Learnability and the Statistical Structure of Language: Poverty of Stimulus Arguments Revisited

John D. Lewis and Jeffrey L. Elman
UC San Diego & McGill University, and UC San Diego

1. Introduction

Statistical learning, and "*any account which assigns a fundamental role to segmentation, categorization, analogy, and generalization*" is rejected in Chomskyan linguistics as "*mistaken in principle*" (Chomsky, 1975). Acquisition is viewed, rather, as a search through the set of possible grammars for natural language, guided by successive inputs; or alternatively, as a parameter setting process in which the inputs serve as triggers. The stochastic nature of the input is thus ignored — supposedly the learner is oblivious to the distributional frequencies of lexical items, grammatical constructions, and utterance types.

Recent acquisition research, however, has shown that children, and even infants, *are* sensitive to the statistical structure of their linguistic input (Saffran et al., 1996; Aslin et al., 1998; Gomez and Gerken, 1999; Newport and Aslin, 2000).

The situation with respect to learnability is thus significantly different from that which has been assumed. Stochastic languages may be learnable from positive examples alone, while their non-stochastic analogues require negative evidence (Gold, 1967; Horning, 1969; Angluin, 1988). Indeed, as Chomsky (1981) observed, distributional information can provide "*a kind of 'negative evidence'*" in that expectations can be formed which may then be violated. And so, in at least some cases, the so-called 'logical problems' associated with the *no negative evidence* hypothesis may be solved by admitting the stochastic information.

Thus, if UG is to account for all and only those properties of language "*that can reasonably be supposed not to have been learned*" (Chomsky, 1975) we must adopt a learning theory which is sensitive to the statistical properties of the input, and reassess poverty of stimulus arguments under those theoretical assumptions.

This paper illustrates this by showing that the "*parade case of an innate constraint*" (Crain, 1991) — *i.e.*, Chomsky's (1975) poverty of stimulus argument that *structure dependence* must be a principle of UG — fails to hold once stochastic information is admitted; the property of language in question is shown to be learnable with a statistical learning algorithm.

Chomsky (1975) suggests that it is reasonable to suppose that *aux*-questions are derived from declaratives, and so children, presumably exposed to only simple forms of either type, should be free to generate either of two sorts of rules: a structure-independent rule — *i.e.* move the first '*is*' — or the correct structure-dependent rule. Chomsky argues that since "*cases that distinguish the hypotheses rarely arise*," (Piatelli-Palmarini, 1980) at least some children can be assumed not to encounter the relevant evidence for a considerable portion of their lives. Thus, since the structure-independent hypothesis generates ungrammatical forms like (2) in place of the correct (1), children should be expected to make

such mistakes. Since they do not (Crain and Nakayama, 1987; Crain, 1991), despite that the correct rule is supposedly more complex, Chomsky suggests

1) *Is the man who is smoking crazy?*

2) **Is the man who smoking is crazy?*

that "*the only reasonable conclusion is that* UG *contains the principle that all such rules must be structure-dependent*" (Chomsky, 1975) — *i.e.* that during the course of language acquisition, children must entertain only hypotheses which respect the abstract structural organization of language, though the data may also be consistent with structure-independent hypotheses.

This conclusion depends, of course, on more assumptions than just that the input available to children does not reliably contain questions like "*Is the jug of milk that's in the fridge empty?*" — an assumption that has been noted to be somewhat questionable (Cowie, 1998). It is apparently also assumed that the learner makes use of no form of generalization whatsoever; for as Sampson (1989) has pointed out, evidence to distinguish the two hypotheses is provided by any utterance in which *any* auxiliary precedes the main clause auxiliary. And so evidence is also provided by questions like "*Is the ball you were speaking of in the box with the bowling pin?*", and "*Where's this little boy who's full of smiles?*", and even "*While you're sleeping, shall I make the breakfast?*" — all of which are from CHILDES (MacWhinney, 2000)[1], and presumably instances of structures which are not overly rare in child-directed speech. Indeed, Pullum and Scholz (2001) estimate that such examples make up about one percent of a typical corpus. Thus, since learners receive approximately three years of language exposure before they exhibit the knowledge in question, it is perhaps unrealistic to assume that they will not encounter evidence of this sort.

Here we show, however, that even in the total absence of the above sort of evidence, the stochastic information in data uncontroversially available to children is sufficient to allow for learning. Building on recent work with simple recurrent networks (SRNs; Elman 1990), we show that the correct generalization emerges from the statistical structure of the data.

Figure 1 shows the general structure of an SRN. The network comprises a three-layer feed-forward network — made up of the input, hidden, and output layers — augmented by a context layer, identical in size to the hidden layer, and connected to it bi-directionally. Each layer consists of a number of simple processing units which are connected to the units in other layers as indicated in the figure; *i.e.*, each of the units in the input and context layers connects to every unit in

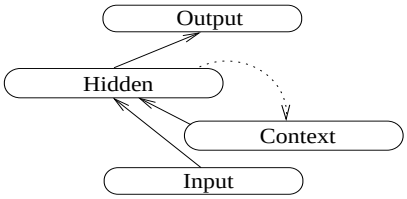

Figure 1: An SRN. Solid lines represent full connectivity; the dashed line indicates a copy connection.

[1]These examples are from Brown's Adam, Korman's St, and Manchester's Anne, respectively.

the hidden layer, and each unit in the hidden layer connects to its corresponding unit in the context layer, as well as to every unit in the output layer.

An input to the network is a set of activation values on the input units, and the corresponding network output is the set of activation values that result on the output units. These values are computed in the same manner as with all units in the network: the activation value of a unit is the sum of the activation values of every unit that connects to it — *i.e.*, its *net input* — squashed by the logistic function to fall between 0 and 1.

The recurrent connections between the hidden and context layers provide a one-step state memory. At each time step the activation value of each of the hidden units is copied to the corresponding unit in the context layer, and the connections from the context layer back to the hidden layer make these values available as additional inputs at the next time step.

The network receives its input sequentially, and learns through prediction: it attempts to predict the next input at each step, and utilizes its prediction errors to correct its connection weights. At the outset of training, the connection weights and activation values are random, but to the extent that there are sequential dependencies in the data, the network will reduce its prediction error by building abstract representations that capture these dependencies. Structured representations thus emerge over time as a means of minimizing error; and such representations, to the extent that they provide for accurate predictions, can be thought of as a grammar.

Networks of this sort have been shown capable of learning solutions to two problems that we suggest are relevant to Chomsky's structure-dependence argument.

Elman (1991, 1993) provided such a network[2] with a corpus of language-like sentences which could be either simple (transitive or intransitive), or contain multiply embedded relative clauses (in which the head noun could be either the subject or object of the subordinate clause). The input was presented as word sequences, where words were represented as orthogonal vectors — a localist representation — so that no information about either the words or the grammatical structure was supplied; thus the network had to extract all information (*e.g.*, the grammatical categories, number agreement, subcategorization frames, and selectional restrictions) from regularities in the input. The network learned the structure of such sentences so as to predict the correct agreement patterns between subject nouns and their corresponding verbs, even when the two were separated by a relative clause with multiple levels of embedding, *e.g.* *boys who like the girl who Mary hates hate Mary*.[3,4]

[2]The network used differed only in that reduction layers were added between the input and hidden layers, and the hidden and output layers. This addition allowed the localist representations — strings of 0s with a single bit set to 1 — used for the inputs and outputs, to be re-represented in distributed form for the mappings to and from the hidden layer, and also reduced the overall number of connections in the network.

[3]The network succeeded only if either the input was structured, or the network's memory was initially limited, and developed gradually.

[4]An SRN's performance with such recursive structures has also been shown to fit well to the human data (Christiansen and Chater, 1999).

Such networks have also been shown to go beyond the data in interesting ways. Elman (1998) and Morris et al. (2000) showed that SRNs induce abstract grammatical categories which allow both distinctions such as *subject* and *object*, and generalizations such that words which have never occurred in one of these positions are nonetheless predicted to occur, if they share a sufficient number of abstract properties with a set of words which have occurred there.

Together these results suggest that an SRN might be able to learn the structure of relative clauses, and generalize that structure to subject position in *aux*-questions — and thus to learn the aspect of grammar in question despite not having access to the sort of evidence that has been assumed necessary. This paper reports on simulations which show that this is indeed the case. An initial experiment verifies that the two results combine in the required way; then an SRN is shown to predict (1), but not (2), from training sets based on CHILDES data. This result clearly runs counter to Chomsky's argument, and thus indicates that the amended view of the input necessitates a re-evaluation of all previous poverty of the stimulus arguments — and that neural networks provide a means of doing this.

2. A Hint of Verification

To verify that the sort of generalization suggested is in fact possible, we trained an SRN on data from an artificial grammar which generated only *a) aux*-questions of the form 'AUX NP ADJ ?', and *b)* sequences of the form 'A_i NP B_i', where A_i and B_i stand for sets of inputs with random content and length. Proper names, pronouns, and NPs of the form 'DET (ADJ) N (PP)' were generated in both types, and NPs with relative clauses were generated for the 'A_i NP B_i' type, but were restricted from appearing in *aux*-questions. Some representative examples are given in Figure 2.

Representing the input in this way ensures 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; any other structure in the data has been abstracted away.

The training regime was similar to that used by Elman (1991, 1993). A three-stage training set was generated from the grammar, with the degree of complexity in NPs increasing at each stage, and the percentage of *aux*-questions decreasing — crudely approximating the structure of child-directed speech. Names and pronouns constituted 80% of the NPs in the first set, and the remaining 20% was

A_i *Mummy* B_i	*is Mummy beautiful?*
A_i *she* B_i	*is she happy?*
A_i *the dog* B_i	*is the dog hungry?*
A_i *the little girl* B_i	*is the little girl pretty?*
A_i *the cat on the mat* B_i	*is the cat on the mat fat?*
A_i *the big dog in the car* B_i	*is the big dog in the car scary?*
A_i *the boy who is smiling* B_i	* *is the boy who is smiling nice?*

Figure 2: Examples of various types of utterances generated by the artificial grammar. The asterisk indicates the absence of *aux*-questions with NPs that contain relative clauses.

shared among the other NP forms (such that the more complex the form, the fewer the instances of it), with relative clauses making up only 1%; there were 40% *aux*-questions, and 60% 'A_i NP B_i' forms. In the second set, names and pronouns constituted 70% of the NPs, relative clauses made up 2.5% of the remainder, and the percentage of *aux*-questions decreased to 30%. And in the third set, 60% of the NPs were names or pronouns, relative clauses made up 5% of the remainder, and the percentage of *aux*-questions decreased to 20%. Each training set consisted of 50,000 examples. An SRN was trained on each set successively, for 10 epochs each, and tested with the structures in (1) and (2) after each epoch.[5] The network received the input one word at a time, and in the same form as used by Elman (1991, 1993) — *i.e.*, a localist representation was used.

Figure 3 shows the network's predictions for successive words of the question "*Is the boy who is smoking crazy?*" after the third stage of training. The sequence of predictions is shown from left to right with the target words beneath the predictions, and the average activation value for each category represented vertically. The leftmost columns show that the network predicts an AUX as a possible first word, a name, pronoun, or DET as a continuation when presented with '*is*', and a noun or an adjective as possibilities after '*is the*'. This is not surprising since these sequences all occur in the training sets. Following presentation of '*is the boy*', however, not only is an adjective or a preposition predicted, but also a relativizer — a sequence which never occurs in the training sets. And upon presentation of '*who*' the network predicts an AUX, followed by the prediction of a participle when given '*is*'. The network has thus generalized to predict the

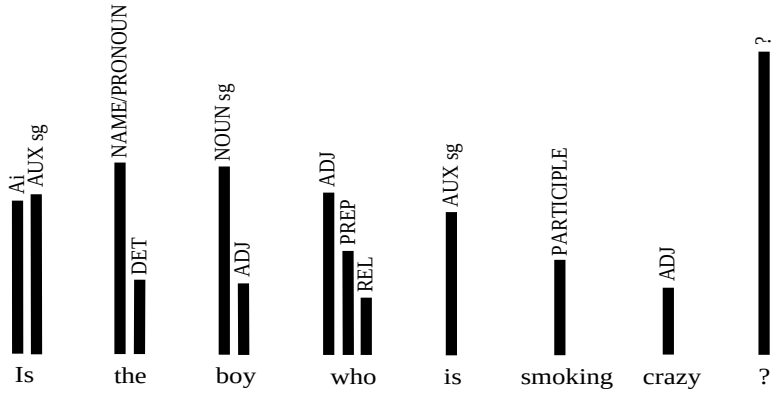

Figure 3: The SRN's categorized predictions for the test sentence "*Is the boy who is smoking crazy?*" Target words appear under the network's predictions; and the strength of the predictions is represented vertically.

[5]The networks were simulated with *LENS* (Rohde, 1999), and trained with a fixed learning rate of 0.01, using a variation of cross entropy which assigned smaller errors for predicting incorrectly than for failure to predict. The architecture shown in Figure 1 is used, with 100 input and output units, 50 units in the reduction layers, and 500 units in both the hidden and context layers.

relative clause.[6] The network does not make the predictions corresponding to the ungrammatical form in (2) — *i.e.*, the network does not predict a participle following '*who*' — and should not be expected to, of course, since the training sets do not contain copula constructions, and so there can be no hypothesis of a movement derivation. Rather, the network has apparently formed an abstract representation of NPs which includes NPs with relative clauses. That this is so is shown by the networks prediction of an adjective when presented with '*is the boy who is smoking*'; the sequence '... PARTICIPLE ADJ ...' never occurs in the training sets, and thus the prediction indicates that the network has formed an abstract representation of *aux*-questions, and generalized over the NP forms.

That the data available to children are sufficient to provide for this generalization, however, remains to be shown.

3. Child-Directed Speech

There are a number of features of child-directed speech that appear to be important for language acquisition, and particularly for the issue at hand. Complexity increases over time — which has been shown to be a determinant of learnability (*e.g.* Elman, 1991, 1993) — and there are also arguably meaningful shifts in the distribution of types, and the limitations on forms.

The increasing complexity of the child's input is especially relevant to the problem here, since it is directly linked to the frequency of occurrence of relative clauses. Complexity in the child's input is introduced in a way akin to the staged presentation of data used to train the network in the experiment described above; Figure 4 charts the occurrences of tagged relative clauses — *i.e.* marked

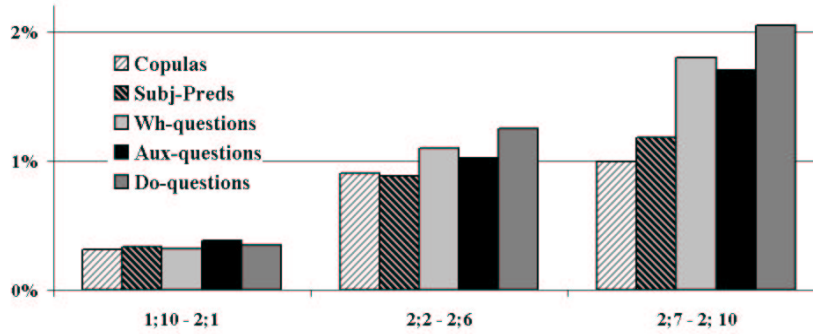

Figure 4: The percentage of the NPs, in each utterance type, that contain relative clauses (averaged over all twelve children, and over each third of the period covered by the corpus).

[6]The fact that the network predicts '*who*' given '*is the boy*' is, on its own, not enough — early in training, the network will make this prediction, but when presented with '*who*' will predict a '*?*', apparently mistaking the relativizer for an adjective. That the network *is* predicting a relative clause is shown by the fact that it predicts '*is*' when subsequently given '*who*', and a participle when then given '*is*'. Since participles are restricted to only occur in relative clauses, the latter is decisive.

with '*who*' or '*that*' — found in child-directed speech in the CHILDES' Manchester corpus (Theakston et al., 2000). Pronominal relatives (*e.g.*, '*the girl you like*') show a similar increase, and occur approximately as frequently. And prepositional phrases increase in frequency slightly more dramatically; they seem to occur approximately twice as often as relatives.[7]

The difference between the distribution of types in child-directed speech and speech between adults is also potentially significant. Child-directed speech contains a much greater proportion of questions — estimated at about one third of the child's input (Hart and Risley, 1995; Cameron-Faulkner et al., 2001) — and thus there is more of a balance between types. This may be critical in establishing the multiple roles that, *e.g.*auxiliaries, can take on; and also to reserve representational space for the the large variety of question forms. Figure 5 shows the changes in the percentages of copula constructions, subject-predicate forms (*e.g.*, transitives and intransitives), and *wh*-, *do*-, and *aux*-questions across the time period covered by the Manchester corpus.

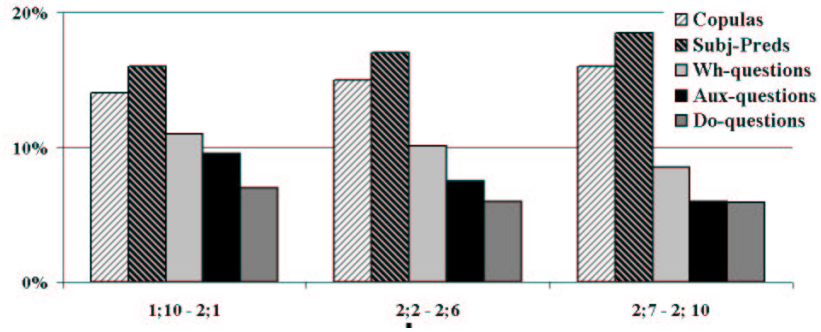

Figure 5: The percentage occurrence, averaged over all children, of various forms for each third of the period covered by the corpus.

And, of potentially even greater significance for the problem at hand, the child's input not only lacks relative clauses in subject position in *aux*-questions, but as Figure 6 shows, seldom contains forms in which the subject NP *could* have a relative clause. The *aux*-questions in child-directed speech overwhelmingly use proper names, pronouns, deictics, and other such forms which do not provide the correct context for a relative clause. This is an aspect of child-directed speech that may both give rise to the absence of relative clauses in *aux*-questions — given the low frequency of relative clauses in general — and also be necessary for the correct generalization to be formed. If this were not the case, and questions like '*Is the boy crazy?*' were common, then the generalization would be threatened: each

[7] A precise count of the prepositional phrases has not been made — in part because of the lesser significance to the current research issue, and in part because it is considerably more problematic to determine whether or not a prepositional phrase is within a noun phrase. But, (Cameron-Faulkner et al., 2001) analyzed a sample from this same corpus, and they report that prepositional phrases make up about 10% of all fragments, which may be indicative of their general frequency.

such occurrence would produce a false prediction, and so provide negative evidence against the wanted generalization.

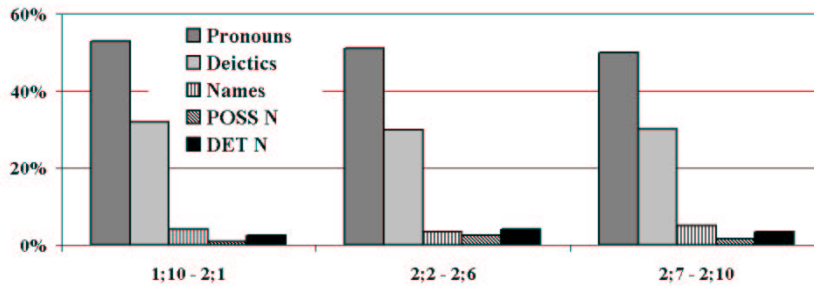

Figure 6: The composition of the subject NPs in *aux*-questions, *i.e.*, the percentage, averaged over all twelve children, of subject NPs that are of each type.

4. Motherese and the Generalization

To determine if an SRN would generalize to predict (1), but not (2), from input of the sort provided to children, the above analysis was used as the basis of a new set of training data, and the simulations repeated. As before, the training sets contained *aux*-questions of the form 'AUX NP ADJ?'; but here the 'A_i NP B_i' forms were eliminated, and copula constructions, subject-predicate forms, and *wh*- and *do*-questions were added. The prohibition on NPs with relative clauses in *aux*-questions extended also to *wh*- and *do*-questions — *i.e.* NPs with relative clauses could occur in object position in these forms, but not in subject position. Thus these training sets also contained no evidence of the sort assumed to distinguish the structure-dependent hypothesis. Some examples from these training sets are given in Figure 7. The proportions of these general types, and the frequency of relative clauses and prepositional phrases, were manipulated in each portion of the training set to match with successive portions of the Manch-

Mummy is beautiful.	*is Mummy beautiful?*
the little boy bites.	*is the little boy nice?*
the cat on the couch scratches.	*is the cat on the couch mangy?*
the boy who is smiling smokes.	* *is the boy who is smiling mean?*
...	...
does Mary smoke?	*is Mary pretty?*
does the dog bite?	*is the dog mangy?*
does the cat on the mat scratch?	*is the cat on the mat fat?*
does he like the girl who is dancing?	* *is the girl who is dancing clumsy?*
...	...
where did the man go?	*is the man scary?*
why did the little girl go home?	*is the little girl happy?*
what about the piece with the dot?	*is the piece with the dot broken?*
who is the boy who is smoking?	* *is the boy who is smoking silly?*

Figure 7: Utterances generated by the artificial grammar.

ester data — *e.g.*, the type distributions can be read directly from figure 5. And, as per the observation of the previous section, noun phrases in *aux*-questions were restricted to be, almost exclusively, pronouns, deitics, and names. The three training sets again consisted of 50,000 examples each; and again the network was trained for 10 epochs on each set, and was tested with the structures in (1) and (2) after each epoch.

Figures 8 and 9 chart the sum-squared error for (1) and (2) after each stage of training. As the figures show, the network succeeds in generalizing to predict (1),

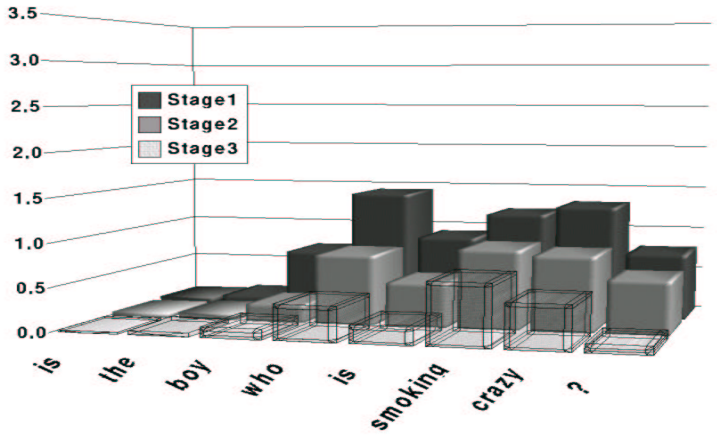

Figure 8: The sum-squared error after each word of the test sentence "*Is the boy who is smoking crazy?*" at the end of each stage of training.

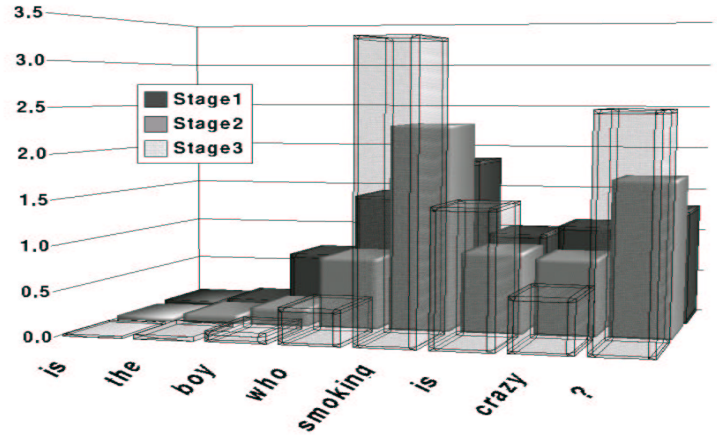

Figure 9: The sum-squared error after each word of the test sentence "*Is the boy who smoking is crazy?*" at the end of each stage of training.

and generates significant error — and progressively larger error — at several points, when presented with (2).[8] The reasonably small error generated by the network when presented with '*who*' in the context of '*is the boy* _' shows that the relativizer is predicted. And the contrast in the errors generated by the subsequent presentation of either '*is*' or '*smoking*' shows clearly that the network has learned to predict an AUX after a relativizer, rather than entertaining the possibility of it's extraction, as in (2). Note, as well, that this contrast is monotonically increasing — at no point in training does the network predict a participle to follow the relativizer. And, for (1), the network's error is quite low for each successive word, including the presentation of the adjective after the participle, despite that '... PARTICIPLE ADJ ...' never occurs in the training sets. In contrast, for (2), as well as the error produced by the presentation of '*smoking*', the network also generates a substantial error upon the subsequent presentation of '*is*'; And though when presented with '*is the boy who smoking is* ' the network successfully predicts an adjective, the success is illusory: when subsequently presented with '*crazy*' the network's predictions are somewhat random, but a period is predicted more strongly than a question mark.

The network does, however, have some difficulties with this input. Although the grammar restricts relative clauses to the form 'REL AUX VERB*ing*', the network persists in predicting noun phrases and adjectives after the auxiliary — presumably because the '*is*' that occurs in initial position in *aux*-questions, followed by a noun phrase, and the '*is*' in declaratives, followed by an adjective, are relatively more frequent in the data than the '*is*' in relative clauses. These erroneous predictions, however, gradually erode. And it is worth noting that they would be correct for a more realistic grammar.

The error associated with the adjective following the participle most likely has a similar source. Relative clauses occur only in either sentence final position, or preceding an auxiliary or a verb; thus the network initially expects participles to be followed by either a verb, a period, a question mark, or most prominently, an auxiliary. Again the problem is somewhat persistent, but is gradually resolved; by the end of the third stage such predictions, though remaining, are substantially weaker than the correct predictions — thus, arguably, not truly problematic. And it is plausible that such errors would not arise were the grammar to be made yet more realistic. The grammar used here contained little variation in terms of either NP types, syntactic structures, or lexical items, and thus generalizations were based on a quite limited set of distributional cues. Lifting the artificial limitations on the grammar might also help to eliminate such errors: questions like '*what's the lady who was at the house called?*' — in Manchester's *ruth28a.cha* — are not only evidence of the sort assumed not to be available, but also data which discourage these sorts of false predictions.

But, such errors are also potentially meaningful. The most prominent and persistent of the errors is the prediction of an auxiliary following the participle, *i.e.*, '*is the boy who is smoking* <u>*is*</u> ...'; in fact an auxiliary is predicted as a possible

[8]The SRN responsible for these results incorporates a variant of the developmental mechanism from (Elman, 1993). That version reset the context layer at increasing intervals; the version used here is similar, but does not reset the context units unless the network's prediction error is greater than a set threshold value.

continuation after any NP, *e.g.*,'*is the boy is* . . . '. And this is an error that children make as well (Crain and Thornton, 1998).

5. Discussion

Assumptions as to the nature of the input, and the ability of the learner to utilize the information therein, clearly play a critical role in determining which properties of language to attribute to UG. These assumptions must be accurate if UG is to be attributed with all and only those properties of language "that can reasonably be supposed not to have been learned" (Chomsky, 1975). An overestimate of either the learner or the input will attribute too little to UG; an underestimate will attribute properties to UG which are, in fact, learned.

The objective here was to demonstrate the necessity of taking into account — amidst a growing body of evidence that children use it — the stochastic information in child-directed speech. To be convincing we have taken on Chomsky's celebrated argument that structure-dependence must be a principle of UG; have been careful to avoid providing the network with input that could be controversial with respect to its availability; and have represented the input in a way that encodes no grammatical information beyond what can be determined from its statistical regularities. This thus substantially under-represents the information actually available to children (since contextual cues, phonological similarity, and other sources of information are abstracted away), and so the fact that a neural network generalizes to make the correct predictions, from data modeled in this way, shows that learnability claims based on a non-stochastic model of the input must be reassessed.

The statistical structure of language provides for far more sophisticated inferences than those which can be made within a theory that considers only whether or not a particular form appears in the input. But determining the potential worth of the stochastic information is difficult. This work shows that neural networks provide a means of dealing with this problem. As demonstrated here, neural networks can be used to assess just how impoverished the stimulus really is, and so can be invaluable to linguists in establishing whether or not a property of language can reasonably be assumed not to have been learned.

References

Angluin, D. (1988). Identifying languages from stochastic examples. Technical Report YALEU/DCS/RR-614, Yale University, New Haven, CT.

Aslin, R., Saffran, J., and Newport, E. (1998). Computation of conditional probability statistics by 8-month old infants. *Psychological Science*, 9:321–324.

Cameron-Faulkner, T., Lieven, E., and Tomasello, M. (2001). A construction based analysis of child directed speech. forthcoming.

Chomsky, N. (1975). *Reflections on Language*. Pantheon Books, New York.

Chomsky, N. (1981). *Lectures on Government and Binding*. Foris Publishers, Dordrecht, Holland.

Christiansen, M. and Chater, N. (1999). Toward a connectionist model of recursion in human linguistic performance. *Cognitive Science*, 23(2):157–205.

Cowie, F. (1998). *What's Within? Nativism Reconsidered*. Oxford University Press.

Crain, S. (1991). Language acquisition in the absence of experience. *Behavioral and Brain Sciences*, 14:597–650.

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

Crain, S. and Thornton, R. (1998). *Investigations in Universal Grammar: A Guide to Experiment's on the acquisition of Syntax and Semantics*. MIT Press.

Elman, J. (1990). Finding structure in time. *Cognitive Science*, 14:179–211.

Elman, J. (1991). Distributed representations, simple recurrent networks, and grammatical structure. *Machine Learning*, 7:195–225.

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

Elman, J. (1998). Generalization, simple recurrent networks, and the emergence of structure. In Gernsbacher, M. and Derry, S., editors, *Proceedings of the 20th Annual Conference of the Cognitive Science Society*, Mahway, NJ. Lawrence Erlbaum Associates.

Gold, E. (1967). Language identification in the limit. *Information and Control*, 10:447–474.

Gomez, R. and Gerken, L. (1999). Artificial grammar learning by one-year-olds leads to specific and abstract knowledge. *Cognition*, 70:109–135.

Hart, B. and Risley, T. (1995). *Meaningful Differences in the Everyday Experiences of Young Children*. Paul H. Brookes, Baltimore, MD.

Horning, J. (1969). *A study of grammatical inference*. PhD thesis, Stanford.

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk. Third Edition*. Lawrence Erlbaum Associates, Mahwah, NJ.

Morris, W., Cottrell, G., and Elman, J. (2000). A connectionist simulation of the empirical acquisition of grammatical relations. In Wermter, S. and Sun, R., editors, *Hybrid Neural Systems*. Springer Verlag, Heidelberg.

Newport, E. and Aslin, R. (2000). Innately constrained learning: Blending old and new approaches to language acquisition. In Howell, S., Fish, S., and Keith-Lucas, T., editors, *Proceedings of the 24th Annual Boston University Conference on Language Development*, Somerville, MA. Cascadilla Press.

Piatelli-Palmarini, M. (1980). *Language and Learning: The debate between Jean Piaget and Noam Chomsky*. Harvard University Press, Cambridge, MA.

Pullum, G. and Scholz, B. (2001). Empirical assessment of stimulus poverty arguments. *The Linguistic Review*. to appear.

Rohde, D. (1999). Lens: The light, efficient network simulator. Technical Report CMU-CS-99-164, Carnegie Mellon University, Department of Computer Science, Pittsburgh, PA.

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

Sampson, G. (1989). Language acquisition: Growth or learning? *Philosophical Papers*, 18:203–240.

Theakston, A., Lieven, E., Pine, J., and Rowland, C. (2000). The role of performance limitations in the acquisition of 'mixed' verb-argument structure at stage 1. In Perkins, M. and Howard, S., editors, *New Directions in Language Development and Disorders*. Plenum.