is crying*>	<*crying hurt*>
(3) ungrammatical	<*who crying*>	<*crying is*>	<*is hurt*>

The majority of correct choices (80 out of 87) were found to be due to the contribution of the distinguishing bigram <*who is*> or <*that is*> in the grammatical version. As prescribed by template (4a), one or other of these bigrams appeared in every grammatical test sentence. These bigrams had a higher smoothed probability than most of the other bigrams in the test sentences, and so they tended to dominate the outcomes. Note that the bigram <*who is*> or <*that is*> (henceforth <*who*|*that is*>) is a local cue at the lexical level. It does not presuppose assignment of hierarchical structure to the word string or recognition of any syntactic dependency between the uninverted and inverted positions of the auxiliary. In other words, the bigram model is able to succeed on this test without needing to learn about the structural properties of PIRCs at all; it can rely instead on a simple co-occurrence of two words. However, this raises the question of whether it can recognize the grammatical form of PIRCs which lack this useful <*who*|*that is*> bigram. Are there, in those cases, some alternative local cues which are equally effective? This is what the experiments in the Kam et al. study were designed to establish.

The English aux-inversion rule (front the highest aux) applies quite generally to polar interrogatives, regardless of whether the auxiliary is "is" or "will" or "can", or whether there is a single auxiliary in the relevant clause or two or more (e.g., "must have been"), or whether the auxiliary in the RC immediately follows the relative pronoun or is separated from it by an intervening subject (e.g., "Is the boy who Jill is helping tired?"). The templates in (4) that defined the test sentences for R&C's experiment and our replication of it pick out just a subset of PIRCs. They specify that the auxiliary in both clauses of the construction is "is", and that the RC has a subject gap, i.e., the relative pronoun fills the subject role in the clause. This ensures that the grammatical version of every test pair contains <*who*|*that is*>. However, Table 2 shows a few of the many kinds of well-formed examples that do not contain <*who*|*that is*>.

Table 2: Examples of English auxiliary inversion missing <*who*|*that is*>.

Sub-types of PIRC	Examples
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?

Kam et al. tested is-is object gap and do-support PIRCs and found poor discrimination in both (see Table 3 below), indicating that the bigram model had

not found any reliable alternative cues to compensate for the lack of the <*who*|*that is*> bigram which served as a reliable marker of the grammatical form in the original experiments with is-is subject gap PIRCs.

Table 3: Discrimination by the bigram model, for three varieties of PIRC.

Kam et al.'s experiments	% correct	% incorrect	% undecided
Is-is subject gap	87	13	0
Is-is object gap	35	15	50
Do-support	49	51	0

4. Augmenting the resources for learning

It is not known at what age children have mastered the full auxiliary inversion generalization in English; there have been no empirical studies investigating this. (See Ambridge et al., under revision, for acquisition of PIRCs with "can".) But normal adults have general competence with regard to auxiliary inversion, so it evidently is learnable by humans. The bigram model in the Kam et al. experiments failed to learn it, but that of course does not mean that all learning models of its type must fail. Some relatively minor adjustment of the learning mechanism or the linguistic input might tip it over into successful performance. Recall 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. Or 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 expanded corpora and with a trigram learning model, aimed to improve the 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]

4.1. Enriching the corpus: Experiments 1 and 2

Experiment 1 employed 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 2 used an even larger corpus of 110,629 adult utterances, 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

[3] Linguistic analyses of *do*-support questions differ with respect to whether what moves is the word "do" itself, or a tense morpheme or feature which then needs insertion of "do" to support it. I believe that nothing relevant to n-gram-based learning hangs on this.

corpora were from Brown (1973), MacWhinney (2000), Menn & Gleason (1986), Suppes (1974), Valian (1991) and Warren-Leubecker (1984). In both 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 similar to those of Kam et al. except for some word replacements so that all test sentence unigrams occurred in the training corpus.

Results are shown in Table 4. The model's discrimination of grammatical PIRCs did improve under these conditions, but there were still from 30% to 53% of errors, a far cry from the performance on the original is-is subject-gap PIRCs.

Table 4: Discrimination by the bigram model in Experiments 1 and 2.

Test sentence type	% correct	% incorrect	% undecided
OBJECT-GAP PIRC			
Original corpus	35	15	50
Older corpus	47	45	8
Larger and older corpus	63	30	7
DO-SUPPORT PIRC			
Original corpus	49	51	0
Older corpus	55	45	0
Larger and older corpus	70	30	0

One important improvement was for the object-gap PIRCs, where the older/ larger corpora greatly reduced the number of undecided cases. Undecided cases occur when none of the bigrams in a test pair are attested in the corpus; their smoothing factors may then cancel out across the grammatical and ungrammatical versions. (This can happen in the object-gap items but not in the do-support items; see Kam et al., submitted, for details.) Since an older and/or larger corpus is likely to contain more bigram types as well as tokens, undecided cases arise less often. However, note that within the decided cases, the older/larger corpora actually yielded a lower proportion of correct judgments for object-gap PIRCs (51% and 68%) than with the original corpus (70%). For do-support, the richer corpora brought some benefit. But all in all, the PIRC constructions that were troublesome in the Kam et al. study still did not show decisively good learning even when the size of the corpus was increased tenfold and the children addressed were up to 8 years old.

Though perhaps unexpected, this outcome is explicable. A larger corpus tends to yield higher counts of the bigrams in the test sentences, but this is true of bigrams in *both* sentence versions so it would not necessarily favor the grammatical version. It is also relevant that there is no bigram that appears systematically in these test sentences (like the <*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. 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 RC 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. Bigrams containing open-class items (such as <*sister is*> or <*who plays*>) tend not to be numerous even in large corpora. Enriching the corpus would typically increase their counts less dramatically than for a bigram consisting of only closed class items, such as <*who*|*that is*>. Moreover, a higher bigram count does not necessarily entail a higher smoothed probability for that bigram as established by formula (5a) 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.

4.2. Providing syntactic category information: Experiments 3 and 4

It appears that the bigram model may be buffeted by unstable frequencies of specific words and word combinations because of its reliance on input in the form of superficial word strings. Another way to improve its performance might therefore be to inject into the input corpus some more syntactically relevant word-class information.

In Experiments 3 and 4 the learning model and procedure were the same as in the Kam et al. study but the sentence representations differed. Starting with the same corpus and test sentences as Kam et al., all the words in them were replaced in Experiment 3 by their part-of-speech tags, using the MOR program (available in the Childes database; MacWhinney, 2000) with 117 distinct part-of-speech tags. For instance, sentence (6) from the corpus was converted into the string (7). In experiment 4, 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. Thus sentence (6) was replaced by (8).

(6) you want to see the book
(7) pro v inf v det n
(8) you v to v the n

There is a double rationale for representing sentences in terms of the syntactic categories of their words, rather than the words themselves. One advantage is that helps to solve the problem of the sparseness of data, which was not greatly improved by the shift to a considerably larger corpus in Experiment 2. When specific words are aggregated into more general syntactic categories, bigram counts are increased and become more reliable. Moreover, the greatest beneficiaries are bigrams containing open-class items; e.g., <*sister is*> would be

represented in Experiment 3 as <*n v:aux&3S*>, a much more frequent bigram. Bigrams containing content words can then 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 <*who|that is*>. Another reason for moving to the more abstract part-of-speech representation is that syntactic patterns in natural language are defined in terms of such categories, not in terms of specific lexical items. Therefore, providing syntactic category information to the learning model could be expected to increase its ability to capture a general structural pattern for auxiliary inversion.

The motive for keeping function words unchanged in Experiment 4 derives from the finding by Mintz (2003) that these 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 4 could be especially helpful since it groups lexical items into broader categories without losing the specific information carried by individual function words.

Table 5: Discrimination by the bigram model in Experiments 3 and 4.

Test sentence type	% correct	% incorrect	% undecided
OBJECT-GAP PIRC			
Kam et al. (word level)	35	15	50
POS-tags only	51	46	3
POS-tags + function words	41	55	4
DO-SUPPORT PIRC			
Kam et al. (word level)	49	51	0
POS-tags only	70	30	0
POS-tags + function words	64	36	0

Results are shown in Table 5. Contrary to expectations, the mixed representation resulted in somewhat worse performance than the full part-of-speech tagging for both object-gap and do-support PIRCs. This suggests that gains due to formation of larger groupings with increased bigram counts tend to outweigh gains due to the specific signposts to syntactic structure provided by the function words in the mixed representation.

Both representation schemes fared better than the purely word level representation of the original experiments, but performance was still not strong, presumably because of the varied effects on bigram probabilities as corpus frequencies increase, as noted for Experiments 1 and 2. Sometimes a distinguishing bigram in the grammatical test sentence benefited most, and

sometimes one in the ungrammatical test sentence. Error rates were substantial, ranging from 30% to 59%, too high to be explained away as due to occasional performance errors by a learner that has mastered the basic rule. Also as for Experiments 1 and 2, while the object-gap PIRCs benefited by losing most of their 'undecided' cases, the proportion of decided cases that were decided correctly actually declined, from 70% in the original experiment to 53% and 43% in Experiments 3 and 4 with part of speech tags. (Circumstances in which more abstract sentence representations can be less effective for learning than lexical representations have been noted by Chang et al. 2006 and Perfors et al. 2006, so this finding may repay further attention.) While exact outcomes may vary with the particular set of categories employed, these lackluster improvements due to part of speech representation seem to suggest that the essential core of auxiliary inversion cannot be captured as a sequence of lexical categories any more than it can as a sequence of lexical items.

5. Enriching the learning model: Experiment 5

Augmenting the original corpus as in the previous four experiments yielded some improvement but did not result in high levels of discrimination. Additional enhancements of the input might boost performance further, but it is reasonable to consider the possibility that it is not the input that is holding back performance levels, but the use that the bigram model makes of the input information available. Therefore, another step in the attempt to facilitate PIRC learning is to upgrade the learning model.

Experiment 5 reverted to the original 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. Many other enhancements of the learning algorithm could be tried (e.g., different smoothing techniques, or more elaborate computations over bigrams), but this modest upgrade from bigrams to trigrams is a first step forward, in keeping with the research strategy of incremental supplementation of the learning situation. Further additions to the computational power of the learning mechanism are planned for future research.

In fact, trigram-based discrimination proved to be no more successful than bigram-based performance. Table 6 below shows that the results were very similar, with respect to both percent correct and percent undecided. The explanation appears to be that the broader scope of the trigram statistics was counterbalanced by their lower frequency in the corpus. 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. When a trigram does not occur, the smoothing factor in formula (5b) substitutes half of the smoothed bigram probability. The results therefore differ very little.

Table 6: Discrimination by the trigram model in Experiment 5, compared with bigram model discrimination for the same constructions.

Test sentence type	% correct	% incorrect	% undecided
SUBJECT-GAP PIRC			
Kam et al. 2005: bigrams	87	13	0
Experiment 5: trigrams	80	20	0
OBJECT-GAP PIRC			
Kam et al. 2005: bigrams	35	15	50
Experiment 5: trigrams	36	16	48
DO-SUPPORT PIRC			
Kam et al. 2005: bigrams	49	51	0
Experiment 5: trigrams	48	52	3

This suggests that the move from bigram to trigram statistics would be more beneficial in combination with a much larger corpus containing a higher proportion of the test sentence trigrams. However, 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 in any corpus. More significantly, perhaps, the object-gap and do-support PIRCs do not contain any distinguishing trigram that reliably favors the grammatical version. As noted above, in such a case there may be little to be gained from increased corpus counts; but this needs to be confirmed in a further experiment.

6. General Discussion

If a simple n-gram model had acquired a general command of auxiliary inversion from a fairly modest corpus, that would have been a striking result and would have made it clear that neither innate linguistic knowledge (Universal Grammar) nor sophisticated learning mechanisms are needed for natural language syntax acquisition. But this was not the case in the original experiments of R&C and Kam at al, and it was still not so with the enrichments examined in the present series of experiments. With or without more substantial input, the n-gram models achieved solid knowledge of only *one* sub-variety of PIRCs, arbitrarily culled from the much wider array of auxiliary inversion sentences in English. Once a more representative set of examples was included among the target sentences for acquisition, weaker performance was observed. It can be concluded that the input sample did not provide the n-gram models with sufficient cues for grammatically correct auxiliary inversion in the do-support or object-gap contexts; either such information was lacking in the corpus or else the statistical measures were too weak to extract it. It can also be concluded that these models had not acquired any general rule for auxiliary inversion: they were unable to *project* a pattern learned for one sub-type of the construction (is-is with subject-gap) to other instantiations of it (is-is with object-gap; do-

support). Thus neither piecemeal learning of specific forms nor generalization across instances was available. Most theories of child language acquisition, by contrast, assume that both these mechanisms are available.

Thus, the search for the minimum resources necessary for natural language syntax acquisition must continue in future research. Even within the class of data-driven statistical learners with no innate linguistic knowledge, there are many kinds of models to be explored that would be more powerful than those examined here. The question of interest is how far up the scale of computational power it is necessary to go before learning comparable to that of children is attainable. At the farther end of the scale, connectionist models such as simple recurrent networks (SRNs) have been shown to be capable of mastering advanced language phenomena such as self-embedded constructions (Elman, 1993). An SRN can successfully learn to favor grammatical over ungrammatical is-is subject-gap PIRCs (Lewis & Elman, 2001; Reali & Christiansen, 2005), but SRNs have not been tested, to date, on a wider range of PIRCs. The Kam et al. experiments and those presented here show clearly that competence with respect to is-is subject-gap PIRCs does not entail the ability to acquire auxiliary inversion in general. Thus the syntactic capabilities of neural networks in this regard will not be known until more stringent tests on a broader range of examples have been conducted.

It should be emphasized that although acquisition was by and large unsuccessful in these experiments, this does not mean that the experiments were without value. Informative conclusions can be drawn from documented learning failures just as they can from learning successes. For obvious reasons, failures are less often published. This is unfortunate insofar as it deprives acquisition research of knowledge about which learning models fail on which language phenomena under what conditions. For the goal of circumscribing the mental mechanisms by which children acquire language, such information is equal in importance to information about when and how learning succeeds.

Finally, one might speculate on what properties *would* enable a statistical learning model to identify abstract syntactic patterns in a corpus. Chomsky (1980) observed that once phrase structure is assigned to word strings, a simple rule can be formulated which applies very generally to generate auxiliary inversion constructions of all kinds. Other linguistic theories concur on this, even if they disagree about precisely how the rule should be formulated. A learner that acquires the general pattern would not need to seek out different cues for auxiliary inversion in different contexts, such as in PIRCs with subject-gaps versus object-gaps in their relative clauses. It is a plausible conjecture, therefore, that a statistical learning system could succeed on auxiliary-inversion-in-general only if it is powerful enough to acquire phrase structure.

References

Ambridge, B., Rowland C.F. & Pine, J.M. (2006 ms., under revision) Structure dependence: An innate constraint? *Cognitive Science.*

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. (2006). Using child utterances to evaluate syntax acquisition algorithms. *Proceedings of the Cognitive Science Society*, Vancouver, Canada

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

Crain, S., & Nakayama, M. (1987). Structure dependence in grammar formation. *Language, 63, 3,* 522–543.

Elman, J. L. (1993). Learning and development in neural networks: The importance of starting small. *Cognition, 48*, 71–99.

Kam, X. N. C., Stoyneshka, I., Tornyova, L., Fodor, J. D. & Sakas, W. G. (2005). Non-robustness of syntax acquisition from n-grams: A cross-linguistic perspective. Paper presented at the 18th Annual CUNY Sentence Processing Conference.

Kam, X. N. C., Stoyneshka, I., Tornyova, L., Fodor, J. D. & Sakas, W. G. (2006 ms., submitted). Bigrams and the richness of the stimulus.

Lewis, J. D., & Elman, J. L. (2001). Learnability and the statistical structure of language: Poverty of stimulus arguments revisited. *Proceedings of the 26th Boston University Conference on Language Development*. Somerville, MA: Cascadilla Press.

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.

Mintz, T. H. (2003). Frequent frames as a cue for grammatical categories in child directed speech. *Cognition, 90,* 91–117.

Perfors, A., Tenenbaum, J. & Regier, T. (2006). Poverty of the stimulus? A rational approach. *Proceedings of the Cognitive Science Society*, Vancouver, Canada.

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

Saffran, J. R., & Wilson, D. P. (2003). From syllables to syntax: Multi-level statistical learning by 12-month-old infants. *Infancy*, *4,* 273–284.

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

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

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