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

The principal goal of Chomskyan linguistics is to account for language acquisition; the core postulate is that the child has available linguistically detailed innate knowledge, known as Universal Grammar (UG). Critical to this postulate is the poverty of the stimulus argument — that the child’s target grammar is “hopelessly underdetermined by the fragmentary evidence available” (Chomsky, 1968). UG is assumed to account for those “properties of language that can reasonably be supposed not to have been learned” (Chomsky, 1975). 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 the most celebrated argument from the poverty of the stimulus — Chomsky’s structure-dependence argument (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

Based on the apparent paucity of input, and the non-obvious nature of linguistic generalizations, Chomsky (1968; 1971; 1975, etc.) has argued extensively that language acquisition must rest on linguistically detailed innate knowledge. This innate knowledge, known as Universal Grammar (UG), is taken to explain how it is that “every child comes to know facts about the language for which there is no decisive evidence from the environment. In some cases, [apparently] no evidence at all.” — i.e. cases of “language acquisition in the absence of experience” (Crain, 1991). UG is attributed with the principles required to account for those “properties of language that can reasonably be supposed not to have been learned” (Chomsky, 1975).

The most celebrated example of the application of this logic is Chomsky’s argument for the innateness of the principle of structure-dependence (Chomsky, 1975). 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 empirically tested

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

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); and that “the belief that each child encounters relevant evidence strains credulity” (Chomsky, 1975). The fact that children do not produce questions like (2), despite that the correct rule is supposedly more complex, and that the learner might proceed “through a considerable portion of his life without ever facing relevant evidence” (Chomsky, 1975) 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 has been argued by a number of researchers (e.g. Sampson, 1989; Freidin, 1991; Pullum, 1996; Pullum and Scholz, 2001), the premiss that the relevant evidence is not available to children is most likely false. Questions like “Is the jug of milk that’s in the fridge empty?” do not seem to be of the sort that a person might go through
their entire lives without encountering. And as Sampson (1989) points out, evidence is provided by any utterance in which an auxiliary precedes the main clause auxiliary, and that question types involving inversion, e.g. do-, care, and wh-questions, as well as a variety of other utterance types, are often of this form; Pullum and Scholz (2001) estimate that such cases constitute about one percent of the child’s input. It remains to be shown, however, that such data are sufficient to account for children’s success with the problem in question.

Recent work with neural networks suggests the possibility of a conclusive argument — a proof that the correct form of aux-questions is learnable. Elman (1991, 1993) demonstrated that a simple recurrent network (SRN; Elman 1990) can learn the structure of an artificial language with embedded relative clauses, and Elman (1998) showed that an SRN will induce subject and object categories, and generalize between them. 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 grammar in question despite not having access to the sort of evidence that has been assumed necessary.

This paper reports on the investigation of this possibility. An analysis of child-directed speech from the CHILDES Manchester Corpus (MacWhinney, 2000; Theakston et al., 2000) is used in order to create realistic training data; and an SRN is shown to generalize from this 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.

**Learning AUX-questions with an SRN**

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 multiple embedded relative clauses (in which the noun could be either the subject or object of the subordinate clause). The network’s task was to learn 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 (possibly with multiple levels of embedding), e.g. boys who like the girl who Mary hates hate Mary. The sentences were input as a sequence of words, where each word was represented as a string of 0s with a single bit set to 1 — a localist representation. Thus no information about either the words or the grammatical structure was supplied; the network had to extract all information from the grammatical regularities that underlay the input (e.g. the grammatical categories, number agreement, subcategorization frames, and selectional restrictions). An SRN was shown to succeed at the task if either a) the input to the net initially contains primarily simple structures, and complexity is introduced gradually; or b) the net undergoes development, beginning with a limited internal memory which gradually expands over time.1

The current research builds on this result by showing a) that it combines with the Elman (1998) result — i.e. that an SRN will generalize, not only words, but also full phrases, between subject and object positions — to yield an SRN that predicts the grammatical aux-question in (1), but not the ungrammatical (2); and b) that (a) holds for realistic training data.

**An SRN Generalizes to Predict (1), but not (2).**

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) simple declaratives and aux-questions, b) declaratives with relative clauses or prepositional phrases in

---

1 Though not relevant to the issue here, the key point of that research was that, unless the network undergoes development, it fails on the task when the corpus is presented in its entirety from the outset, and thus in agreement with Newport’s (1990) less is more hypothesis, it seems that learning is enhanced by the initial limitations on the network.
either subject or object position, and c) aux-questions with prepositional phrases in either position, but relative clauses in object position only. Examples of each sort are given in Figure 2. A three-stage training set was then generated: the first with 10% of either utterance type containing the complex forms, the second with 25% complex forms, and the third 50%; each training set consisted of 50,000 examples, with approximately 50% declaratives and 50% aux-questions. An SRN was then trained on each set successively, for 10 epochs each, and tested with the structures in (1) and (2) after each epoch. 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.

- The dog bites.
- the man with Daddy is nasty.
- the girl walks the dogs.
- the boy with the cat likes Mommy.
- the woman who is laughing is happy.
- Mommy likes the boy with the dog.
- the boy likes the girl who is laughing.
- is the man crazy?
- is the girl with Mommy pretty?
- is Daddy the man with Mommy?
- is Mommy the woman who is smiling?

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

The SRN succeeded at the task — i.e. it generalized to recognize questions like (1), but to exclude ungrammatical forms like (2). Figure 3 shows the networks predictions after training. Following presentation of ‘is the boy’, a relativizer is predicted as a possibility; and an auxiliary or verb is predicted after the subsequent presentation of ‘who’. The false generalization that Chomsky suggests should occur, and lead to the ungrammatical question in (2), does not occur; relative clauses containing an auxiliary are always of the form ‘REL AUX VERBING’, but the network does not predict a gerund following ‘who’. And though substantial training is required before a relativizer is predicted when presented with ‘is the boy _____’, the network quickly learns to predict an auxiliary or a verb when presented with ‘is the boy who _____’. And at no point during training is a gerund predicted to follow the relativizer.

Two portions of the task, however, are problematic for the SRN: a) the prediction of the gerund, and b) the prediction of the word to follow the gerund. The grammar restricted gerunds to relative clauses, and since 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, the network has a tendency to predict noun phrases and adjectives following the ‘is’ in relative clauses. This is a somewhat persistent problem, but these erroneous predictions gradually erode. And it is worth noting that these predictions would be correct for a more realistic grammar. Problem (b) has a similar source. Since relative clauses occur only in declaratives, either in sentence final position, or preceding a verb, and in final position in aux-questions, the network initially expects gerunds to be followed by either a verb, a period, or a question mark. Again the problem is somewhat persistent, but is gradually resolved.

The training data, however, are not terribly realistic — particularly for child-directed speech — and thus the fact that an SRN forms the correct generalization over these data shows that the Elman (1998) result applies at the appropriate level, but not that the data available to children are sufficient to account for the acquisition of the grammar with respect to aux-questions.

**Child-Directed Speech**

Researchers have reported a number of features of child-directed speech that, in opposition to the claim that the child’s input is “meager and degenerate” (Chomsky, 1968), appear to be important for language acquisition. Disfluencies in child directed speech are extremely rare (Newport et al., 1977), and it is, in fact, sufficiently structured to have prompted the claim that mothers are essentially providing principled language lessons (e.g."
Levitt, 1975; Snow and Ferguson, 1977, but see Newport et al. 1977). Complexity increases over time — which is a determinant of learnability — 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 particularly relevant to the problem at hand 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; an analysis of the child-directed speech in CHILDES’ Manchester corpus, however, reveals that complex forms in the child’s data occur far less frequently than in even the first training set, and the increase in the complexity of the child’s data is far more gradual. Figure 4 charts the occurrences of tagged relative clauses — *i.e.* marked with ‘who’ or ‘that’ — over the period of time covered by the corpus. The increase is extremely gradual, and even by the end of the period considered, tagged relative clauses constitute less than 3% of the data. Pronominal relatives — *e.g.* ‘the girl you like’ — show a similar increase, and occur approximately as frequently. Prepositional phrases also increase in frequency, but are substantially more common; they seem to be approximately twice as frequent as relatives, though an exact count remains to be done.4

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 — *i.e.* ‘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 backpropagation would attempt to eliminate.

**An SRN Learns from Child-Directed Speech.**

An SRN was presented with new training sets generated on the basis of this analysis. Noun phrases in subject position in aux-questions were restricted as per the above observation; and three new training sets were generated — with the proportions of declaratives and aux-questions, and the frequency of relative clauses and prepositional phrases, as per successive portions of the Manchester data. The new training sets thus provided far fewer questions, proportionally, than did the first sets, and were a massive reduction with respect to relative
clauses. The SRN that had previously succeeded now failed — also with larger hidden and context layers, and even when the network started small as per Elman (1991, 1993)\(^5\). Substantial training was required for the network to make correct predictions on the latter part of even simple questions.

To provide a better balance between declaratives and question forms, and to encourage generalization by increasing the number of contexts in which noun phrases with relative clauses occur, the original grammar was augmented with \textit{wh} and \textit{do}-questions (and the training sets were regenerated with the distribution of types, and the frequencies of relative clauses for each type, as per the Manchester analysis). SRNs performed significantly better on these training sets — particularly when the networks started small — and by modifying Elman’s original developmental mechanism to take into account the network’s level of confusion, a variant was produced that succeeded in forming the correct generalization.\(^5\) The sum-squared errors for (1) and (2) are charted in figures 6 and 7. As the figures show, the network generalizes to predict the possibility of a relative clause — \textit{i.e.} it predicts ‘who’ given ‘is the boy’ \(^7\) — and that a relative clause must contain an auxiliary or a verb. Presenting the network with ‘is the boy who smoking’ generates a substantially larger error than presenting it with ‘is the boy who is’, as can be seen by comparing figures 6 and 7. And the network’s subsequent predictions, though better, still reflect confusion. 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 among them is a period, and not a question mark. As the figures show, the network generalizes to recognize questions like (1), and at no point in development does it predict the absence of the auxiliary, \textit{i.e.} the ungrammatical (2).

But, as before, the network has difficulties with a) the gerund, and b) the word following the gerund. Despite that the grammar restricts relative clauses with auxiliaries to the form ‘REL AUX VERBING’, presumably for the reasons discussed above, the network persists in predicting noun phrases and adjectives after the auxiliary — though this slowly erodes with extended training. Problem (b) is exacerbated here by training-data like ‘who is the boy who is smoking?’, which encourage the false prediction that ‘?’ will follow the gerund; but again, these errors erode with training, and 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; questions like ‘what’s the lady who was at the house called?’ — in Manchester’s \textit{ruth28a.cha} — are not only evidence of the sort assumed not to be available, but also are data which discourage these sorts of false predictions.

\(^5\)Elman (1991, 1993) showed that an SRN could learn an artificial language with relative clauses from unstructured data only if the network started with a limited memory which increased over time, \textit{e.g.} by resetting its context units at increasing intervals.

\(^6\)The successful variant, rather than resetting the context layer at set intervals, does so only when the network is both wildly erroneous in its predictions, and the number of inputs received exceeds the lower limit set by the memory-length parameter. This eliminates a tremendous amount of noise, since training data which are relatively long, but nonetheless unproblematic, reinforce correct predictions, whereas, on the original scheme, the loss of the inputs from the context layer led to errors in subsequent predictions, and thus to erroneous weight adjustments.

\(^7\)Actually, 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 ‘?’? That the network is predicting a relative clause is shown by the fact that it predicts ‘is’ when subsequently given ‘who’, and a gerund when then given ‘is’. Since gerunds are restricted to only occur in relative clauses, the latter is decisive.
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 modelled 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


