A Connectionist Investigation of Linguistic Arguments from the Poverty of the Stimulus: Learning the Unlearnable

John D. Lewis (jlewis@crl.ucsd.edu)
Department of Linguistics, McGill University; 1085 Dr. Penfield Avenue
Montreal, Quebec H3A1A7 Canada
&
Center for Research in Language, UC San Diego; 9500 Gilman Dr.
La Jolla, CA 92093-0526 USA

Jeffrey L. Elman (elman@crl.ucsd.edu)
Center for Research in Language, UC San Diego; 9500 Gilman Dr.
La Jolla, CA 92093-0526 USA

Abstract

Based on the apparent paucity of input, and the non-obvious nature of linguistic generalizations, Chomskyan linguists assume an innate body of linguistically detailed knowledge, known as Universal Grammar (UG), and attribute to it principles required to account for those "*properties of language that can reasonably be supposed not to have been learned*" (Chomsky, 1975). A definitive account of learnability is lacking, but is implicit in examples of the application of the logic. Our research demonstrates, however, that important statistical properties of the input have been overlooked, resulting in UG being credited for properties which are demonstrably learnable; in contradiction to Chomsky's celebrated argument for the innateness of structure-dependence (*e.g.* Chomsky, 1975), a simple recurrent network (Elman, 1990), given input modelled on child-directed speech, is shown to learn the structure of relative clauses, and to generalize that structure to subject position in *aux*-questions. The result demonstrates that before a property of language can *reasonably* be supposed not to have been learned, it is necessary to give greater consideration to the indirect positive evidence in the data — and that connectionism can be invaluable to linguists in that respect.

Introduction

Chomskyan linguists argue that language acquisition cannot strictly be a matter of learning — the child's target grammar is "*hopelessly underdetermined by the fragmentary evidence available*" (Chomsky, 1968) — rather it must rest on a set of innate linguistic principles; the goal of the Chomskyan linguist is to determine the contents of this set, known as Universal Grammar (UG). The idea is to attribute to UG all and only the principles required to account for those "*properties of language that can reasonably be supposed not to have been learned*" (Chomsky, 1975). Learning theory is thus of central importance to the enterprise, but, oddly, a definitive account of the notion of learning that Chomskyans adopt is lacking, and is given only implicitly in the examples of the principles attributed to UG. Statistical approaches, however, and the notions of generalization and analogy have been explicitly rejected as irrelevant (Chomsky, 1975). In this paper we demonstrate that this rejection is a serious error — that UG has been attributed with principles to account for properties of language that are demonstrably learnable from the statistical properties of the input.

Chomsky's celebrated argument for the innateness of the principle of structure-dependence (Chomsky, 1975) serves as an example. Chomsky claims that, during the course of language acquisition, children entertain only hypotheses which respect the abstract structural organization of language, though the data may also be consistent with structure-independent hypotheses, *i.e.* relationships over utterances considered only as linearly ordered word sequences. As support for this claim, Chomsky notes that though questions like (1) are apparently absent in the child's input, questions like (2) are never erroneously produced — a claim subsequently

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

empirically tested and substantiated by Crain and Nakayama (1987, also see Crain 1991). Chomsky suggests that it is reasonable to suppose that children derive *aux*-questions from declaratives, and exposed to only simpler structures, might hypothesize either of two sorts of rules: a structure-independent rule — *i.e.* move the first '*is*' — or the correct structure-dependent rule. Chomsky claims that "*cases that distinguish the hypotheses rarely arise; you can easily live your whole life without ever producing a relevant example to show that you are using one hypothesis rather than the other one*" (Piatelli-Palmarini, 1980). The fact that children do not produce questions like (2), despite that the correct rule is supposedly more complex, and that the learner might not encounter the relevant evidence leads Chomsky to suggest that "*the only reasonable conclusion is that* UG *contains the principle that all such rules must be structure-dependent*" (Chomsky, 1975).

As a number of researchers have noted, however, there are several weaknesses in this argument. Slobin (1991), for instance, points out that the conclusion rests on the assumption that *aux*-questions are derived from declaratives by movement — an assumption which lacks justification — as well as on the equally questionable assumption of the autonomy of syntax. The argument has also been widely criticized for its reliance on the extremely limited conception of learning as hypotheses generation and testing. And the premise that the relevant evidence is not available to children has repeatedly been argued to most likely be false. As Sampson (1989) points out, evidence to distinguish the two hypotheses is provided by any utterance in which *any* auxiliary precedes the main clause auxiliary; thus evidence is available not only in questions like "*Is the jug of milk that's in the fridge empty?*" (from Cowie, 1998), but also "*Is the ball you were speaking of in the box with the bowling pin?*", or "*Where's this little boy who's full of smiles?*", or even "*While you're sleeping, shall I make the breakfast?*" None of these forms seem to be of the sort that a person might go for long without encountering; the latter three examples, in fact, are taken from the CHILDES database,[1] and Pullum and Scholz (2001) estimate that such examples make up about one percent of a typical corpus.

These are strong criticisms, but a conclusive counter-argument, or an alternate account of the acquisition of *aux*-questions remains to be given. This paper builds on recent work with simple recurrent networks (SRNs; Elman 1990) to close this gap — *i.e.* to provide a proof that the correct form of *aux*-questions is learnable from data uncontroversially available to children.

Figure 1 shows the general structure of an SRN. The recurrent connections from the hidden layer to the context layer provide a one-step state memory. At each time step the activation values 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 at each step attempts to predict the next input. 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.

Elman (1991, 1993) provided such a network 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*.[2,3]

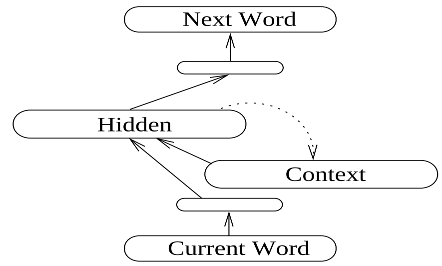

Figure 1: An SRN. Solid lines represent full connectivity; the dashed line indicates unit-to-unit connections. The unlabeled layers are reduction layers.

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 the case. An initial experiment verifies that the two results combine in the required way; then an SRN is shown to generalize from training sets based on CHILDES data to predict (1), but not (2). This result clearly runs counter to Chomsky's argument, and thus both draws into question the validity of poverty of the stimulus arguments in general, and shows that neural networks provide a means of assessing just how impoverished the stimulus really is.

Abstractions and Generalization

Training sets similar to those used by Elman (1991, 1993) were used to test whether an SRN would generalize to predict relative clauses in subject position in *aux*-questions from data which contained no such questions. An artificial grammar was created such that it generated *a)* *aux*-questions of the form 'AUX NP ADJ?',

[1] The second through fourth examples are from Brown's Adam, Korman's' St, and Manchester's Anne, respectively.

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

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

and *b)* sequences of the form '$\mathrm{A_i}$ NP $\mathrm{B_i}$', where $\mathrm{A_i}$ and $\mathrm{B_i}$ were of varying content and length. Proper names and NPs of the form 'DET (ADJ) N (PP)' were generated in both types, and NPs with relative clauses were generated for the '$\mathrm{A_i}$ NP $\mathrm{B_i}$' type, but were restricted from appearing in *aux*-questions. Some representative examples are given in Figure 2.

$\mathrm{A_i}$ *Mummy* $\mathrm{B_i}$	*is Mummy beautiful?*
$\mathrm{A_i}$ *the dog* $\mathrm{B_i}$	*is the dog hungry?*
$\mathrm{A_i}$ *the little girl* $\mathrm{B_i}$	*is the little girl pretty?*
$\mathrm{A_i}$ *the cat on the mat* $\mathrm{B_i}$	*is the cat on the mat fat?*
$\mathrm{A_i}$ *the boy who is smiling* $\mathrm{B_i}$	*

Figure 2: Examples of the various types of utterances generated by the artificial grammar.

A three-stage training set was generated from this 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 constituted 80% of the NPs in the first set, and the remaining 20% was 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% '$\mathrm{A_i}$ NP $\mathrm{B_i}$' forms. In the second set, names 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, 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.[4] The network received the input in the same form as used by Elman (1991, 1993), *i.e.* a localist representation was used, and the data was presented one word at a time.

Figure 3 shows the networks predictions (after the third stage of training) for successive words of the question "*Is the boy who is smoking crazy?*" As should be expected, the network predicts an AUX as a possible first word, a name or a DET as a continuation when presented with '*is*', and a noun or an adjective as possibilities after '*is the*'. These sequences all occur in the training sets. But, following presentation of '*is the boy*', 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, and when given '*is*', predicts a participle; the network has thus generalized to predict the

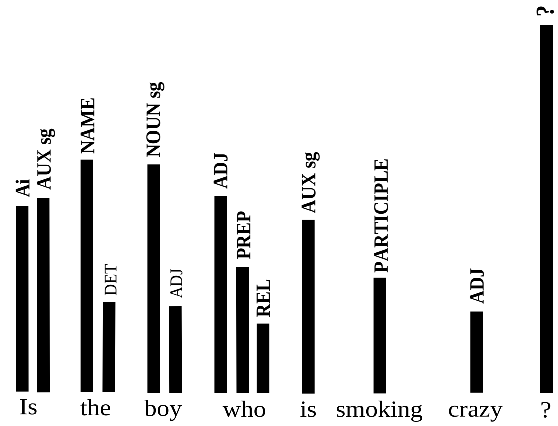

Figure 3: The SRN's categorized predictions for the test sentence "*Is the boy who is smoking crazy?*" Target words appear under the network's predictions; and the strength of the predictions is represented vertically.

relative clause.[5] The network does not, of course, make the predictions corresponding to the ungrammatical form in (2) — *i.e.* the network does not predict a participle following '*who*'; 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.

Child-Directed Speech

There are a number of features of child-directed speech that run counter to the notion that the child's input is "*meager and degenerate*" (Chomsky, 1968) — *i.e.*, 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.

[4]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.

[5]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.

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 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.[6]

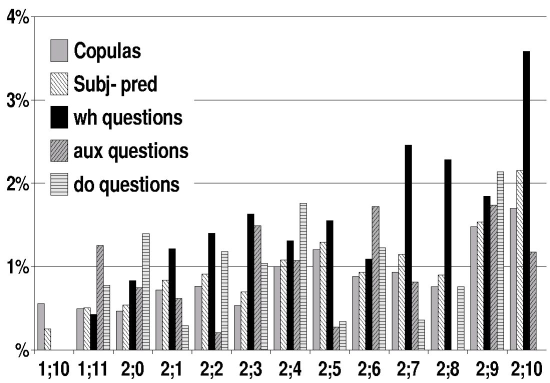

Figure 4: The percentage of NPs that contain relative clauses, for each month, averaged over all twelve children in the Manchester corpus.

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 percentages of copula constructions, subject-predicate forms (*e.g.*, transitives and intransitives), and *wh*-, *do*-, and *aux*-questions for representative months near the beginning, middle, and end of the time period covered by the Manchester corpus.

And finally, *aux*-questions in the child's input not only lack relative clauses in subject position, but are limited in a way that both predicts this absence, and potentially allows for the correct generalization to be formed. In child-directed speech, *aux*-questions with a determiner in the subject noun phrase — like '*Is the boy crazy?*' — are almost never used; the *aux*-questions in child-directed speech overwhelmingly use proper names, pronouns, deictics, *e.g.* '*Is that …*', and other such forms which do not provide the correct context for a relative clause. Thus, given the low frequency of relative clauses in general, one should expect them to almost never occur in subject position.

These are ideal conditions for an SRN. The target generalization is supported by the appearance of relative clauses in all other positions in which noun phrases occur, and making the generalization incurs little cost since the context in which the generalization applies seldom occurs. If this were not the case, and questions like '*Is the boy crazy?*' were common, then the generalization would be threatened — each such occurrence would produce a false prediction which backpropogation would attempt to eliminate.

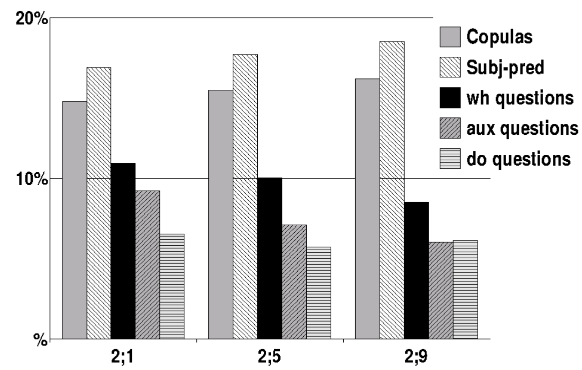

Figure 5: The percentage occurrence of various forms, at three stages, averaged over all children.

Motherese and the Generalization

Training sets generated on the basis of this analysis were used to determine if an SRN would generalize to predict (1), but not (2) from input of this sort. As before, the training sets contained *aux*-questions of the form 'AUX NP ADJ?'; but here the '$\mathrm{A_i}$ NP $\mathrm{B_i}$' 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 6. 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 Manchester 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, names. The three training sets again consisted of 50,000

[6] 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.

Mummy is beautiful.	*is Mummy beautiful?*
the little boy bites.	*is the little boy nice?*
the dog likes Mummy.	*is the dog hungry?*
does Mary smoke?	.
who likes Mary?	.
who does Mary like?	.
who likes the cat on the mat?	
who does the girl at the shop like?	
does the cat on the mat scratch?	
does the little girl like the boy who is smiling?	

Figure 6: Examples of the various types of utterances generated by the artificial grammar.

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 7 and 8 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), and generates significant error — and progressively larger error — at several points, when presented with (2).[7] 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.

[7] 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.

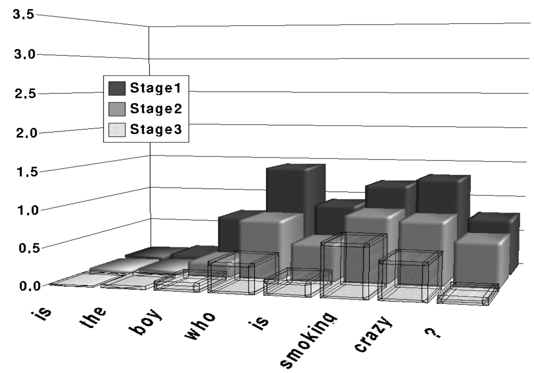

Figure 7: The sum-squared error after each word of the test sentence "*Is the boy who is smoking crazy?*" at the end of each stage of training.

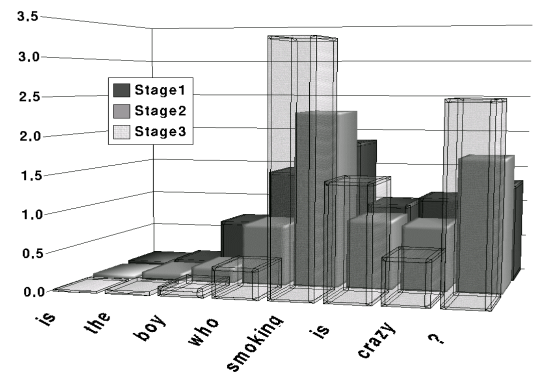

Figure 8: The sum-squared error after each word of the test sentence "*Is the boy who smoking is crazy?*" at the end of each stage of training.

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

Discussion

The objective here was to provide a proof that the structure of aux-questions is learnable from the input available to children. To make the results convincing, we 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.

The fact that a neural network generalizes to make the correct predictions from input represented in this way, and modeled on child-directed speech — but limited to contain no data of what has been considered the relevant sort — shows that poverty of the stimulus arguments must give greater consideration to the indirect evidence available to the child. 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. And there is a growing body of evidence that children, not only neural networks, make use of the statistical properties of the input in acquiring the structure of language (*e.g.* Aslin et al., 1998; Gomez and Gerken, 1999). Thus learnability arguments cannot ignore those properties.

But discovering what those properties are, and determining their potential worth in language acquisition 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

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. (1968). *Language and Mind*. Brace & World, New York.

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

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.

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.

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.

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.

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

Slobin, D. (1991). Can Crain constrain the constraints? *Behavioral and Brain Sciences*, 14:633–634.

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.