Incremental learning, or The importance of starting small

Jeffrey L. Elman

CRL Technical Report 9101
March 1991

Center for Research in Language

University of California, San Diego
La Jolla, CA 92093-0126

Incremental learning, or The importance of starting small

Jeffrey L. Elman

CRL Technical Report 9101

March 1991

University of California, San Diego

La Jolla, CA 92093-0126

Incremental learning, or The importance of starting small

Jeffrey L. Elman
Departments of Cognitive Science and Linguistics
University of California, San Diego

ABSTRACT

Most work in learnability theory assumes that both the environment (the data to be learned) and the learning mechanism are static. In the case of children, however, this is an unrealistic assumption. First-language learning occurs, for example, at precisely that point in time when children undergo significant developmental changes. In this paper I describe the results of simulations in which network models are unable to learn a complex grammar when both the network and the input remain unchanging. However, when *either* the input is presented incrementally, *or*—more realistically—the network begins with limited memory that gradually increases, the network is able to learn the grammar. Seen in this light, the early limitations in a learner may play both a positive and critical role, and make it possible to master a body of knowledge which could not be learned in the mature system.

INTRODUCTION

One of the things which makes language learning such an interesting phenomenon is what has been called the 'projection problem'. The idea is just that, if the problem of the language learner is to figure out the underlying regularities—that is, the grammar—which are responsible for the language he or she hears, then the data which are available to the learner may not be sufficient to uniquely determine the correct grammar.[1]

This problem of the apparent insufficiency of the data has been discussed in many contexts (e.g., Baker, 1979; Bowerman, 1987; Pinker, 1989; Wexler & Cullicover; 1980) but one of the simplest demonstrations comes from Gold's (1967) work. Gold shows that if a language learner is presented with positive-only data (what he calls 'text presentation'), only regular languages can be learned. Regular languages are languages which can be generated by finite state automata. The rub is that, on the one hand, natural languages appear to belong to a more powerful class than this (Chomsky, 1957); and on the other, there is no good evidence that children receive or use negative data during learning (Brown & Hanlon, 1970; Hirsh-Pasek, Treiman, and Schneiderman, 1984; Braine, 1971).

Gold advances several suggestions in order to account for the fact that, despite his findings, children *do* learn language. Although children do not appear to receive explicit negative evidence, they may receive indirect negative evidence. Or possibly, some of what children know is innate; thus they need not infer the grammar solely on the basis of positive data.[2]

Almost certainly both of the possibilities outlined by Gold are true to some extent. That is, the child is not an unconstrained learning mechanism in the sense of being able to learn any and all possible languages. Rather, innate predispositions narrow the range of what can be learned. Of course, it is very much an open (and controversial) question exactly what form that innate knowledge takes. A number of investigators have also proposed that although *direct* negative evidence may not

1. I say 'may' because much hinges on exactly what one believes the nature of the input to be: bare sentence strings? strings accompanied by semantic interpretations? strings accompanied by information about the environment in which they were uttered?

2. Gold mentions a third possibility, which is that if the text is ordered, then positive-only presentation is sufficient to learn even the most complex set of languages he considers. The details of this proposal are not well-developed however.

be available, there are subtler forms of negative evidence. For example, the non-occurrence of an expected form constitutes an indirect sort of negative evidence. Just how far this sort of evidence can be used has been challenged (Pinker, 1989). Thus, although innateness and indirect evidence plausibly participate in the solution of the learnability problem, their contribution is not known and remains controversial.

In this paper, I want to pursue what may be a third factor in helping account for the apparent ability of learners to 'go beyond the data'. This factor hinges on the simple fact that first language learners (children) are themselves undergoing significant developmental changes during precisely the time that they learn language. Indeed, language learning after these developmental changes have completed seems to be far less successful. This is often attributed to the passing of a 'critical period' for language learning. But this is no more than a restatement of facts. What I would like to consider here is the question of what it is about the so-called critical period that might facilitate learning language.

Interestingly, almost all learnability work ignores this basic fact. It is typically assumed that both the learning device and training input are static. One might wonder what the consequences are of having either the learning device (network or child) or the input data not be constant during learning? Plunkett and Marchman (1990) have shown that while the basic influences of type/token frequency and phonological predictability are similar to the condition of non-incremental learning, better overall learning is achieved when the training corpus for a connectionist model is allowed to slowly grow in size. We might also ask what the consequences are when the learning mechanism itself changing.

In this paper, I will report the effect of staged input on learning in a connectionist model. The network fails to learn the task when the entire data set is presented all at once, but succeeds when the data are presented incrementally. I then show how similar effects can be obtained by the more realistic assumption that the the input is held constant, but the learning mechanism itself undergoes developmental changes. Finally, I examine the network to see what the mechanism is which allows this to happen and suggest what conditions are necessary for incremental learning to be useful.

Simulations

This work was originally motivated by an interest in studying ways in which connectionist networks might use distributed representations to encode complex, hierarchically organized information. By this I mean just the sort of relationships which typically occur in language. For example, in the sentence *The girls who the teacher has picked for the play which will be produced next month practice every afternoon*, there are several events which are described. Some are backgrounded or subordinate to the main event. This has grammatical consequences. Thus, the main verb (*practice*) is in the plural because it agrees with *the girls* (not *the teacher*, nor *the play*). And although *picked* is a transitive verb which often takes a direct object following it, no noun appears after the verb because the direct object (*the girls*) has already been mentioned.

These sorts of facts (specifically, the recursive nature of embedded relative clauses) led many linguists to conclude that natural language could not be modeled by a finite state grammar (Chomsky, 1957), and that statistical inference as a learning mechanism for language was untenable (Miller & Chomsky, 1963). These conclusions about the representational and learnability requirements of natural language seem to pose real problems for connectionist networks, which typically rely heavily (though not necessarily exclusively) on statistical inference, and which are more similar to finite state machines than other sorts of computational devices (e.g., pushdown automata; but see Pollack, 1990).

My first approach was to try construct a semi-realistic artificial language which had some of the crucial properties that were cited by Chomsky and his colleagues as being problematic for FSA's and statistical learning, and to train a neural network to process sentences from this language. The network was a simple recurrent network (Elman, 1990, in press; Jordan, 1986; Servan-Schreiber, Cleeremans, & McClelland, 1988). The salient property of this architecture is that it is a kind of dynamical system which allows inputs to be processed in sequence, and in which the internal states are fed back at every time step to provide an additional input. The network must learn to develop internal states (i.e., the hidden unit activation patterns) which encode temporal information in ways which enable the network to produce the correct outputs. The network architecture that was used is shown in Figure 1.

The input corpus consisted of sentences which were generated by a grammar that had certain critical properties: (a) there was number agreement between subject nouns and their verbs; (b) verbs differes with regard to argument expectations (some required direct objects, others optionally permitted objects; others precluded direct objects); (c) sentences could contain multiple embeddings in the form of relative clauses (in which the head could be either the subject or object of the subordinate cluase). The existence of these relative clauses considerably complicated the set of agreement and verb argument facts. (See Elman, in press, for details of this language.)

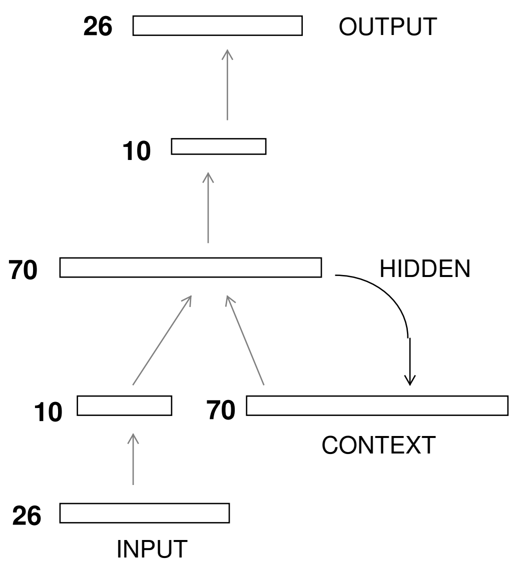

Figure 1

The results of the first trials were quite disappointing. The network failed to master the task, even for the training data. Performance was not uniformly bad. Indeed, in some sentences, the network would correctly coordinate the number of the main clause subject, mentioned early in a sentence, with the number of the main clause verb, mentioned after many embedded relative clauses. But it would then fail to get the agreement correct on some of the relative clause subjects and verbs, even when these were close together. (For example, it might produce *The boys who the girl chase see the dog*, getting the number agreement of *boys* and *see* right, but failing on the more proximal—and presumably, easier—*girl chase*.) But even this pattern was idiosyncratic.

This result, of course, is exactly what might have been predicted by Chomsky, Miller, and Gold.

Incremental input

In an attempt to understand where the breakdown was occurring, and just how complex a language the network might be able to learn, I devised a regimen in which the training input was organized into corpora of increasing complexity, and the network was trained first with the simplest input. There were five phases in all. In the first phase, 10,000 sentences consisting solely of simple sentences were presented. The network was trained on five exposures ('epochs') to this database. At the conclusion of this phase, the training data were discarded and the network exposed to a new set of sentences. In this second phase, 7,500 of the sentences were simple, and 2,500 complex sentences were also included. As before, the network was trained for 5 epochs, after which performance was also quite high, even on the complex sentences. In phase three, the mixture was 5,000 simple/5,000 complex sentences, for 5 epochs. In phase four, the mixture was 2,500 simple/7,500 complex. And in phase five, the network was trained on 10,000 complex sentences.

Since the prediction task—given this grammar—is non-deterministic, the best measure of performance is not the extent to which the literal prediction is correct (measured thus, 0 error would require the network to memorize the training data) but rather the degree to which the network's predictions approximate the empirical probability distributions. Performance using this metric was high at the conclusion of all phases of training, including the final phase: final performance had an error of 0.177, with network output measured against the empirically derived likelihood estimates. (Alternatively, one can measure the cosine of the angle between these two vectors. Mean cosine at the end of traing was 0.852; perfect performance would have been 1.00.) Furthermore, the network's high performance generalized to a variety of novel sentences which systematically test the capacity to predict grammatically correct forms across a range of different structures.

This result contrasts strikingly with the earlier failure of the network to learn when the full corpus was presented at the outset.[1] Put simply, the network was unable to learn the complex grammar when trained from the outset with the full 'adult' language. However, when the training data were selected such that simple sentences were presented first, the network succeeded in not only mastering these, but then going on to master the complex sentences as well.

In one sense, this is a pleasing result, because the behavior of the network partially resembles that of children. Children do not begin by mastering the adult language in all its complexity. Rather, they begin with the simplest of structures, and build incrementally until they achieve the adult language.

1. Both this result and the earlier failure were replicated several times with different starting conditions, a variety of different architectures, and various settings of the learning parameters (learning rate, momentum, bounds on beginning random weight initialization).

There is an important disanalogy, however. In this simulation, the network was placed in an environment which was carefully constructed so that it only encountered the simple sentences at the beginning. As learning and performance progressed, the environment was gradually enriched by the inclusion of more and more complex sentences. This is not a good model for the situation in which children learn language. Although there is evidence that adults modify their language to some extent when interacting with children, it is not clear that these modifications affect the grammatical structure of their speech. Unlike the network, children hear exemplars of all aspects of the adult language from the beginning.

If it is not true that the child's environment changes radically (as in this first simulation), what is true is that the *child* changes during the period he or she is learning language. A more realistic network model would have a constant learning environment, but some aspect of the network itself would undergo change during learning.

Incremental memory

One developmental change which is plausibly relevant to learning is the gradual increase in memory and attention span which is characteristic of children. In the network, the analog of memory is supplied by the access the network has (via the recurrent connections) to its own prior internal states. The network can be given a more limited memory by depriving it of access, periodically, to this feedback. The network would thus have only a limited temporal window within which patterns could be processed.

A second simulation was therefore carried out with the goal of seeing what the effect would be, not of staging the input, but of beginning with a limited memory and gradually increasing memory span. The rationale was that this scenario more closely resembled the conditions under which children learn language.

In this simulation, the network was trained from the outset with the full adult language (i.e., the target corpus that had previously been shown to be unlearnable when it was presented from the beginning). However, the network itself was modified such that during the first phase, the recurrent feedback was eliminated after every third or fourth word (randomly).[1] In the second phase, the network continued with another set of sentences drawn from the the adult language (the first set was discarded simply so the network would not be able to memorize it); more importantly, the memory window was increased to 4-5 words. In the third phase, the memory window was increased to 5-6 words; in the fourth phase, to 6-7 words; and in the fifth phase, the feedback was not interfered with at all.

Under these conditions, it turned out that the first phase had to be extended to much longer than in the previous simulation in order to achieve a comparable level of performance (12 epochs rather than 5; for purposes of comparison, performance was measured only on the simple sentences even though the network was trained on complex sentences as well). However, once this initially prolonged stage of learning was over, learning proceeded as quickly through the remaining stages (5 epochs per stage). At the end, performance on both the training data, and also on a wide range of novel data, was as good as in the prior simulation. If the learning mechanism itself was allowed to undergo 'maturational changes' (in this case, increasing its memory capacity) during learning, then outcome was just as good as if the environment itself had been staged.

Before discussing some of the implications of this finding, it is important to try to understand exactly what the basic mechanism is which results in the apparently paradoxical finding that learning can be improved under conditions of limited capacity. One would like to know, for example, whether this outcome is always to be expected, or whether this result might be obtained in only special circumstances.

We begin by looking at the way the network eventually solved the problem of representing complex sentences. The network has available to it, in the form of its hidden unit patterns, a high-dimensional space for internal representations. It is well known that in such networks these internal representations can play a key role in the solution to a problem. Among other things, the internal representations permit the network to escape the tyranny of a form-based interpretation of the world. Sometimes the *form* of an input is not a reliable indicator of how it should be treated; put another way, appearances can deceive. In such cases, the network uses its hidden units to construct a *functionally-based* representational scheme. Thus, the similarity structure of the internal representations can be more reliable indicator of 'meaning' than the similarity structure of the bare inputs.

In this simulation, the network utilized the various dimensions of the internal state to represent a number of different factors which were relevant to the task. These include: individual lexical item; grammatical category (noun, verb, relative pronoun, etc.); number (singular vs. plural); grammatical role (subject vs. object); level of embedding (main clause, subordinate, etc.); and verb argument type (transitive, intransitive, optional).

[1] This was done by setting the context units to values of 0.5.

Principle component analysis (Gonzalez & Wintz, 1977) can be used to identify the specific dimensions associated with each factor. The internal representations of specific sentences can then be visualized as movements through this state space (one looks at selected dimensions or planes, chosen to illustrate the factor of interest). Example trajectories for several sentences are shown in Figures 2-4.

One can also visualize the representational space more globally by having the network process a large number of sentences, and recording the positions in state space for each word; and then displaying the overall positions. This is done in Figure 5b. Three dimensions (out of the 70 total) are shown; the *x* and *y* coordinates together encode depth of embedding and the *z* coordinate encodes number. (See Elman, in press, for details.)

At the outset of learning, of course, none of these dimensions have been assigned to these functions. If one passes the same sentences through a network prior to training, the internal representations have no discernible structure. These internal representations are the important outcome of learning; they are also the necessary basis for good performance.

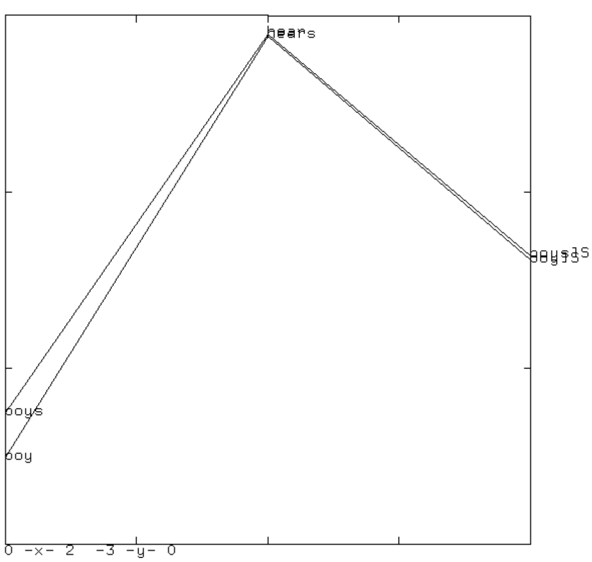

Figure 2
Trajectories through state space for *boy chases boys* and *boys chase boys.* Each point marks the position along the second principle component of hidden unit space, after the indicated word has been input. Magnitude of the second principle component is measured along the ordinate; time (i.e., order of word in sentence) is measured along the abscissa. In this example, differences in the magnitude of the second PC encodes the singular/plural distinction in the subject.

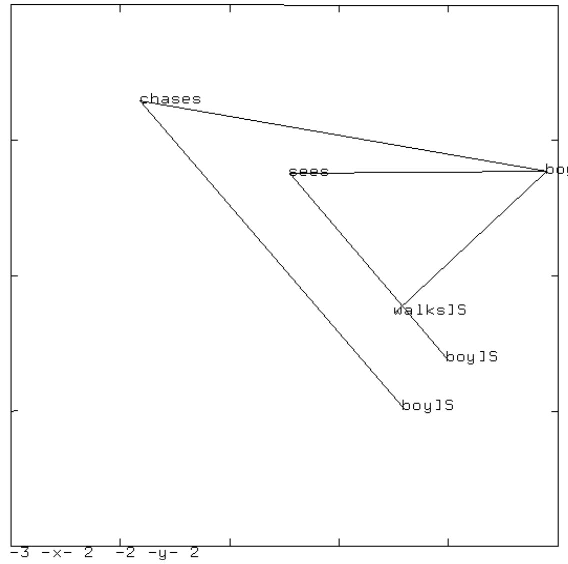

Figure 3
Trajectories through state space for sentences *boy chases boy, boy sees boy, boy walks.* Principal component 1 is plotted along the abscissa; principal component 3 is plotted along the ordinate. These two PC's together encode differences in verb-argument expectations.

The state-space graph shown in Figure 5b was produced under conditions of incremental training, which, we have seen, was crucial for successful learning. What does the state-space look like under conditions of failure, such as when we train a fully-mature network on the adult corpus from the beginning? Figure 5a shows such a plot.

Unlike Figure 5b, Figure 5a reveals a less clearly organized use of the state space. There is far greater variability, and words have noisier internal representations. We do not see the kind of sharp distinctions which are associated with the encoding of number, verb argument type, and embedding as we do when the network has succeeded in mastering the language. Why might this be?

When the network is confronted from the beginning with the entire adult corpus the problem is this. There are actually a relatively small number of sources of variance (number, grammatical category, verb-argument type, and level of embedding). However, these sources of variance interact in complex ways. Some of the interactions involve fairly long-distance dependencies. For example, in the sentence *The girls who the dogs that I chased down the block frightened, ran away*, the evidence that the verb *frightened* is transitive is a bit obscure, because the direct object (*the girls*) not only does not occur after the verb (the normal position for a direct object in simple English sentences), but occurs 10 words earlier; and there are several other nouns and verbs in between. The simple recurrent network does not have perfect memory. All things being

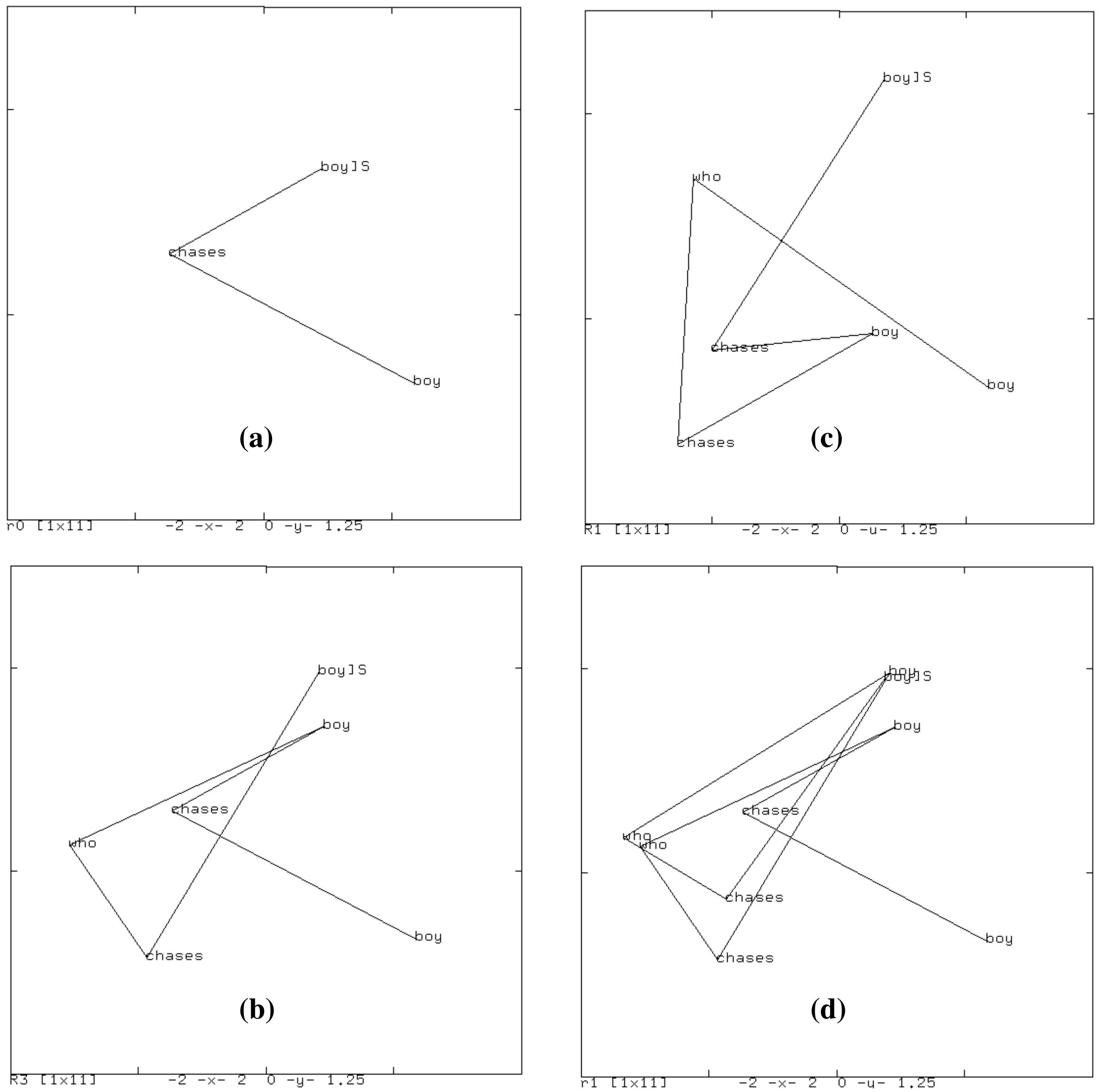

Figure 4

Movement through state space for sentences with relative clauses. Principal component 1 is displayed along the abscissa; principal component 11 is displayed along the ordinate. These two PC's encode depth of embedding in relative clauses.

equal, information decays exponentially. What happens is that the network finds a solution to the task which works enough of the time to yield reasonable performance. However, the solution is imperfect and results in a set of internal representations which do not reflect the true underlying sources of variance. The outcome is, as already pointed out, consistent with the claims of Chomsky (1957) Miller & Chomsky (1963), and Gold (1967).

When learning proceeds in an incremental fashion—either because the environment has been altered or because the network itself is initially handicapped—the result is that the network only sees a subset of the data. When the input is staged, the data are just the simple sentences. When the network is given a limited temporal window, the data are the full adult language but the *effective* data are only those sentences, and portions of sentences, which fall within the window. These are the simple sentences. (Now we see why the initial phase of learning takes a bit longer in this condition; the network also has to wade through a great deal of input which it is

(a)

(b)

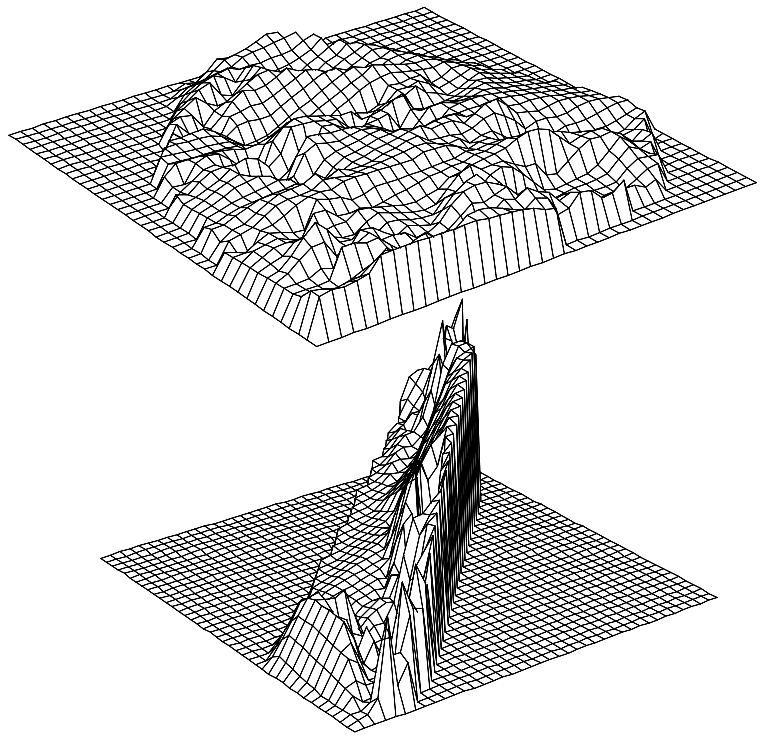

Figure 5

(a) Graph of dimensions which encode embedding (x,y) and number (z) from a network which has failed to learn the grammar. (b) The same graph from a network which has succeeded in the learning task.

essentially noise.)

This subset of data, the simple sentences, contain three of the four sources of variance (grammatical category, number, and verb argument type) and there are no long-distance dependencies. As a result, the network is able to develop internal representations which encode these sources of variance. When learning advances (either because of new input, or because improvements in the network's memory capacity give it a larger temporal window), all additional changes are constrained by this early commitment to the basic grammatical factors.

The effect of early learning, thus, is to constrain the solution space to a much smaller region. The solution space is initially very large, and contains many false solutions (in network parlance, local minima). Whether or not it is really the case that the data truly underdetermine the solution, it does seem to be true that the chances of stumbling on the correct solution are small. However, by selectively focusing on the simpler set of facts, the network appears to learn the basic distinctions—noun/verb/relative pronoun, singular/plural, etc.—which form the necessary basis for learning the more difficult set of facts which arise with complex sentences.

Seen in this light, the early limitations on memory capacity assume a more positive character. It is natural to believe that the more powerful a network, the greater its ability to learn a complex domain. However, this appears not always to be the case. If the domain is of sufficient complexity, and if there are abundant “false solutions”, then the opportunities for failure are great. What is required is some way to artificially constrain the solution space to just that region which contains the true solution. The initial memory limitations fill this role; they act as a filter on the input, and focus learning on just that subset of facts which lay the foundation for future success.

Higher primates are distinguished by their extended period of development, and by their relatively diminished capacities at early stages of development. Ultimately, of course, the higher primates achieve a high level of cognitive functioning and exert extraordinary control over their environment. But that environment is a complex domain, and not easily mastered. It may well be that the early developmental limitations play a key role in learning. The so-called critical period may be critical not by virtue of special capacities which are present during childhood and magically and lamentably lost at puberty. Rather, the critical period may be special because the later abilities which are found in adults have not yet fully developed. The simpler view of the world which this affords makes learning tractable.

Finally, it is worth pointing out that the advantage which accrues to incremental learning does not arise in all circumstances. Consider the extreme case where what is to be learned is a completely random collection of facts. In such circumstances, undersampling the data (which is what incremental learning involves) runs the risk of establishing faulty generalizations. Incremememtal learning can only be useful if (a) the environment contains structure; and (b) the material learned early embodies the major generalizations in a simpler form. In practice, the world is not a random place, and the sorts of things children have to learn about typically contain a great deal of structure. In the specific case of language, it may also be that the filtering which occurs as a result of limited memory picks out just that evidence which is required for successful language learning, and which allow the child to indeed ‘go beyond the data.’

ACKNOWLEDGEMENTS

This research was supported by contract DAAB-07-87-C-H027 from Army Avionics, Ft. Monmouth. I m am grateful to Dick Aslin, Mary Hare, Mike Maratsos, and Virginia Marchman for their comments on an earlier draft of this paper. The author can be reached via email as *elman@crl.ucsd.edu*

REFERENCES

Baker, C.L. (1979). Syntactic theory and the projection problem. *Linguistic Inquiry,* 10, 533-581.

Bowerman, M. (1987). The ‘no negative evidence’ problem: How do children avoid constructing an overly general grammar? In J.A. Hawkins (Ed.), *Explaining language universals.* Oxford: Basil Blackwell.

Braine, M.D.S. (1971). On two types of models of the internalization of grammars. In D.I. Slobin (Ed.), *The ontogenesis of grammar: A theoretical perspective.* New York: Academic Press.

Brown, R., & Hanlon, C. (1970). Derivational complexity and order of acquisition in child speech. In J.R. Hayes (Ed.), *Cognition and the development of language.* New York: Wiley.

Chomsky, N. (1957). *Syntactic structures.* The Hague: Mouton.

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

Elman, J.L. (in press). Distributed representations, simple recurrent networks, and grammatical structure. *Machine Learning.*

Gold, E.M. (1967). Language identification in the limit. *Information and Control*, 16, 447-474.

Gonzalez, R.C., & Wintz, P. (1977). *Digital image processing.* Reading, MA: Addison-Wesley.

Hirsh-Pasek, K., Treiman, R., & Schneiderman, M. (1984). Brown and Hanlon revisited: Mothers' sensitivity to ungrammatical forms. *Journal of Child Language,* 11, 81-88.

Jordan, M. I. (1986). Serial order: A parallel distributed processing approach. Institute for Cognitive Science Report 8604. University of California, San Diego.

Miller, G.A., & Chomsky, N. (1963). Finitary models of language users. In R.D. Luce, R.R. Bush, & E. Galanger (Eds.), *Handbook of Mathematical Psychology, Volume II.* New York: John Wiley.

Pinker, S. (1989). *Learnability and cognition.* Cambridge, MA: MIT Press.

Plunkett, K., & Marchman, V. (1990). From rote learning to system buildling. Center for Research in Language, TR 9020. University of California, San Diego.

Pollack, J.B. (1990). Language acquisition via strange automata. *Proceedings of the Twelfth Annual Conference of the Cognitive Science Society.* Hillsdale, NJ: Erlbaum.

Rumelhart, D.E., Hinton, G.E., & Williams, R.J. (1986). Learning internal representations by error propagation. In D.E. Rumelhart & J.L. McClelland (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition (Vol. 1).* Cambridge, MA: MIT Press.

Servan-Schreiber, D., Cleeremans, A., & McClelland, J.L. (1988). Encoding sequential structure in simple recurrent networks. CMU Technical Report CMU-CS-88-183. Computer Science Department, Carnegie-Mellon University.

Wexler, K., & Cullicover, P. (1980). *Formal principles of language acquisition.* Cambridge, MA: MIT Press.