COGNITION

Cognition 72 (1999) 67–109

Language acquisition in the absence of explicit negative evidence: how important is starting small?

Douglas L.T. Rohde, David C. Plaut*

Carnegie Mellon University and the Center for the Neural Basis of Cognition, Pittsburgh, PA, USA

Received 10 October 1997, received in revised form 29 January 1999; accepted 4 May 1999

Abstract

It is commonly assumed that innate linguistic constraints are necessary to learn a natural language, based on the apparent lack of explicit negative evidence provided to children and on Gold's proof that, under assumptions of virtually arbitrary positive presentation, most interesting classes of languages are not learnable. However, Gold's results do not apply under the rather common assumption that language presentation may be modeled as a stochastic process. Indeed, Elman (Elman, J.L., 1993. Learning and development in neural networks: the importance of starting small. Cognition 48, 71–99) demonstrated that a simple recurrent connectionist network could learn an artificial grammar with some of the complexities of English, including embedded clauses, based on performing a word prediction task within a stochastic environment. However, the network was successful only when either embedded sentences were initially withheld and only later introduced gradually, or when the network itself was given initially limited memory which only gradually improved. This finding has been taken as support for Newport's 'less is more' proposal, that child language acquisition may be aided rather than hindered by limited cognitive resources. The current article reports on connectionist simulations which indicate, to the contrary, that starting with simplified inputs or limited memory is not necessary in training recurrent networks to learn pseudo-natural languages; in fact, such restrictions hinder acquisition as the languages are made more English-like by the introduction of semantic as well as syntactic constraints. We suggest that, under a statistical model of the language environment, Gold's theorem and the possible lack of explicit negative evidence do not implicate innate, linguistic-specific mechanisms. Furthermore, our simulations indicate that special teaching methods or maturational constraints may be unnecessary in learning the structure of natural language.

Keywords: Explicit negative evidence; Language acquisition; Connectionist simulations

* Corresponding author. School of Computer Science, Carnegie Mellon University, 4400 Fifth Avenue, Pittsburgh, PA 15213-2583, USA. Fax: +1-412-2685060.
E-mail address: plaut@cmu.edu (D.C. Plaut)

0010-0277/99/$ - see front matter
PII: S0010-0277(99)00031-1

1. Introduction

Traditionally, the problem of language acquisition has been treated as a problem of learning to identify and produce the valid sentences in one's language. The idealized speaker is presumed to possess a set of rules, or *competence grammar*, capable of generating all well-formed sentences or determining whether any sentence is valid or invalid. The learning process is driven both by the learner's innate endowment of structured linguistic knowledge and by the learner's exposure to language. Fundamental questions thus concern the nature of these sources of information, how they are utilized, and the extent to which each is responsible for the eventual attainment of language skill.

The standard approach in linguistics has tended to view the input to the child learner simply as a sequence of valid sentences. Statistical properties of this input are generally overlooked or thought to bear little relevance to learning. Indeed, some consider this a feature of the approach as attention to statistics potentially places a tremendous computational burden on the learner (see Allen and Seidenberg, 1999, for discussion). Additionally, Baker (1979), among others, has argued that children receive negligible explicit negative feedback following production errors.[1]

One virtue of a simple model of the language environment is that it facilitates the investigation of formal proofs of the learnability or unlearnability of certain problems. In particular, the theoretical findings of Gold (1967) have led to the widely accepted hypothesis that the burden of language learning lies primarily on our genetic endowment and only secondarily on actual language exposure. In short, Gold proved, under certain assumptions, that no superfinite class of languages is learnable by any learner without negative examples. Among the superfinite classes of languages is the set of regular languages, recognizable by finite-state machines, as well as the classes of context-free and context-sensitive languages, which are believed to be more closely related to natural languages. A critical assumption in Gold's model is that the language input consists of a nearly arbitrary sequence of positive examples, subject only to the constraint that no sentence may be withheld from the learner indefinitely.

Gold recognized the problem his findings posed for natural language acquisition and offered three solutions. The first is that the child may make use of some subtle or covert negative evidence in the parental responses to the child's utterances. Researchers who emphasize the role of environmental input in language acquisition have principally focused on this issue, arguing that subtle feedback is available to the child and is correlated with improved long-term learning (see Sokolov and Snow, 1994 for review). Although the extent to which parents do indeed provide

[1]We will use the term *explicit* negative evidence to refer to feedback given to the child in response to the child's utterances. One can further distinguish between *overt* explicit negative evidence, such as direct statements that a particular sentence is ungrammatical, and *subtle* or *covert* explicit evidence, such as a greater tendency for parents to rephrase ungrammatical compared with grammatical utterances. In contrast, we will use *implicit* negative evidence to refer to distributional properties of the input which do not depend on the language production of the learner. Implicit negative evidence is sometimes referred to as *indirect*, although we favor the former term.

either overt or covert explicit feedback is a matter of ongoing debate, it seems unlikely that this feedback would be sufficiently robust to overcome Gold's problem.

The second solution proposed by Gold is that the class of possible natural languages is smaller than expected and that the child has some innate knowledge identifying this class. This is the solution that has been most readily accepted in the linguistics community and is associated with the theories of Universal Grammar and the innate Language Acquisition Device. Given the apparent lack of explicit negative evidence provided to children, strong innate linguistic constraints are regarded by many authors (e.g. Berwick, 1985; Morgan and Travis, 1989; Marcus, 1993; Morgan et al., 1995) to be an inescapable solution to the learnability problem. On the surface, it seems perfectly reasonable to hypothesize that the set of natural languages is limited: It is unlikely that *every* regular or every context-free language is a possible natural language. However, even under this assumption, most interesting subsets of these language classes would still be unlearnable under Gold's model. It remains to be seen what degree of constraints, if any, would enable the learning of natural language in Gold's framework.

However, Gold made brief mention of a third possibility: that his assumption regarding the possible texts (or sequences of positive examples) for a language was too general and that 'there is an a priori restriction on the class of texts which can occur' (p. 454). In Gold's model, a fair text is a series of positive examples from the language in which every legal sentence will eventually occur. Superfinite languages were found to be unlearnable only if texts are arbitrary or are produced by the powerful class of recursive functions. Such a function can prohibit learning by producing a series of examples designed specifically to confuse the learner indefinitely. However, this hardly seems an appropriate model for a child's linguistic environment. While there is ongoing debate on the extent to which child-directed speech is simplified relative to adult-directed speech (see e.g. Snow and Ferguson, 1977; Gallaway and Richards, 1994) no one would propose that it is tailored specifically to *hinder* language acquisition.

An alternative is to constrain the possible texts by modeling language as a stochastic process – some sentences or grammatical constructions are more frequent than others and language is generated by a relatively stationary distribution over these strings (see Seidenberg, 1997; Seidenberg and MacDonald, 1999). The statistical structure of a stochastically generated text provides an implicit source of negative evidence. Essentially, if a particular grammatical construction is not observed during some extended but finite exposure, one can safely assume that it is not part of the language.[2] With more exposure, the probability of making an error decreases. Note, though, that deriving evidence from non-occurrence within a finite

[2]The term 'construction' here refers to grammatical distinctions, abstractions or rules rather than to specific sentences. Thus, for example, the famous sentence of Chomsky (1957), 'Colorless green ideas sleep furiously', is supported by the input as one of many simple active SVO sentences. Although connectionist networks might not instantiate such constructions as explicit, distinct data structures, these systems nonetheless have the capability of developing internal distributed representations that support effective generalization across sentences with similar grammatical structure (in the classic sense).

sample is invalid without a more limited source than Gold's text. The difficulty in learning from an arbitrary text derives largely from the possibility that a construction that is important to the language has been withheld from all prior sentences. However, given a stochastic text, a construction that does not appear for a very long time has a very small chance of being an important part of the language and can thus be ignored at little cost.

While a stochastic model of text generation is perhaps still overly weak, as it neglects the influence of context on sentence selection, it is nonetheless sufficient to allow learnability. Indeed, Horning (1969) and Angluin (1988) have proved, under slightly different criteria for convergence, that stochastic context-free languages are learnable from only positive examples. Angluin notes that there is an important similarity between this result and Gold's positive finding that even recursively enumerable languages are learnable from texts generated by primitive recursive functions, as opposed to fully recursive functions. If we accept that a stochastic text is a more reasonable approximation to a child's linguistic input than an arbitrary text, Gold's findings no longer pose a 'logical problem' (Baker and McCarthy, 1981) for language acquisition.

It is important to note, though, that a stochastic view of language leads to a rather different definition of what it means to learn a language. On the traditional view, learning a language involves converging on the single, correct grammar of the language; any deviation from this grammar in the actual behavior of language users must be ascribed to performance factors. Moreover, given that all learners of a language must acquire competence in equivalent grammars, it is critical to have formal guarantees that this will happen. From a stochastic perspective, by contrast, the grammars acquired by members of a language community need not be identical but only sufficiently similar to permit effective communication. The degree of agreement among individuals in, for example, making grammaticality judgments would thus be expected to be very high but not perfect. It is still possible to formulate explicit bounds on learnability, but these bounds are probabilistic rather than absolute. Moreover, on this view, the study of actual language performance plays a more central role than on traditional views because such performance is taken to reflect underlying language knowledge more directly.

This leads to a serious practical problem. The human brain is considerably restricted as a learning device due to its limited memory and analytical abilities. The principal mechanisms of language acquisition seem to operate online with relatively little storage and subsequent analysis of the actual inputs. In contrast, the learning mechanisms proposed by Horning, Angluin, and others rely on repeated evaluation and re-evaluation of vast sets of complete, candidate grammars. They are thus unlikely to lead to reasonable computational models of our language acquisition mechanism.

Given restrictions of limited memory and online learning with iterative updates of a small set of candidate grammars, one way the statistical structure of a language can be approximated is through the formulation and testing of implicit predictions. By comparing one's predictions to what actually occurs, feedback is immediate and

negative evidence derives from incorrect predictions. Although not emphasizing online prediction, Chomsky (1981) followed Gold (1967) in pointing out the potential importance to language acquisition of 'expectations':

> A not unreasonable acquisition system can be devised with the operative principle that if certain structures or rules fail to be exemplified in relatively simple expressions, where they would be expected to be found, then a possibly marked) option is selected excluding them in the grammar, so that a kind of 'negative evidence' can be available even without corrections, adverse reactions, etc. (p. 9; emphasis added).

The ability to predict utterances in a language is surprisingly powerful. Accurate prediction is equivalent to possessing a grammar able to produce a language or to decide the grammaticality of any sentence. Prediction must be based on a *language model*, which has been found to be essential in many forms of automated natural language processing, such as speech recognition (Huang et al., 1990). More generally, in learning complex, goal-directed behavior, prediction can provide the feedback necessary to learn an internal *forward model* of how actions relate to outcomes (Jordan, 1992; Jordan and Rumelhart, 1992). Such a model can be used to convert 'distal' discrepancies between observable outcomes and goals into the 'proximal' error signals necessary for learning, thereby obviating the need for externally provided error signals. An important additional feature of prediction is that feedback is available immediately; the learner need not perform a re-analysis of previously observed positive evidence (cf. Marcus, 1993). Again, it should be emphasized that theoretical proposals involving expectation or prediction are precluded under Gold's model because past experience with the language is not necessarily representative of future experience.

It remains, then, to be demonstrated that a computational system can acquire a language under stochastic text presentation without relying on inappropriate memory or time requirements. Towards this end, Elman (1991, 1993) provided an explicit formulation of how a general connectionist system might learn the grammatical structure of a language on the basis of performing a prediction task. He trained a simple recurrent network (Elman, 1990; sometimes termed an 'Elman' network) to predict the next word in sentences generated by an artificial grammar exhibiting number agreement, variable verb argument structure, and embedded clauses. Although word prediction is a far cry from language comprehension, it can be viewed as a useful component of language processing to the extent that learning a grammar is useful, given that the network can make accurate predictions only by learning the structure of the grammar. Elman found that the network was unable to learn the prediction task – and, hence, the underlying grammar – when presented from the outset with sentences generated by the full grammar. The network was, however, able to learn if it was trained first on only simple sentences (i.e. those without embeddings) followed by an increasing proportion of complex sentences, or if the network's memory span was initially reduced and gradually allowed to improve. The fact that learning was successful only under conditions of restricted

input or restricted memory is what Elman (1993) referred to as 'the importance of starting small.'

Elman's finding that simplifying a network's training environment or limiting its computational resources was necessary for effective language learning accords well with Newport's 'less is more' proposal (Newport, 1990; Goldowsky and Newport, 1993) that the ability to learn a language declines over time as a result of an *increase* in cognitive abilities. This hypothesis is based on evidence that early and late learners seem to show qualitative differences in the types of errors they make. It has been suggested that limited abilities may force children to focus on smaller linguistic units which form the fundamental components of language, rather than memorizing larger units which are less amenable to recombination. In terms of Elman's network, it is possible that staged input or limited memory similarly caused the network to focus early on simple and important features, such as the relationship between nouns and verbs. By 'starting small', the network had a better foundation for learning the more difficult grammatical relationships which span potentially long and uninformative embeddings.

We set out in the current work to investigate whether the need for starting small in learning a pseudo-natural language might be less critical if the language incorporated more of the constraints of natural languages. A salient feature of the grammar used by Elman is that it is purely syntactic, in the sense that all words of a particular class, such as the singular nouns, were identical in usage. A consequence of this is that embedded material modifying a head noun provides relatively little information about the subsequent corresponding verb. Earlier work by Cleeremans et al. (1989), however, had demonstrated that simple recurrent networks were better able to learn long-distance dependencies in finite-state grammars when intervening sequences were partially informative of (i.e. correlated with) the distant prediction. The intuition behind this finding is that the network's ability to represent and maintain information about an important word, such as the head noun, is reinforced by the advantage this information provides in predicting information within embedded phrases. As a result, the noun can more effectively aid in the prediction of the corresponding verb following the intervening material.

One source of such correlations in natural language are distributional biases, due to semantic factors, on which nouns typically co-occur with which verbs. For example, suppose dogs often chase cats. Over the course of training, the network has encountered chased more often after processing sentences beginning *The dog who...* than after sentences beginning with other noun phrases. The network can, therefore, reduce prediction error within the embedded clause by retaining specific information about the dog (beyond it being a singular noun). As a result, information on dog becomes available to support further predictions in the sentence as it continues (e.g. *The dog who chased the cat barked*).

These considerations led us to believe that languages similar to Elman's but involving weak semantic constraints might result in less of an advantage for starting small in child language acquisition. We began by examining the effects of an incremental training corpus, without manipulating the network's memory. In the first simulation study reported here, we found, somewhat surprisingly, that the

addition of semantic constraints not only resulted in less of an advantage for starting small but in a significant advantage for starting with the full complexity of the language. Moreover, and in accordance with the results of Cleeremans and colleagues, the advantage for 'starting large' increased as the language was made more English-like by strengthening the semantic constraints.

In order to better understand the discrepancy between our results and those of Elman (1991, 1993), in a second study we attempted a more direct replication of Elman's grammar and methods. Using a similar grammar but our own training methods, we again found a disadvantage for starting small. With parameters similar to those used by Elman, however, the network failed to learn the task well in either condition. Altering these methods by increasing the range of the initial connection weights resulted in much-improved performance but a clear advantage for starting with the full grammar. In fact, we found no advantage for starting with a simplified training corpus even when the target language contains no simple sentences. Only in extreme conditions involving no simple sentences and embedded clauses which are unrelated to the word being modified did we find an advantage for starting small. It thus appears that the benefit of starting with simplified inputs is not a robust result for the acquisition of such languages by simple recurrent networks.

There remained the possibility that an advantage for starting small would hold for networks with initially restricted memory, which is the condition Elman (1993) interpreted as a more appropriate approximation to child language acquisition. To test this possibility, we carried out a third simulation study involving the same memory manipulation as Elman, using two different grammars and several combinations of training parameters. Under no circumstances did we find a significant difference between the results with full memory and the results with initially limited memory. Therefore, although early memory impairments do not significantly hinder language learning, they do not seem to provide any advantage in our experiments.

Based on the results of these simulation studies, we argue that, in learning the structure of pseudo-natural languages through prediction, it is an inherent property of simple recurrent networks that they extract simple, short-range regularities before progressing to more complex structures. No manipulation of the training corpus or network memory is necessary to induce this bias. Thus, the current work calls into question whether effective child language acquisition depends on, or even benefits from, initially limited cognitive resources or other maturational constraints. In the General Discussion we address open issues in early versus late exposure to language and question the necessity of either explicit negative evidence or innate linguistic constraints in language acquisition under a model of language that promotes the importance of statistical information.

2. Simulation 1: Progressive inputs

Elman (1991) was interested in demonstrating how, and indeed if, a recurrent network could represent complex structural relations in its input. A task was chosen in which sentences were presented one word at a time, and the network was trained

to predict each successive word. The ability of the network to perform well is indicative of its ability to represent and use the structural relations in the grammar.

A notable limitation of Elman's grammar was that it was purely syntactic. The goal of our initial simulation was to extend Elman's work to apply to a more naturalistic language. In particular, we set out to study the effect of making the grammar more natural through the addition of semantic constraints (i.e. restrictions on noun-verb relationships). Given the findings of Cleeremans et al. (1989), that even subtle information in an embedding can aid the learning of long-distance dependencies, we hypothesized that the addition of semantic constraints might reduce the advantage for starting small.

2.1. Method

The methods used in the simulation are organized below in terms of the grammar used to generate the artificial language, the network architecture, the training corpora generated from the grammar, the procedures used for training the network, and the way in which the performance of the network was tested. In general, these methods are very similar to those used by Elman (1991, 1993); differences are noted explicitly throughout.

2.1.1. Grammar

The pseudo-natural language used in the current simulation was based on the grammar shown in Table 1. The grammar generates simple noun-verb and noun-verb-noun sentences with the possibility of relative clause modification of nouns. The grammar involved 10 nouns and 14 verbs, as well as the relative pronoun *who* and an end-of-sentence marker (here denoted '.'). Four of the verbs were transitive, four were intransitive, and five were optionally transitive. Six of the nouns and seven of the verbs were singular, the others plural. Finally, number agreement was enforced between subjects and verbs, where appropriate. Relative clauses could be nested, producing sentences such as:

girls who cat who lives chases walk dog who feeds girl who cats walk .

Although this language is highly simplified from natural language, it is nonetheless of interest because, in order to learn to make accurate predictions, a network must form representations of potentially complex syntactic structures and remember

Table 1
The context-free grammar used in Simulation[a]

S	→	NP VI . | NP VT NP .
NP	→	N | NRC
RC	→	who VI | who VT NP | who NP VT
N	→	boy | girl | cat | dog | Mary | John | boys | girls | cats | dogs
VI	→	barks | sings | walks | bites | eats | bark | sing | walk | bite | eat
VT	→	chases | feeds | walks | bites | eats | chase | feed | walk | bite | eat

[a]Transition probabilities are specified and additional constraints are applied on top of this framework.

information, such as whether the subject was singular or plural, over lengthy embeddings. The grammar used by Elman was nearly identical to the current one, except that it had one fewer mixed transitivity verb in singular and plural form, and the two proper nouns, *Mary* and *John*, could not be modified.

In the current work, several additional constraints were applied on top of the grammar in Table 1. Primary among these was that individual nouns could engage only in certain actions, and that transitive verbs could act only on certain objects. For example, anyone could walk, but only humans could walk something else and the thing walked must be a dog. The full set of constraints are listed in Table 2.

Another restriction in the language was that proper nouns could not act on themselves. For example, *Mary chases Mary* would not be a legal sentence. Finally, constructions which repeat an intransitive verb, such as *Boys who walk walk*, were disallowed because of redundancy. These and the above constraints will be referred to as semantic constraints. In the simulation, semantic constraints always applied within the main clause of the sentence as well as within any subclauses. Although number agreement affected all nouns and verbs, the degree to which the semantic constraints applied between a noun and its modifying phrase was controlled by specifying the probability that the relevant constraints would be enforced for a given phrase. In this way, effects of the correlation between a noun and its modifying phrase, or of the level of information the phrase contained about the identity of the noun, could be investigated.

Two other parameters were used to control the behavior of the grammar. First, the framework depicted in Table 1 was modified to allow the direct specification of the percentage of simple and complex sentences produced. Second, the probability of noun phrase modification was adjusted to control the average length of sentences in the language.

When probabilities are specified for the productions in the grammar, it becomes a stochastic context-free grammar (SCFG). A grammar of this form is convenient not only for generating example sentences, but also because it allows us to calculate the optimal prediction behavior on the language. Given the stochastic nature of the language, the network cannot in general predict the actual next word in a sentence accurately. Rather, over the course of training, we expect the network to increasingly

Table 2
Semantic constraints on verb usage[a]

Verb	Intransitive subjects	Transitive subjects	Objects if transitive
chase	–	any	any
feed	–	human	animal
bite	animal	animal	any
walk	any	human	only dog
eat	any	animal	human
bark	only dog	–	–
sing	human or cat	–	–

[a]Columns indicate legal subject nouns when verbs are used intransitively or transitively and legal object nouns when transitive.

approximate the theoretically correct prediction given the sentence context up to the current point, in the form of a probability distribution over the 26 words in the vocabulary. One advantage of expressing the language as an SCFG is that this probability distribution can be computed exactly. However, the above mentioned number agreement and semantic constraints are difficult to incorporate into the basic grammar shown in Table 1. Therefore, a program was developed (Rohde, 1999) which takes the grammar, along with the additional constraints, and produces a new, much larger SCFG with the constraints incorporated into the stochastic, context-free transitions. In this way, a single SCFG could be produced for each version of the grammar and then used to generate sentences or to specify optimal predictions.

2.1.2. Network architecture

The simple recurrent network used in both Elman's simulations and in the current work is shown in Fig. 1. Inputs were represented as localist patterns or basis vectors: Each word was represented by a single unit with activity 1.0, all other units having activity 0.0. This representation was chosen to deprive the network of any similarity structure among the words that might provide indirect clues to their grammatical properties. The same 1-of-n representation was also used for outputs, which has the convenient property that the relative activations of multiple words can be represented independently. Although Elman reserved two of the input and output units for another purpose, all 26 units were used in Simulation 1. The two small 10-unit hidden layers were provided to allow the network to re-represent localist inputs in a distributed fashion and to perform a more flexible mapping from the main hidden layer to the output. These layers have the additional benefit of reducing the total number of connections in the model; a direct projection from 26 units to 70 units requires 1820 connections, whereas the same projection via 10 intermediate units requires only 970 connections.

On each time step, a new word was presented by fixing the activations of the input layer. The activity in the main hidden layer from the previous time step was copied

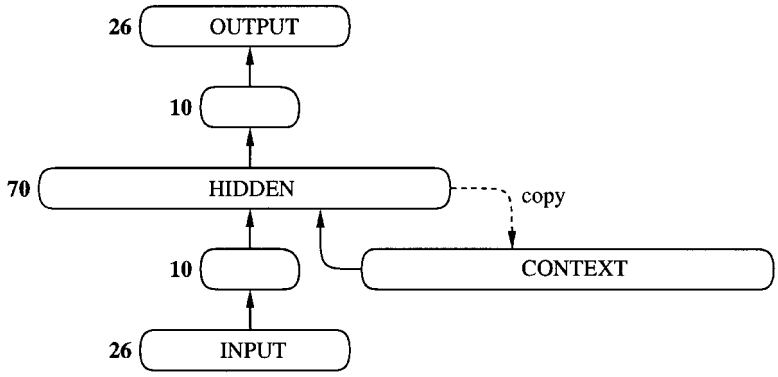

Fig. 1. The architecture of the network used in the simulations. Each solid arrow represents full connectivity between layers (with numbers of units next to each layer). Hidden unit states are copied to corresponding context units (dashed arrow) after each word is processed.

to the *context* layer. Activation then propagated through the network, as in a feed-forward model, such that each unit's activation was a smooth, non-linear (logistic) function of its summed weighted input from other units. The resulting activations over the output units were then compared with their target activations. In a simple recurrent network, errors are not back-propagated through time (cf. Rumelhart et al., 1986) but only through the current time step, although this includes the connections from the context units to the hidden units. These connections allow information about past inputs – as encoded in the previous hidden representation copied onto the context units – to influence current performance. Although the target outputs used during training were the encoding for the actual next word, typically a number of words were possible at any given point in the sentence. Therefore, to perform optimally the network must generate, or predict, a probability distribution over the word units indicating the likelihood that each word would occur next. Averaged across the entire corpus, this distribution will result in the lowest performance error on most any measure, including squared error and Kullback-Leibler divergence (see Rumelhart et al., 1995). Table 3 contains the formulae used to calculate these and the other error measures discussed in the current work.

Sentences in the corpora were concatenated together and context units were not reinitialized at sentence boundaries. Note, however, that it is trivial for the network to learn to be sensitive to the start of a sentence, as the end-of-sentence marker is a perfectly reliable indicator of sentence breaks.

2.1.3. Corpora

Initially, Elman produced a corpus of 107 000 sentences, 75% of which were 'complex' in that they contained at least one relative clause. Despite experimenting with various architectures, starting conditions, and learning parameters, Elman

Table 3
Error measures used in testing the network

Error measure	Formula
City-Block	$\sum_i \lvert t_i - o_i \rvert$
Squared Error	$\sum_i (t_i - o_i)^2$
Cosine	$\sum_i t_i o_i \left(\sum_i t_i^2 \sum_i o_i^2 \right)^{-1/2}$
Divergence	$\sum_i t_i \log(t_i / o_i)$

o_i is the activation of the ith output unit on the current word and t_i is its target or desired activation.

(1991) reported that 'the network was unable to learn the task when given the full range of complex data from the beginning' (p. 100). In response to this failure, Elman designed a staged learning regimen, in which the network was first trained exclusively on simple sentences and then on an increasing proportion of complex sentences. Inputs were arranged in four corpora, each consisting of 10 000 sentences. The first corpus was entirely simple, the second 25% complex, the third 50% complex, and the final corpus was 75% complex, as was the initial corpus that the network had failed to learn when it alone was presented during training. An additional 75% complex corpus, generated in the same way as the last training corpus, was used for testing the network.

In order to study the effect of varying levels of information in embedded clauses, we constructed five grammar classes. In class A, semantic constraints did not apply between the clause and its subclause, only within a clause. In class B, 25% of the subclauses respected the semantic constraints, 50% in class C, 75% in class D, and 100% in class E. Therefore, in class A, which was most like Elman's grammar, the contents of a relative clause provided no information about the noun being modified other than whether it was singular or plural, whereas class E produced sentences which were the most English-like. We should emphasize that, in this simulation, semantic constraints always applied within a clause, including the main clause. This is because we were interested primarily in the ability of the network to perform the difficult main verb prediction, which relied not only on the number of the subject, but on its semantic properties as well. In the second simulation, we will investigate a case in which all the semantic constraints were eliminated to produce a grammar essentially identical to Elman's.

As in Elman's work, four versions of each class were created to produce languages of increasing complexity. Grammars A_0, A_{25}, A_{50}, and A_{75}, for example, produce 0%, 25%, 50%, and 75% complex sentences, respectively. In addition, for each level of complexity, the probability of relative clause modification was adjusted to match the average sentence length in Elman's corpora, with the exception that the 25% and 50% complex corpora involved slightly longer sentences to provide a more even progression, reducing the large difference between the 50% and 75% complex conditions apparent in Elman's corpora. Specifically, grammars with complexity 0%, 25%, 50%, and 75% had 0%, 10%, 20%, and 30% modification, respectively. The average sentence lengths for each of the training corpora used in the current simulation, as well as Elman's, are given in Table 4.

Table 4
Average length of sentences generated by grammar classes

	Grammar class						
% Complex	A	B	C	D	E	R[a]	Elman
0	3.50	3.50	3.50	3.50	3.50	3.46	3.46
25	4.20	4.19	4.20	4.19	4.18	3.94	3.92
50	5.04	5.07	5.07	5.06	5.06	4.39	4.38
75	6.05	6.04	6.04	6.06	6.06	6.02	6.02

[a]Used in Simulation 2.

For each of the 20 grammars (five levels of semantic constraints crossed with four percentages of complex sentences), two corpora of 10 000 sentences were generated, one for training and the other for testing. Corpora of this size are quite representative of the statistics of the full language for all but the longest sentences, which are relatively infrequent. Sentences longer than 16 words were discarded in generating the corpora, but these were so rare ($<0.2\%$) that their loss should have had negligible effects. In order to perform well, the network cannot possibly 'memorize' the training corpus but must learn the structure of the language.

2.1.4. Training procedure

In the condition Elman referred to as 'starting small', he trained his network for five epochs on each of the four corpora, in increasing order of complexity. During training, weights were adjusted to minimize the summed squared error between the network's predicted next word and the actual next word, using the back-propagation learning procedure (Rumelhart et al., 1986) with a learning rate of 0.1, reduced gradually to 0.06. No momentum was used and weights were updated after each word presentation. Weights were initialized to random values sampled uniformly between ± 0.001.

For each of the five language classes, we trained the network shown in Fig. 1 using both incremental and non-incremental training schemes. In the *complex* regimen, the network was trained on the most complex corpus (75% complex) for 25 epochs with a fixed learning rate. The learning rate was then reduced for a final pass through the corpus. In the *simple* regimen, the network was trained for five epochs on each of the first three corpora in increasing order of complexity. It was then trained on the fourth corpus for 10 epochs, followed by a final epoch at the reduced learning rate. The final six epochs of training on the fourth corpus, not included in Elman's design, were intended to allow performance with the simple regimen to approach asymptote.

Because we were interested primarily in what performance level was possible under optimal conditions, we searched a wide range of training parameters to determine a set which consistently achieved the best performance overall.[3] We trained our network with back-propagation using momentum of 0.9, a learning rate of 0.004 reduced to 0.0003 for the final epoch, a batch size of 100 words per weight update, and initial weights sampled uniformly between ± 1.0 (cf. ± 0.001 for Elman's network). Network performance for both training and testing was measured in terms of divergence (see Table 3). In addition to being an appropriate measure of the difference between two distributions from an information theoretic standpoint (see Rumelhart et al., 1995), divergence has the feature that, during training, error is injected only at the unit representing the actual next word. This is perhaps more plausible than functions which provide feedback to every word in the vocabulary.

Because divergence is well-defined only over probability distributions (which sum to 1.0), normalized Luce ratios (Luce, 1986), also known as *softmax* constraints,

[3]The effects of changes to some of these parameter values, in particular the magnitude of initial random weights, will be evaluated in a later simulation.

were applied to the output layer. In this form of normalization, the activation of output unit i is calculated by $o_i = e^{x_i}/\sum_j e^{x_j}$, where x_i is the unit's net input and j ranges over all of the output units. The remaining units in the network used the standard logistic activation function, $o_i = (1 + e^{-x_i})^{-1}$, as in Elman's network.

2.1.5. Testing procedure

Although the network was trained by providing feedback only to the actual next word in the sentence, the prediction task is probabilistic. Consequently, the network cannot possibly achieve perfect performance if evaluated against the actual next word. Optimally, the network should produce a distribution over its outputs indicating the likelihood of each word occurring next given the sentence context encountered so far. Because our grammars were in standard stochastic, context-free form, it was possible to generate the theoretically correct next-word distributions given any sentence context. Such distributions were calculated for each word in the final testing corpus and the performance of our network was evaluated against these optimal predictions. By contrast, it was not possible to generate such optimal predictions based on Elman's grammar. In order to form an approximation to such predictions, Elman trained an empirical language model on sentences generated in the same way as the testing corpus. Predictions by this model were based on the observed next-word statistics given every sentence context to which it was exposed. This can be thought of as an n-gram model or a k-limited Markov source whose context can extend back to the beginning of the sentence, but no further.

2.2. Results and discussion

Although Elman did not provide numerical results for the *complex* condition, he reported that his network was unable to learn the task when trained on the most complex corpus from the start. However, learning was effective in the *simple* regimen, in which the network was exposed to increasingly complex input. In this condition, Elman found that the network achieved an overall error of 0.177 when compared against the empirical model (using, we believe, city-block distance; see Table 3). However, this type of criterion is not a particularly good measure of the difference between two probability distributions. A better indicator is the mean cosine of the angle between the prediction vectors, by which the network achieved a value of 0.852 (SD = 0.259) where 1.0 is optimal.

Fig. 2 shows, for each training condition, the mean divergence error per word on the testing corpora of our network when evaluated against the theoretically optimal predictions given the grammar. To reduce the effect of outliers, and because we were interested in the best possible performance, results were averaged over only the best 16 of 20 trials. Somewhat surprisingly, rather than an advantage for starting small, the data reveals a significant advantage for the complex training regimen ($F_{1,150} = 53.8$, $P < 0.001$). Under no condition did the simple training regimen outperform the complex training. Moreover, the advantage in starting complex increased with the proportion of fully constrained relative clauses. Thus, there was a strong positive correlation across individual runs ($r = 0.75$, $P < 0.001$) between the order of the

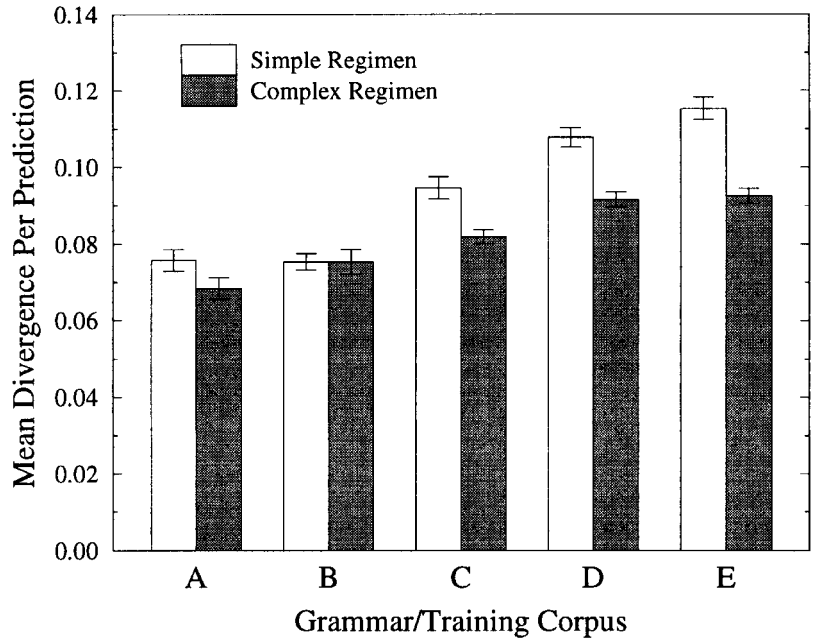

Fig. 2. Mean divergence per word prediction over the 75% complex testing corpora generated from grammar classes A through E (increasing in the extent of semantic constraints) for the simple and complex training regimes. Note that lower values correspond to better performance. Means and standard errors were computed over the best 16 of 20 trials in each condition.

grammars from A to E and the difference in error between the simple versus complex training regimes. This is consistent with the idea that starting small is most effective when important dependencies span uninformative clauses. Nevertheless, against expectations, starting small failed to improve performance even for class A, in which relative clauses did not conform to semantic constraints imposed by the preceding noun.

2.2.1. Has the network learned the task?

In interpreting these results, it is important to establish that the network was able to master the task to a reasonable degree of proficiency in the complex regimen. Otherwise, it may be the case that none of the training conditions produced effective learning, rendering any differences in performance irrelevant to understanding human language acquisition. Average divergence error was 0.068 for the network when trained on corpus A_{75} and 0.093 when trained on corpus E_{75}, compared with an initial error of approximately 2.6. The class E languages yielded slightly higher error because semantic constraints force the network to make use of more information in predicting the contents of relative clauses. Informal inspection revealed that the network appeared to perform nearly perfectly on sentences with up to one relative clause and quite well on sentences with two relative clauses.

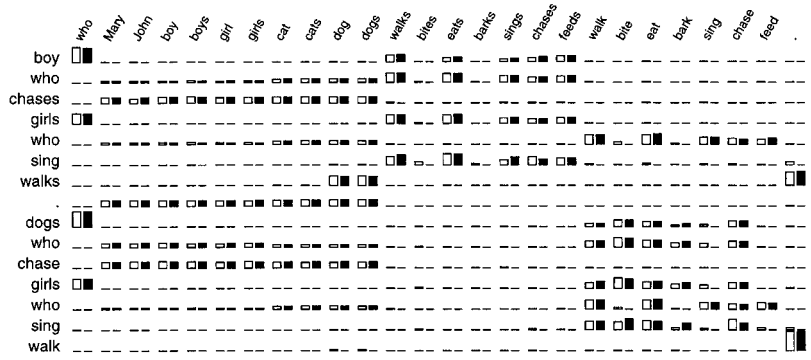

Fig. 3. Predictions of the network trained on corpus E_{75} on two sample sentences (white bars) compared with the optimal predictions given the grammar (filled bars). To enhance contrast, all values shown are the square root of the actual probabilities.

Fig. 3 compares the output activations of a network trained exclusively on corpus E_{75} with the optimal outputs for that grammar. The behavior of the network is illustrated for the sentences *Boy who chases girls who sing walks* and *Dogs who chase girls who sing walk*. Note, in particular, the prediction of the main verb following sing. Predictions of this verb are not significantly degraded even after two embedded clauses. The network is clearly able to recall the number of the main noun and has a basic grasp of the different actions allowed on dogs and humans. It nearly mastered the rule that dogs cannot walk something else. It is, however, unsure across a double embedding that boys are not allowed to bite and that dogs may bark, but not sing. Otherwise, the predictions appear quite close to optimal.

For sentences with three or four clauses, such as *Dog who dogs who boy who dogs bite walks bite chases cat who Mary feeds*, performance of the networks was considerably worse. Note, however, that humans are generally unable to parse such sentences without multiple readings. In addition, fewer than 5% of the sentences in the most complex corpora were over nine words long. This limitation was necessary in order to match the average sentence-length statistics in Elman's corpora, but it did not provide sufficient exposure to such sentences for the network to master them. Interestingly, the network was only 8.2% worse on the testing set than on the training set when trained on corpus E_{75}, and only 5.4% worse when trained on A_{75}. These findings indicate that the network generalized quite well to novel sentences but was still slightly sensitive to the particular characteristics of the training corpus.

However, it should be noted that this analysis is not a clean test of generalization as many of the shorter sentences in the testing corpus appeared in the training corpus as well. Table 5 gives a breakdown of performance of a sample network from the previous analysis, which was trained only on the E_{75} corpus, on those sentences that appeared on both the training and testing set ('Familiar Sentences') and those only in

[4]The comparison for simple sentences and for very complex sentences is unreliable because there were very few novel simple sentences and no very complex sentences that appeared both during training and testing.

Table 5
Analysis of the E_{75} testing corpus and performance of a network on familiar and novel sentences

Relative clauses	Total sentences	Unique sentences	Percent novel	Mean divergence error: Familiar sentences	Mean divergence error: Novel sentences	Example novel sentence
0	2548	230	1.3	0.011	0.019	boy chases dog.
1	5250	2413	53.4	0.043	0.045	dogs who John walks chase girl.
2	1731	1675	94.3	0.110	0.123	dog who chases John who feeds cats bites Mary.
3	395	395	100	0.242	0.247	John feeds cats who bite cats who Mary who walks dog feeds.
4	76	76	100	0.359	0.364	girls who walk dogs who bite Mary who cats who chase Mary chase sing.
Overall	10000	4789	69.8			

the testing set ('Novel Sentences'). The results indicate that the mean divergence error per word of the network was only 3.5% greater on novel versus familiar sentences involving one relative clause and 16.6% greater on novel sentences involving two relative clauses.[4] Thus, the network generalized fairly well, but certainly not perfectly.

A stronger test than predicting individual words for whether a network has learned a grammar is the one standardly employed in linguistic studies: grammaticality judgment of entire sentences. Although the word-prediction networks do not deliver overt yes/no responses to grammatical versus ungrammatical sentences, we assume this decision can be based on the accuracy of its predictions throughout a given sentence (see also Allen and Seidenberg, 1999). Specifically, the word encountered at the point at which a sentence becomes ungrammatical will be poorly predicted and will likely cause poor predictions for subsequent words. Accordingly, as a simple approximation, we selected the two words that were most 'surprising' to the network (those to which the network assigned the least likelihood) and took the log of the product of the two likelihoods as a measure of the 'goodness' of the sentence for the purpose of judging its grammaticality.

In order to obtain grammatical and ungrammatical sentences for this test, we took each sentence in the E_{75} grammar and performed a number of transformations. We used the sentence in its original form, each sentence produced by removing one of its words (not including the period), and each sentence produced by replacing a single word with some other word. A sentence having five words would thus result in 126 derived sentences. Each derived sentence was then classified as *grammatical*, according to the E_{75} grammar, *semantically invalid*, or *syntactically invalid*. Syntactically invalid sentences are those that would not be accepted by the E_{75} grammar even if all of the semantic constraints were removed. For example, *boy chases who*

or *boy who chases cats walk*. Semantically invalid sentences, on the other hand, would be accepted by the grammar with no semantic constraints but are ruled out by the semantic constraints. For example, *boy bites dog*. Note that the invalid sentences are far from random collections of words and differ from valid sentences in only a single word. Often the invalid sentences are valid far beyond the point at which the transformation took place.

The selected network, trained only on the E_{75} corpus, was run on each of the derived sentences and the strength with which it predicted each word recorded. Fig. 4 shows the distribution of the goodness measure for sentences in each of the three categories. It is apparent that the measure does a fairly good job of pulling apart the three distributions. We can now ask how well various judgments can be made given the measure. On the standard grammaticality judgment task of distinguishing correct sentences from those with a syntactic violation, a decision criterion of −3.75 yields highly accurate performance, with only 2.21% false positives and 2.95% misses ($d' = 3.90$). In fact, the network can also distinguish, although somewhat less accurately, syntactically legal sentences with semantic violations (cf. 'Colorless green ideas....') from sentences with true syntactic violations: a decision criterion of 5.40 yields 19.6% false-alarms and 12.7% misses ($d' = 2.00$). Note that, in this latter case, the network never encountered sentences of either type during training. Also note that the syntactically invalid sentences were not simply random word jumbles but differed from a valid sentence by only a single word.

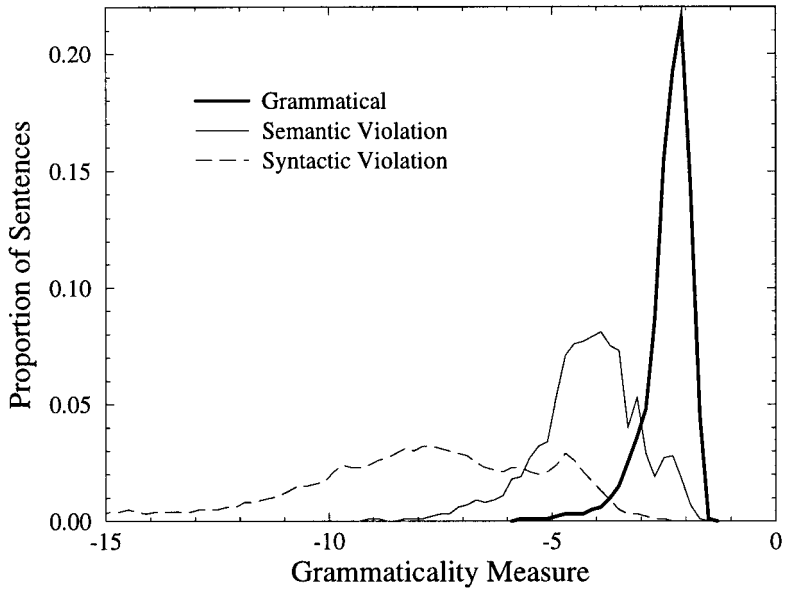

Fig. 4. Distributions of a measure of grammaticality for fully grammatical sentences, for sentences which violate semantic constraints (but obey syntactic constraints), and for sentences which violate syntactic (and semantic) constraints. The measure is the log of the product of the two worst word predictions in the sentence (see text for details).

The 'goodness' measure can also provide a basis for determining what factors influence the relative effectiveness of processing various types of valid sentences. Not surprisingly, goodness generally decreases with the number of embeddings in the sentence (means of −2.35, −2.71, −3.20, −3.72 for sentences with 1, 2, 3, or 4 embeddings, respectively; $P < 0.001$ for all pairwise comparisons). Interestingly, sentences with no embeddings produce somewhat lower values (mean −2.43) than those with one embedding ($t_{7796} = 5.51$, $P < 0.001$), but this is attributable to the unnaturally low proportion of simple sentences in the E_{75} corpus by construction (25.5% simple vs. 52.5% singly-embedded sentences). Among complex sentences, center-embedded sentences have higher goodness than purely right-branching sentences (means −2.40 vs. −2.56; $t_{6219} = 7.40$, $P < 0.001$) but, again, this is highly confounded with frequency (50.9% vs. 11.3% of sentences, respectively). Right-branching sentences have higher goodness than object-relative center-embedded sentences – a subclass with comparable frequency (10.3% of sentences, mean goodness of −2.75; $t_{2157} = 6.947$, $P < 0.001$). This latter finding is more in accord with what would be expected to hold for human subjects, but it should be kept in mind that the current corpora were not designed to match the distribution of syntactic constructions found in English.

Having provided evidence that a representative network has, in fact, learned the grammar reasonably well (although certainly not perfectly) we can return to the question of the basis for our failure to find an advantage for starting small. One possibility is that, although the network trained in the small regimen might have performed more poorly overall, it may nonetheless have learned long-distance dependencies better than when trained with the complex regimen. To test this hypothesis, we computed the total probability assigned to ungrammatical predictions (i.e. words that could not, in fact, come next in the sentence), as a function of sentence position of the predicted word (see Fig. 5). In general, fewer than eight of the 26 words were legal at any point in a sentence produced by grammar E_{75}. Overall, performance declined with word position (except for position 16 which can only be end-of-sentence). This trend is due largely to the fact that early positions are dominated by predictions within simple sentences, whereas later positions are dominated by predictions within complex sentences with multiple embeddings. Even so, 17% of the total output activation spread over 18 illegal words is respectable, considering that randomized weights produce about 71% illegal predictions. More importantly, across word positions, the complex training regimen produced better performance than the simple training regimen ($F_{1.15} = 25.7$, $P < 0.001$).

In summary, starting with simple inputs proved to be of no benefit and was actually a significant hindrance when semantic constraints applied across clauses. The networks were able to learn the grammars quite well even in the complex training regimen. Moreover, the advantage for training on the fully complex corpus increased as the language was made more English-like by enforcing greater degrees of semantic constraints. While it has been shown previously that beginning with a reduced training set can be detrimental in classification tasks such as exclusive-OR (Elman, 1993), it appears that beginning with a simplified grammar can also produce significant interference on a more language-like prediction task. At the very least,

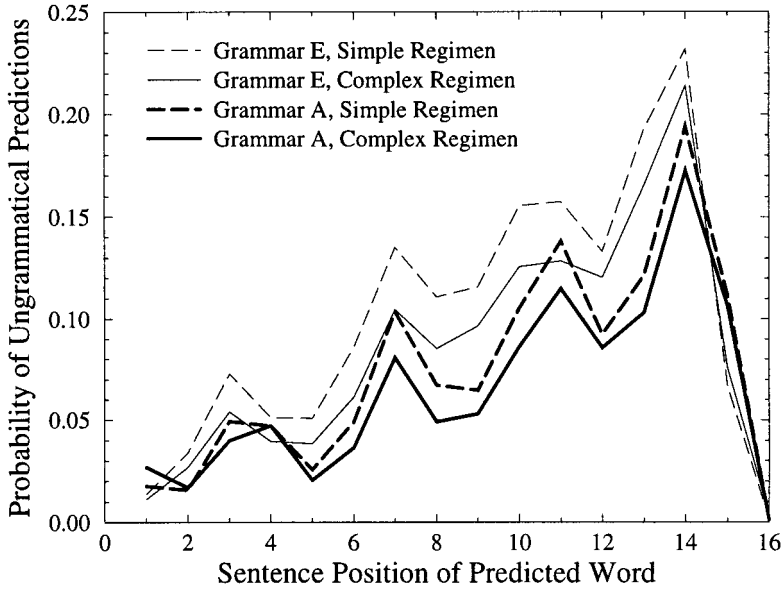

Fig. 5. Total probability assigned by the network to ungrammatical predictions, as a function of the position of the predicted word in sentences from grammars A and E, for the simple and complex training regimes. Values are averaged over all 20 networks trained in each condition.

starting small does not appear to be of general benefit in all language learning environments.

3. Simulation 2: Replication of Elman's (1993) study

Our failure to find an advantage for starting small in our initial work led us to ask what differences between that study and Elman's were responsible for the discrepant results. All of the grammars in the first set of simulations differed from Elman's grammar in that the language retained full semantic constraints within the main clause. It is possible that within-clause dependencies were in some way responsible for aiding learning in the complex training regimen. Therefore, we produced a language, labeled R for *replication*, which was identical to Elman's in all known respects, thus ruling out all but the most subtle differences in language as the source of our disparate results.

3.1. Method

Like Elman's grammar, grammar R uses just 12 verbs: two pairs each of transitive, intransitive, and mixed transitivity. In addition, as in Elman's grammar, the proper nouns *Mary* and *John* could not be modified by a relative clause and the only additional constraints involved number agreement. We should note that, although

our grammar and Elman's produce the same set of strings to the best of our knowledge, the probability distributions over the strings in the languages may differ somewhat. As before, corpora with four levels of complexity were produced. In this case they exactly matched Elman's corpora in terms of average sentence length (see Table 4).[5]

Networks were trained on this language both with our own methods and parameters and with those as close as possible to the ones Elman used. In the former case, we used normalized output units with a divergence error measure, momentum of 0.9, eleven epochs of training on the final corpus, a batch size of 10 words, a learning rate of 0.004 reduced to 0.0003 for the last epoch, and initial weights between ±1. In the latter case, we used logistic output units, squared error, no momentum, five epochs of training on the fourth corpus, online weight updating (after every word), a learning rate of 0.1 reduced to 0.06 in equal steps with each corpus change, and initial weights between ±0.001.

3.2. Results and discussion

Even when training on sentences from a grammar with no semantic constraints, our learning parameters resulted in an advantage for the complex regimen. Over the best 12 of 15 trials, the network achieved an average divergence of 0.025 under the complex condition compared with 0.036 for the simple condition ($F_{1,22} = 34.8$, $P < 0.001$). Aside from the learning parameters, one important difference between our training method and Elman's was that we added six extra epochs of training on the final corpus to both conditions. This extended training did not, however, disproportionately benefit the complex condition in some way. Between epoch 20 and 25, the average divergence error under the simple regimen dropped from 0.085 to 0.061. During the same period, the error under the complex regimen fell only from 0.051 to 0.047.[6]

It is again important to establish that the network was actually learning to perform the task well. Otherwise the apparent advantage for starting large might be an artifact of settling into local minima due to poor training methods. The best measure of network performance would appear to be a direct comparison with the results published by Elman (1991). However, as discussed earlier, Elman evaluated his network using empirically derived probabilities, rather than predictions generated directly from the grammar.

In order to approximate Elman's evaluation methods, we trained an empirical model on the R_{75} testing corpus, as well as on 240 000 additional sentences produced by the same grammar. Elman reported a final error of 0.177 for

[5]To match the average lengths of sentences generated by grammar R as closely as possible to those produced by Elman and rsquo;s grammar, the selection probabilities for intransitive verbs across the levels of complexity (0%, 25%, 50%, and 75%) were increased from 50% for each (as in grammar classes A–E) to 54%, 65%, 75%, and 50%, respectively.

[6]The further drop of these error values, 0.047 and 0.061, to the reported final values of 0.025 and 0.036 resulted from the use of a reduced learning rate for epoch 26.

his network (using, we believe, city-block distance). When trained on corpus R_{75} and evaluated against the empirical model, our network produced an average city-block distance of 0.240 (over the best 12 runs), which would seem to be considerably worse. However, as mentioned earlier, cosine is a more accurate measure of the differences between probability distributions. Our network had an average cosine of 0.942, which is considerably better than the value of 0.852 reported by Elman.

However, the empirical model itself provides a poor match to the theoretically derived predictions and, hence, is not an appropriate basis for evaluating the extent to which a network has learned the structure of a grammar. Specifically, when evaluated against the theoretical predictions, the empirical model had a mean divergence of 0.886, a city-block distance of 0.203, and a cosine of 0.947. These values are all much worse than those for the network which, when compared against the same correct predictions, produced a mean divergence of 0.025, a distance of 0.081, and a cosine of 0.991, even though it was trained on only 10 000 different sentences (cf. over 250 000 sentences for the empirical model). Thus, as far as we can tell, our network learned grammar R at least as well under the complex training regimen as Elman's network did under the simple regimen.

Because grammar R has so few constraints, it might be thought that this is a more difficult task than learning a grammar with full semantics. It is true that the problem space becomes more sparse as we add constraints, and the entropy of the optimal predictions is higher without the constraints because more alternatives are possible. However, the amount of information that must be stored to formulate an accurate prediction is much lower without semantics. Although the prediction error when measured against the actual next word is likely to be higher for the purely syntactic grammar, the error when measured against the optimal distribution is lower. This is reflected by the fact that the network achieved an average divergence error of 0.025 in this simulation versus 0.093 for the class E language with full semantic constraints in Simulation 1.

When the network was trained using parameters similar to those chosen by Elman, it failed to learn adequately, settling into bad local minima. The network consistently reached a divergence error of 1.03 under the simple training regimen and 1.20 under the complex regimen, regardless of the initial random weight values. In terms of city-block distance, these minima fall at 1.13 and 1.32, respectively – much worse than the results reported by Elman. Observation of the network in the simple condition revealed that it was able to learn only the second-order statistics of the language, and even these were not learned particularly well. The network learned that the word *who* could only follow a noun, but not that a singular head noun could never be followed by another noun or a plural verb. On the other hand, in the complex condition, the network learned only the first-order statistics, giving predictions which approximated the overall word frequencies regardless of context. Examination of the connection weights revealed that all input weights and biases to the three hidden layers had approached zero. It is not clear why we find such poor performance with what we believe to be similar training methods to those used by Elman.

We did, however, obtain successful learning by using the same parameters but simply increasing the weight initialization range from ±0.001 to ±1.0, although performance under these conditions was not quite as good as with all of our parameters and methods. Even so, we again found a significant advantage for the complex regimen over the simple regimen in terms of mean divergence error (means of 0.122 vs. 0.298, respectively; $F_{1,22} = 121.8$, $P < 0.001$).

Given that the strength of initial weights appears to be a key factor in successful learning, we conducted a few additional runs of the network to examine the role of this factor in more detail. The networks were trained on 25 epochs of exposure to corpus R_{75} under the complex regimen using parameters similar to Elman's, although with a fixed learning rate of 1.0 (i.e. without annealing). Fig. 6 shows the sum squared error on the testing corpus over the course of training. It is apparent that larger initial weights help the network break through the first-order plateau which lies at an error value of 0.221. Performance was remarkably sensitive to ranges of initial weights around ±0.1. It is interesting that the network can remain at the plateau for up to 20 epochs, processing 200 000 sentences (about 1.2 million words), before successfully breaking through.

3.3. Additional manipulations

Although we have yet to find conditions under which starting with simplified inputs aided successful learning of a simple recurrent network, there are certainly situations in which this is the case. It is possible that the simplicity of our languages

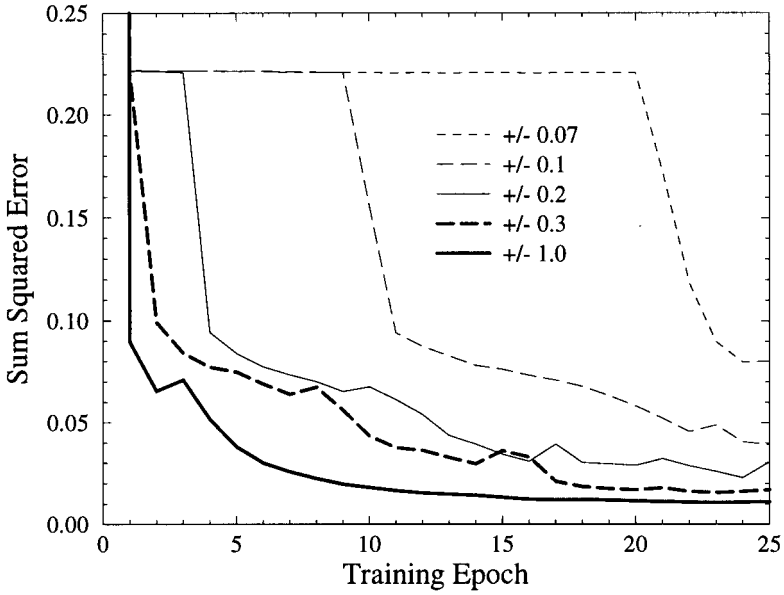

Fig. 6. Sum squared error produced by the network on the testing set at each epoch of training on corpus R_{75} under the complex regimen, as a function of the range of initial random weights.

created an unnatural advantage for the complex regimen. What, then, is required to create a task in which starting small is helpful, and are such tasks reasonable approximations of natural language processing? To answer this question, we performed two additional manipulations, one involving the removal of all constraints on embedded clauses and one extending the task to a language with 100% complex sentences.

3.3.1. Uninformative embeddings

In grammar A, as well as in Elman's grammar, verbs in subject-relative embedded clauses were constrained to agree in number with the modified noun. We might expect that this partial information was responsible for the ability of the networks trained in the complex condition to learn the noun-verb dependencies spanning the embeddings. To test this, we constructed a new grammar, A′, which was similar to A with the exception that all constraints, including number agreement, were removed on the contents of embedded clauses or between nouns and verbs within relative clauses. Full semantic and agreement constraints were left intact only within the main clause. This was done to assess the ability of the network to learn the difficult main verb prediction with no support from preceding words other than the main noun itself. As before, four versions of the grammar were produced, ranging from 0% to 75% complex. A separate testing corpus was generated from the same grammar as the last training corpus. Twenty trials each of the complex and simple conditions were performed. The same training parameters and exposures were used as in Simulation 1.

Analysis of the best 16 of 20 trials revealed an average divergence error of 0.080 in the simple regimen and 0.079 in the complex regimen ($F < 1$, n.s.). Therefore, even in the case where all constraints on the relative clauses are removed, starting small does not prove beneficial, although it is no longer a hindrance.

3.3.2. 100% complex sentences

Although Elman (1991) limited the composition of his corpora to 75% complex, his later paper (Elman, 1993) reports simulations which added a fifth corpus, consisting entirely of complex sentences. While a language composed entirely of complex sentences is not a realistic model of English, it is certainly true that the current grammars overlook many complexities of natural English. Therefore, one might view this 100% complex language as a surrogate for one in which nearly all sentences contain *some* complex grammatical structure, if not a relative clause per se.

In addition to the original four training corpora for grammatical classes E, A, and A′, a fifth, entirely complex corpus was generated for each of these classes (i.e. E_{100}, A_{100}, and A'_{100}), along with corresponding testing corpora. The same learning parameters were used as in Simulation 1. In the simple regimen, the network was trained for five epochs on each of the first four corpora and then for 10 epochs on the all-complex corpus, followed by one more epoch at the reduced learning rate of 0.0003. In the complex regimen, the network was simply trained on the fifth corpus for 30 epochs followed by one epoch at the reduced learning rate.

Despite the elimination of all simple sentences from the final corpus, the network showed no advantage for the simple regimen on grammar classes E and A. For E, the complex regimen produced an average divergence on the best 16 of 20 trials of 0.112 compared with 0.120 for the simple regimen ($F_{1.22} = 1.46$, $P > 0.2$). For A, the complex regimen yielded an error of 0.078 compared with 0.081 for simple regimen ($F_{1.22} = 1.14$, $P > 0.2$). By contrast, for class A′, in which there were absolutely no constraints except in the main clause, the simple regimen outperformed the complex regimen (means of 0.064 vs. 0.105, respectively; $F_{1.22} = 6.99$, $P < 0.05$). Therefore, starting small can be beneficial in certain circumstances. We would, however, argue that A'_{100} is not at all representative of natural language, in which relative clauses are highly dependent on what they are modifying and simple sentences are quite common.

In summary, on a grammar essentially identical to that used by Elman (1991, 1993), we found a robust advantage for training with the full complexity of the language from the outset. Although we cannot directly compare the performance of our network to that of Elman's network, it appears likely that the current network learned the task considerably better than the empirical model that we used for evaluation. By contrast, the network was unable to learn the language in either the simple or the complex condition when we used parameters similar to those employed by Elman. However, increasing the range of the initial connection weights allowed the network to learn quite well, although in this case we again found a strong advantage for starting with the full grammar. It was possible to eliminate this advantage by removing all dependencies between main clauses and their subclauses, and even to reverse it by training only on complex sentences. However, these training corpora bear far less resemblance to the actual structure of natural language than do those which produce a clear advantage for training on the full complexity of the language from the beginning.

4. Simulation 3: Progressive memory

Elman (1993) argued that his finding that initially simplified inputs were necessary for effective language learning was not directly relevant to child language acquisition because, in his view, there was little evidence that adults modify the grammatical structure of their speech when interacting with children (although we would disagree; see e.g. Sokolov, 1993; Gallaway and Richards, 1994; Snow, 1995). As an alternative, Elman suggested that the same constraint could be satisfied if the network itself, rather than the training corpus, was initially limited in its complexity. Following Newport's 'less is more' hypothesis (Newport, 1990; Goldowsky and Newport, 1993), Elman proposed that the gradual maturation of children's memory and attentional abilities could actually aid language learning. To test this proposal, Elman (1993) conducted additional simulations in which the memory of a simple recurrent network (i.e. the process of copying hidden activations onto the context units) was initially hindered and then allowed to gradually improve over the course of training. When trained on the full complexity of the grammar from the outset, but

with progressively improving memory, the network was again successful at learning the structure of the language which it had failed to learn when using fully mature memory throughout training. In this way, Elman's computational findings dovetailed perfectly with Newport's empirical findings to provide what seemed like compelling evidence for the importance of maturational constraints on language acquisition (see e.g. Elman et al., 1996 for further discussion).

Given that the primary computational support for the 'less is more' hypothesis comes from Elman's simulations with limited memory rather than those with incremental training corpora, it is important to verify that our contradictory findings of an advantage for the complex regimen in Simulations 1 and 2 also hold by comparison with training under progressively improving memory.[7] Accordingly, we conducted simulations similar to those of Elman, in which a network with gradually improving memory was trained on the full semantically constrained grammar, E, as well as on the replication grammar, R, using both Elman's and our own training parameters. As for Simulation 1, any differences between our methods and Elman's are mentioned explicitly.

4.1. Method

In his limited-memory simulation, Elman (1993) trained a network exclusively on the complex corpus, which he had previously found to be unlearnable. It is unclear from the text, however, whether he used the corpus with 75% or 100% complex sentences in this second simulation. As a model of limited memory span, the recurrent feedback provided by the context layer was eliminated periodically during processing by setting the activations at this layer to 0.5. For the first 12 epochs of training, this was done randomly after 3-4 words had been processed, without regard to sentence boundaries. For the next 5 epochs the memory window was increased to 4–5 words, then to 5–6, 6–7, and finally, in the last stage of training, the memory was not interfered with at all.

In the current simulation, the training corpus consisted of 75% complex sentences, although, as mentioned above, Elman's may have extended to 100% complexity. Like Elman, we extended the first period of training, which used a memory window of 3–4 words, from five epochs to 12 epochs. We then trained for five epochs each with windows of 4–5 and 5–7 words. The length of the final period of unrestricted memory depended on the training methods. When using our own methods (see Simulation 2), as when training on the final corpus in the simple regimen, this period consisted of 10 epochs followed by one more with the reduced learning rate. When training with our approximation of Elman's methods on grammar R, this final period was simply five epochs long. Therefore, under both conditions, the memory-limited network was allowed to train for a total of 7 epochs more than

[7]Goldowsky and Newport (1993) provide an illustration of how randomly degraded input could aid learning in a morphology-like association task. However, the results appear to depend largely on their use of a learning mechanism that collects *co-occurrence* statistics rather than perhaps more appropriate *correlations*. It is not clear whether similar results could be obtained in a mechanism attempting to learn natural language syntax.

the corresponding full-memory network in Simulations 1 and 2. When using our methods, learning rate was held fixed until the last epoch, as in Simulation 1. With Elman's method, we reduced the learning rate with each change in memory limit.

4.2. Results and discussion

Although he did not provide numerical results, Elman (1993) reported that the final performance was as good as in the prior simulation involving progressive inputs. Again, this was deemed a success relative to the complex, full-memory condition which was reportedly unable to learn the task.

Using our training methods on language R, the limited-memory condition resulted in equivalent performance to that of the full-memory condition, in terms of divergence error (means of 0.027 vs. 0.025, respectively; $F_{1.22} = 2.12$, $P > 0.15$). Limited memory did, however, provide a significant advantage over the corresponding progressive-inputs condition from Simulation 2 (mean 0.036; $F_{1.22} = 24.4$, $P < 0.001$). Similarly, for language E, the limited-memory condition was equivalent to the full-memory condition (mean of 0.093 for both; $F < 1$) but better than the progressive-inputs condition from Simulation 2 (mean of 0.115; $F_{1.22} = 31.5$, $P < 0.001$).

With Elman's training methods on grammar R, the network with limited memory consistently settled into the same local minimum, with a divergence of 1.20, as did the network with full memory (see Simulation 2). Using the same parameters but with initial connection weights in the range ±1.0, the limited-memory network again performed equivalently to the network with full memory (means of 0.130 vs. 0.122, respectively; $F_{1.22} = 2.39$, $P > 0.10$), and significantly better than the full-memory network trained with progressive inputs (mean of 0.298; $F_{1.22} = 109.1$, $P < 0.001$).

To summarize, in contrast with Elman's findings, when training on the fully complex grammar from the outset, initially limiting the memory of a simple recurrent network provided no advantage over training with full memory, despite the fact that the limited-memory regimen involved seven more epochs of exposure to the training corpus. On the other hand, in all of the successful conditions, limited memory did provide a significant advantage over gradually increasing the complexity of the training corpus.

5. General discussion

Based on the apparent lack of abundant explicit negative evidence provided to children during language learning, and the formal learnability results of Gold (1967) and others, it is often assumed that innate linguistic constraints are required for effective language acquisition. However, language learning is possible using implicit negative evidence derived from implicit predictions within a stochastic language environment. In fact, Elman (1991, 1993) demonstrated that a recurrent connectionist network could learn the structure of a pseudo-natural language based on continually predicting the next word to occur in a large corpus of sentences. Learning

was effective, however, only if either the training sentences or the network's memory were initially limited and gradually increased in complexity. Elman's findings seem to imply that standard connectionist assumptions are insufficient for language learning, and additional constraints – perhaps based on maturational factors (Newport, 1990) – must be introduced (see Elman et al., 1996 for discussion).

The first simulation of the current work demonstrated, to the contrary, that it is possible for a standard simple recurrent network to gain reasonable proficiency in a language roughly similar to that designed by Elman without staged inputs or memory. In fact, there was a significant advantage for starting with the full language, and this advantage increased as languages were made more natural by increasing the proportion of clauses which obeyed semantic constraints (see also Cleeremans et al., 1989). There may, of course, be other training methods which would yield even better performance. However, at the very least, it appears that 'starting small' is not a robust phenomenon in simple recurrent networks.

In order to identify the factors that led to the disadvantage for starting small, we returned to a more direct replication of Elman's work in Simulation 2. Using Elman's parameters, we did find what seemed to be an advantage for starting small, but the network failed to sufficiently master the task in this condition. We do not yet understand what led Elman to succeed in this condition where we failed. One observation made in the course of these simulations was that larger initial random connection weights in the network were crucial for learning. We therefore reapplied Elman's training methods but increased the range of the initial weights from ± 0.001 to 1.0. Both this condition and our own training parameters revealed a strong advantage for starting with the full language.

Finally, in Simulation 3 we examined the effect of progressive memory manipulations similar to those performed by Elman (1993). It was found that, despite increased training time, limited memory failed to provide an advantage over full memory in any condition. Interestingly, training with initially limited memory was generally less of a hindrance to learning than training with initially simplified input. In all cases, though, successful learning again required the use of sufficiently large initial weights.

The dependence of learning on the magnitudes of initial weights can be understood in light of properties of the logistic activation function, the back-propagation learning procedure, and the operation of simple recurrent networks. It is generally thought that small random weights aid error-correcting learning in connectionist networks because they put unit activations within the linear range of the logistic function where error derivatives, and hence weight changes, will be largest. However, the error derivatives that are back-propagated to hidden units are scaled by their outgoing weights; feedback to the rest of the network is effectively eliminated if these weights are too small. Moreover, with very small initial weights, the summed inputs of units in the network are all almost zero, yielding activations very close to 0.5 regardless of the input presented to the network. This is particularly problematic in a simple recurrent network because then context representations (copied from previous hidden activations) contain little if any information about previous inputs. Consequently, considerably extended training may be required to

accumulate sufficient weight changes to begin to differentiate even the simplest differences in context (see Fig. 6). By contrast, starting with relatively large initial weights not only preserves the back-propagated error derivatives but also allows each input to have a distinct and immediate impact on hidden representations and, hence, on context representations. Although the resulting patterns may not be particularly good representations for solving the task (because the weights are random), they at least provide an effective starting point for beginning to learn temporal dependencies.[8]

In the remainder of this article, we discuss other apparent demonstrations of the importance of starting small, and why recurrent networks can learn effectively without introducing this constraint. We then consider the implications of our findings for arguments concerning the use of implicit negative evidence and the need for maturational constraints on language acquisition.

5.1. Previous replications

There have been a number of informal reports of replications of Elman's basic finding of an advantage for starting small. However, a common factor in these simulations appears to be that networks trained exclusively on complex inputs were not allowed sufficient training time given the initial random weights. As we showed in Fig. 6, it is possible for a network in the complex condition to remain seemingly entrenched in a local minimum for some time before breaking through and attaining better ultimate performance than a network trained in the simple condition for an equivalent period. It may be that, in such apparent replications, networks trained in the complex condition were terminated before this breakthrough could occur.

Another problem may be that the learning parameters chosen resulted in poor overall performance for both training regimens, in which case, it would be unwise to conclude that apparent differences in performance reflect meaningful advantages for one regimen over the other. For example, Joyce (1996) claimed to have successfully replicated Elman's results, but his networks obtained a final cosine error of only 0.785 (evaluated against empirically derived probabilities), compared with values of 0.852 obtained by Elman and 0.942 obtained using our parameters in Simulation 2. In evaluating these numbers, note that assigning uniform probability across words gives a cosine of 0.419 against the empirical model from Simulation 1. Using first-order statistics (i.e. word frequencies) yields a cosine of 0.476, and using second-order statistics (i.e. including the previous word) yields a cosine of 0.780. Thus, Joyce's model is doing only about as well as the second-order statistics. The performance of Elman's network (0.852) is not quite as good as when using third-order statistics (0.873). Also note that the networks we trained with small initial weights in

[8]There is the potential complementary problem of using initial weights so large that unit activations are pinned at the extremes of the logistic function where its derivative vanishes. However, this problem is mitigated to some extent by the use of an error function like divergence that grows exponentially large as the derivative for a unit on the incorrect side of the logistic function becomes exponentially small.

Simulation 2, which clearly failed to learn the task, nevertheless obtained cosine scores of 0.604. Thus, Joyce's networks may not, in fact, have mastered the task sufficiently to make a meaningful comparison between the training regimes.

Certainly there are situations in which starting with simplified inputs is necessary for effective learning in a recurrent network. For example, Bengio et al. (1994) (see also Lin et al., 1996) report such results for tasks requiring a network to learn contingencies which span 10–60 entirely unrelated inputs. Such tasks are, however, quite unlike the learning of natural language. Similarly, in an extension of Simulation 2, we introduced a language in which absolutely no constraints existed between a noun and its relative clause. In this case, both starting small and starting large were equally effective. We also created a final corpus involving no simple sentences. At this point, we did find a significant advantage in starting small on the language with no constraints on the relative clauses. Thus, starting with simplified inputs is indeed advantageous at times, though we argue that this advantage disappears as an artificial language is made to be more like natural language.

5.2. Learning in recurrent networks

The intuition behind the importance of starting with properly chosen simplified inputs is that it helps the network to focus immediately on the more basic, local properties of the language, such as lexical syntactic categories and simple noun-verb dependencies. Once these are learned, the network can more easily progress to harder sentences and further discoveries can be based on these earlier representations.

Our simulation results indicate, however, that such external manipulation of the training corpus is unnecessary for effective language learning, given appropriate training parameters. The reason, we believe, is that recurrent connectionist networks already have an inherent tendency to extract simple regularities first. A network does not begin with fully formed representations and memory; it must learn to represent and remember useful information under the pressure of performing particular tasks, such as word prediction. As a simple recurrent network learns to represent information about an input over the hidden units, that information then becomes available as context when processing the next input. If this context provides important constraints on the prediction generated by the second input, the relevant aspects of the first input will be re-represented over the hidden units and, thus, be available as context for the third input, and so on. In this way, the network first learns short-range dependencies, starting with simple word transition probabilities for which no deeper context is needed. At this stage, the long-range constraints effectively amount to noise which is averaged out across a large number of sentences. As the short-dependencies are learned, the relevant information becomes available for learning longer-distance dependencies. Very long-distance dependencies, such as grammatical constraints across multiple embedded clauses, still present a problem for the network in any training regimen. Information must be maintained across the intervening sequence to allow the network to pick up on such a dependency. However, there must be pressure to maintain that information or the hidden representa-

tions will encode more locally relevant information. Long-distance dependencies are difficult because the network will tend to discard information about the initial cue before it becomes useful. Adding semantic dependencies to embedded clauses aids learning because the network then has an incentive to continue to represent the main noun, not just for the prediction of the main verb, but for the prediction of some of the intervening material as well (see also Cleeremans et al., 1989).[9]

It might be thought that starting with simplified inputs would facilitate the acquisition of the local dependencies so that learning could progress more rapidly and effectively to handling the longer-range dependencies. There is, however, a cost to altering the network's training environment in this way. If the network is exposed only to simplified input, it may develop representations which are overly specialized for capturing only local dependencies. It then becomes difficult for the network to restructure these representations when confronted with harder problems whose dependencies are not restricted to those in the simplified input. In essence, the network is learning in an environment with a non-stationary probability distribution over inputs. In extreme form, such non-stationarity can lead to so-called *catastrophic interference*, in which training exclusively on a new task can dramatically impair performance on a previously learned task that is similar to but inconsistent with the new task (see e.g. McCloskey and Cohen, 1989; Ratcliff, 1990; McClelland et al., 1995). A closely related phenomenon has been proposed by Marchman (1993) to account for critical period effects in the impact of early brain damage on the acquisition of English inflectional morphology. Marchman found that the longer a connectionist system was trained on the task of generating the past tense of verbs, the poorer it was at recovering from damage. This effect was explained in terms of the degree of *entrenchment* of learned representations: as representations become more committed to a particular solution within the premorbid system, they become less able to adapt to relearning a new solution after damage. More recently, McClelland (1999) and Thomas and McClelland (1997) have used entrenchment-like effects within a Kohonen network (Kohonen, 1984) to account for the apparent inability of non-native speakers of a language to acquire native-level performance in phonological skills (see e.g. Logan et al., 1991), and why only a particular type of retraining regimen may prove effective (see also Merzenich et al., 1996; Tallal et al., 1996). Thus, there are a number of demonstrations that connectionist networks may not learn as effectively when their training environment is altered significantly, as is the case in the incremental training procedure employed by Elman (1991).

Periodically disrupting a network's memory during the early stages of learning has relatively little effect because only very local information is lost, and this information would have influenced the processing of only the next word or two in

[9]It should be pointed out that this positive result applies only to the ability to *accept* a language rather than to *decide* the language. Deciding a language indicates the ability to judge, in finite time, the grammaticality of any sentence, whereas accepting a language requires only the ability to say 'yes' in finite time if a sentence is grammatical; an accepter might never respond if given an ungrammatical sentence.

any case. As the network develops in its ability to represent and use information across longer time spans, the memory is interfered with less, again leading to minimal impact on learning. Therefore, this manipulation tends neither to help nor hinder learning.

There has been much debate on the extent to which children experience syntactically simplified language (see e.g. Richards, 1994; Snow, 1994, 1995 for discussion). While child-directed speech is undoubtedly marked by characteristic prosodic patterns, there is also evidence that it tends to consist of relatively short, well-formed utterances and to have fewer complex sentences and subordinate clauses (Newport et al., 1977; see also Pine, 1994). The study by Newport and colleagues is instructive here, as it is often interpreted as providing evidence that child-directed speech is not syntactically simplified. Indeed, these researchers found no indication that mothers carefully tune their syntax to the current level of the child or that aspects of mothers' speech styles have a discernible effect on the child's learning. Nonetheless, it was clear that child-directed utterances, averaging 4.2 words, were quite unlike adult-directed utterances, averaging 11.9 words. Although child-directed speech included frequent deletions and other forms that are not handled easily by traditional transformational grammars, whether or not these serve as complexities to the child is debatable.

If children do, in fact, experience simplified syntax, it might seem as if our findings suggest that such simplifications actually impede children's language acquisition. We do not, however, believe this to be the case. We have only been considering the acquisition of syntactic structure (with some semantic constraints), which is just a small part of the overall language learning process. Among other things, the child must also learn the meanings of words, phrases, and longer utterances in the language. This process is certainly facilitated by exposing the child to simple utterances with simple, well-defined meanings. We support Newport and colleagues' conclusion that the form of child-directed speech is governed by a desire to communicate with the child and not to teach syntax. However, we would predict that language acquisition would ultimately be hindered if particular syntactic or morphological constructions were avoided altogether in child-directed speech.

To this point, our simulation results have served to broaden the applicability of connectionist networks to language acquisition by calling into question the need for additional, maturation-based constraints. In this respect, our conclusions contrast with those of Elman (1991, 1993). At a more general level, however, we are in complete agreement with Elman (and many others; see Seidenberg, 1997; Seidenberg and MacDonald, 1999) in adopting a statistical approach to language acquisition. That is, we believe that language learning depends critically on the *frequency* with which forms occur in the language and not simply on whether or not they occur at all. As discussed in the Introduction, this approach is based on assumptions about the nature of language that are fundamentally different from those traditionally adopted within linguistics. It is thus important to consider carefully the relationship between our work and alternative proposals concerning learnability and the role of negative evidence.

5.3. Learnability

At the core of the results of Gold (1967) is a proof that no interesting classes of languages are learnable from a text consisting of only valid sentences if the text is generated by the powerful class of recursive functions, which are all functions that can be computed by a Turing machine. The reason is essentially that the generating function has the power to confuse the learner indefinitely. Past experience tells the learner relatively little about the future properties of the text because at any point the text could change dramatically. Gold's result has been taken as evidence for the impossibility of language learning without stronger constraints on the learner and the class of possible languages.

However, another of Gold's results is generally ignored: If the text is generated by only a primitive recursive function, even very powerful language classes are learnable.[10] As Gold (1967) puts it, 'the primitive recursive algorithms are a special class of algorithms which are not general in the sense of Turing machines, but are general enough to include all algorithms normally constructed' (p. 474; see Hopcroft and Ullman (1979), p. 175 for a definition of *primitive recursive*). This positive result makes it clear that relaxing the strong assumption that texts are generated by fully recursive functions may alleviate the learnability problem. Along these lines, Gold (1967) suggested that learning may be possible given 'some reasonable probabilistic assumption concerning the generation of text' (p. 461).

Indeed, not long after Gold's results were published, Horning (1969) showed that stochastic context-free grammars are learnable with arbitrarily high probability from only positive examples. Angluin (1988) also showed that a fairly weak computability restriction, that the distributions used to generate the text are drawn from a 'uniformly approximately computable' sequence of distributions, allow the learnability of recursively enumerable sets (see also Osherson et al., 1986). Kapur and Bilardi (1992) proved a similar learnability result under the assumption that the learner has some rather general prior information about the input distribution. An interesting aspect of this model is that the learning is not considered to be the ability to approximate the distribution producing the text but actually learning which sentences are part of the language and which are not in the traditional sense. It is not clear whether Angluin's formalism or Kapur and Bilardi's formalism is more appropriate for the case of natural language. In some sense it is a matter of whether one views performance or competence, respectively, as primary.

One reaction to these results is to argue that a child's language experience cannot be modeled by a stochastic process. For example, Miller and Chomsky (1963) argued that k-limited Markov sources were poor language models. Note that this is precisely the same point that we have made concerning the inadequacy of using an

[10]It should be pointed out that the bias towards learning short- before long-range dependencies is not specific to simple recurrent networks; fully recurrent networks also exhibit this bias. In the latter case, learning long-range dependencies is functionally equivalent to learning an input-output relationship across a larger number of intermediate processing layers (Rumelhart et al., 1986), which is more difficult than learning across fewer layers (see Bengio et al., 1994; Lin et al., 1996).

empirical model to evaluate network performance. It is important, however, not to reject a statistical approach to language based on the inadequacy of a specific, overly simple statistical model. In fact, most empirical work on language relies on the assumption that language can be modeled as a statistical object. Whenever researchers collect a sample of language (e.g. the CHILDES database; MacWhinney and Snow, 1985; MacWhinney, 1991) and argue that the statistical properties of that sample, such as the frequency of various syntactic constructions, are in any way predictive of future samples, they are assuming that the language is generated by a process that is relatively statistically stationary. In doing so, they are, implicitly or otherwise, operating outside the scope of Gold's theorem.

In a similar vein, various proposals have been made for how the child learns language despite Gold's negative results, including acquisition rules such as the 'Uniqueness Principle', 'competition', 'preemption', 'blocking', the 'principle of contrast', 'mutual exclusivity', and the 'M-constraint' (see Wexler and Cullicover, 1980; Pinker, 1984; Bowerman, 1988; Marcus et al., 1992; MacWhinney, 1993). It is important to note that these proposals avoid Gold's problem by making a fundamental change in the assumptions of the model. All of the acquisition rules are based, in one way or another, on some form of implicit negative evidence which, in turn, depends on some degree of statistical stationarity in language. For example, suppose the child has committed a morphological overgeneralization, such as using *goed* instead of *went*. Ruling out the incorrect form based on the observation that adults do not seem to use it, or use another form in its place, is valid only if language is produced according to a reasonably stationary probability distribution over forms or sentences. One way to see this is to consider a verb like *dive* with multiple common past-tense forms (*dived* and *dove*). Marcus et al. (1992) (p. 9) argue that both past-tense forms would be treated as irregular. The problem is that the blocking principle eliminates *dived* as a past tense of *dive* if *dove* occurs first; moreover, *dived* may be withheld arbitrarily long under Gold's assumptions. If *dived* is eventually accepted as an alternative form, then by the same token, *goed* cannot be ruled out because, as far as the learner knows, *go* may be like *dive* and *goed* is just being withheld. By contrast, if the language is relatively stationary, then if the learner often hears went and never hears *goed*, it is reasonable to assume that *go* is not like *dive* and *goed* can be ruled out (or, in a probabilistic framework, made increasingly unlikely).

Thus, our suggestion that implicit negative evidence is critical to language acquisition is largely in agreement with many standard models. Indeed, prediction inherently implements a form of competition because it involves representing some alternatives at the expense of others. Where we differ is that, in our view, adequate sensitivity to the structure of language input can obviate the need for detailed innate linguistic constraints. Whether a 'uniqueness rule' must be explicitly defined as part of our innate language-acquisition constraints, or whether, as we would argue, it emerges from more general information processing mechanisms, is a matter for debate. In either case, however, we must acknowledge that we are no longer within the framework of Gold's theorem or the statistics-free assumptions of traditional approaches to linguistics.

It might be argued that our networks are not general learning mechanisms but that their success, like that of humans, is really due to innate constraints. The networks do, of course, have constraints built into them, including the number of units, the connectivity pattern, the input and output representations, the learning mechanism, the distribution of initial weights, and many other factors. Indeed, there is no such thing as a completely unbiased learning algorithm. At issue is whether the constraints needed to learn language are consistent across many forms of information processing in the brain or whether they apply only to language, and whether the constraints affect language processing very generally or whether they are specific to particular aspects of language (see also Marcus et al., 1992). Critically, none of the constraints embodied in the networks are specifically *linguistic* – given appropriate input, the identical networks could have learned to perform any of a wide range of tasks. In fact, the only critical sensitivity to parameter settings that we discovered – avoiding very small initial random weights – arises from very general characteristics of learning and processing in connectionist networks and applies equally well in non-linguistic domains.

These constraints differ markedly from the very specific rules that some proponents of innate constraints on language suggest are embedded in the genome. Such rules typically make reference to explicit syntactic and lexical abstractions assumed to be involved in language processing. As Crain notes, 'linguists generally find it reasonable to suppose that constraints are innate, domain-specific properties' (p. 598). For example, Marcus et al. (1992) propose the *blocking principle* as, 'a principle specifically governing the relations among the inflected versions of a given stem,' (p. 9) in contrast to a more general mechanism that is sensitive to the frequency with which meanings map to particular forms in the input. Along similar lines, Gropen et al. (1991) pose the universal *object affectedness linking rule*, by which, 'An argument is encodable as the direct object of a verb if its referent is specified as being affected in a specific way in the semantic representation of the verb' (p. 118), and Crain (1991) proposes a rule that contraction may not occur across a trace left behind by Wh-movement. The point here is simply to emphasize that such language-specific constraints are qualitatively distinct from the more general parameters that control, for instance, the flexibility of weights in a neural network.

5.4. Prediction as a source of negative evidence

Robust negative results like Gold's are universal in that they prove that no learning algorithm is guaranteed to succeed given the stated assumptions. By contrast, positive learnability results, such as those obtained by Horning (1969) and Angluin (1988), must be interpreted with more caution because they show only that some system can learn the task. In particular, Horning's and Angluin's methods rely on the ability of the learner to explicitly enumerate and test all possible grammars and rely on essentially unbounded resources. It seems unlikely that such assumptions hold for the language acquisition processes of the human cognitive system. The importance of these results, however, is that they demonstrate that learning is possible in the

absence of strong constraints on the language and the learner, and that a key factor in overcoming the 'logical problem' of language acquisition (Baker and McCarthy, 1981) is the use of implicit negative evidence.

In order to be relevant to human language learning, it must be possible for the limited computational mechanisms of the cognitive system to take advantage of this information. We wish to advance the hypothesis that the principal means by which the cognitive system makes use of implicit negative evidence is through the formation and evaluation of online, implicit predictions (see Jordan and Rumelhart, 1992; McClelland, 1994, for discussion). The type of predictions we are hypothesizing need not be consciously accessible, nor must predictions be over a small set of alternatives. Nor, for that matter, is prediction restricted to a probability distribution over localist lexical units, as in our network model – it is likely that linguistic predictions occur on many levels of representation, across phonemic features, across semantic and syntactic features of words, and across semantic and syntactic features of entire phrases.[11]

On our view, prediction involves the operation of standard processing mechanisms which embody the general computational principle, in interpreting linguistic utterances, of going as far beyond the literal input as possible in order to facilitate subsequent processing (see McClelland et al., 1989). A clear, if simplified, instantiation of this principle is the Cohort model of spoken word recognition (Marslen-Wilson, 1987), in which competing words are eliminated from contention as soon as information is received which is inconsistent with them. A natural (and more robust) extension of this approach would be to propose that the system maintains and updates in real time a probability distribution over words reflecting the likelihood that each word is the one being heard. Such a distribution is exactly what would emerge from attempting to predict the current word as early as possible. More generally, accurate prediction need not and should not be based on the preceding surface forms alone, as in a k-limited Markov source. In order to make accurate predictions and to generalize to novel combinations of surface forms, the system must learn to extract and represent the underlying higher-order structure of its environment.

Fodor and Crain (1987) considered the use of prediction involving syntactic structures, but argued that it is problematic on two accounts. First, they contended that 'it assumes that a learner engages in a vast amount of labor 'on the side', that he does not stop work when he has constructed a set of rules that generate all the constructions he hears and uses' (p. 51). Note, however, that learning based on prediction, on our account, is an on-line procedure that is not 'on the side' but an

[11]It might seem that prediction can operate only over localist representations, but this is not necessarily true. A prediction over distributed representations can take the form of a weighted average of the representations for individual items, with the weighting determined by the posterior probability distribution over the items. Although such a blended pattern would be quite different than the representation for any of the contributing items, it would nonetheless be closer to each of the contributing items (as a function of its weighting) than to virtually any unrelated pattern (Hinton and Shallice, 1991, Appendix 1). Such a prediction would thus provide effective context for processing subsequent input (see e.g. Kawamoto et al., 1994).

inherent part of language processing. It need not rely on memorization of entire utterances, nor on explicit compilation of frequency counts over hypothesized rules or structures, nor on discrete decisions about the grammaticality of those structures. As in the current set of simulations, feedback can be immediate, can operate on a word-by-word or more fine-grained basis, and can be incorporated in a graded fashion into the system's current, working grammar. It is true that prediction mechanisms may not stop work when one has constructed a set of rules that generate all the constructions one hears and uses, but that is a desirable feature. Algorithms that learn only from failure (e.g. Berwick, 1987) have been criticized because they fail to account for changes that are observed after children are parsing sentences competently (Bowerman, 1987). By contrast, learning via prediction applies to both successes and failures, because there are no complete successes unless the next event is predicted with absolute certainty; every prediction is likely to be approximate to some degree.

The second argument of Fodor and Crain (1987) against prediction is that the learner must know how to generalize to appropriate different constructions. This is indeed an important point. However, if predictions are generated based on the representations which form the learner's grammar, feedback will generalize to the extent that these representations generalize over structures. Functionally similar structures will receive similar feedback and will be given similar representations, allowing generalization of subsequent feedback. In contrast, similar representations for different structures are pulled apart by competing feedback. Inferring the grammar of a natural language requires the ability to form broad generalizations without sacrificing sensitivity to subtle distinctions and contradictions. This kind of processing may not be amenable to a clean description in the traditional sense, but it is what connectionist learning systems excel at.

5.5. Late exposure and second languages

The computational findings of Elman (1991, 1993) of the importance of starting small in language acquisition have been influential in part because they seemed to corroborate empirical observations that language acquisition is ultimately more successful the earlier in life it is begun (see Long, 1990). While older learners of either a first or a second language show initially faster acquisition, they tend to plateau at lower overall levels of achievement than do younger learners. The importance of early language exposure has been cited as an argument in favor of either an innate language acquisition device which operates selectively during childhood or, at least, genetically programmed maturation of the brain which facilitates language learning in childhood (Johnson and Newport, 1989; Newport, 1990; Goldowsky and Newport, 1993). It has been argued that the fact that late first- or second-language learners do not reach full fluency is strong evidence for 'maturationally scheduled *language-specific* learning abilities' (Long, 1990, p. 259, emphasis in the original).

We would argue, however, that the data regarding late language exposure can be explained by principles of learning in connectionist networks without recourse to maturational changes or innate devices. Specifically, adult learners may not nor-

mally achieve fluency in a second language because their internal representations have been largely committed to solving other problems – including, in particular, comprehension and production of their native language (see Flege, 1992; Flege et al., 1995). By contrast, the child ultimately achieves a higher level of performance because his or her resources are initially uncommitted. This idea, which accords with the theory of Quartz and Sejnowski (1997) of neural constructivism, is certainly not a new one, but is one that seems to remain largely ignored (although see Marchman, 1993; McClelland, 1999). On this view, it seems unlikely that limitations in a child's cognitive abilities are of significant benefit in language acquisition. While adults' greater memory and analytical abilities lead to faster initial learning, these properties need not be responsible for the lower asymptotic level of performance achieved, relative to children.

Along similar lines, the detrimental impact of delayed acquisition of a first language may not implicate a language-specific system that has shut down. Rather, it may be that, in the absence of linguistic input, those areas of the brain which normally become involved in language may have been recruited to perform other functions (see e.g. Merzenich and Jenkins, 1995 for relevant evidence and discussion). While it is still sensible to refer to a critical or sensitive period for the acquisition of language, in the sense that it is important to start learning early, the existence of a critical period need not connote specific language-acquisition devices or genetically prescribed maturational schedules.

Indeed, similar critical periods exist for learning to play tennis or a musical instrument. Rarely if ever does an individual attain masterful abilities at either of these pursuits unless they begin at an early age. And certainly in the case of learning the piano or violin, remarkable abilities can be achieved by late childhood and are thus not simply the result of the many years of practice afforded to those who start early. One might add that no species other than humans is capable of learning tennis or the violin. Nevertheless, we would not suppose that learning these abilities depends upon domain-specific innate mechanisms or constraints.

While general connectionist principles may explain the overall pattern of results in late language learning, considerable work is still needed to demonstrate that this approach is sufficient to explain the range of relevant detailed findings. For example, it appears that vocabulary is more easily acquired than morphology or syntax, and that second language learners have variable success in mastering different syntactic rules (Johnson and Newport, 1989). In future work, we intend to develop simulations that include comprehension and production of more naturalistic languages, in order to extend our approach to address the empirical issues in late second-language learning and to allow us to model a wider range of aspects of language acquisition more directly.

6. Conclusion

If we accept the assumptions of Gold's model (1967), his theorems seem to imply that natural language should not be learnable. Although explicit negative evidence

may sometimes be available to the child in a form that is successfully utilized, such feedback appears insufficient by itself to overcome Gold's problem. There would thus appear to be two remaining viable solutions, which both involve altering the assumptions of the model: either natural languages are drawn from a highly restricted set and the properties of the possible natural languages are encoded genetically, or there is a restriction on the set of possible texts – in particular, to those that are produced according to reasonably stable probability distributions.

In their most extreme forms, these solutions accord either with the hypothesis that language is learned by a highly constrained mechanism with little reliance on distributional properties of the input, or with the hypothesis that language is learnable by a relatively general mechanism that relies heavily on statistical evidence in the input. We believe that the latter hypothesis is preferable as a starting point in that it embodies weaker initial assumptions, and that its investigation will lead more quickly to an understanding of cognition and the learning mechanisms of the brain more generally. We have already seen that reliance on implicit negative evidence is difficult to avoid in either framework, thus bringing them perhaps that much closer.

Adopting a statistical learning approach raises the issue of how a cognitively and neurally plausible mechanism might actually acquire the relevant knowledge from appropriately structured linguistic input. Following Elman (1991, 1993), we have shown that simple recurrent connectionist networks can learn the structure of pseudo-natural languages based on implicit negative evidence derived from performing a word prediction task in a stochastic environment. Unlike Elman, however, we found that learning was most effective when the network was exposed to the full complexity of the language throughout training, and that the advantage of this approach over 'starting small' increased as the language was made more English-like by introducing semantic constraints.

One major limitation of the task in our simulations is that the networks are not actually comprehending, only learning the syntax of the language. As such, there is no context or meaning to the utterances. This is not representative of what is required for language acquisition, but it may actually make the subtask of learning the grammatical structure of the language more difficult. Because context, whether it is visual or verbal, greatly constrains the set of likely utterances, its addition could significantly facilitate learning of the grammar. Without context, it is difficult to determine whether prediction errors are due to inadequate syntactic knowledge or inadequate semantic knowledge. Familiar contexts clarify the intended semantics, helping the system overcome this bootstrapping problem. We leave it to future research to determine whether the simulation results we have obtained with a mostly syntactic prediction task generalize to more natural comprehension tasks and more realistic languages.

Despite their simplicity, our simulations call into question the proposal that limited cognitive resources are necessary, or even beneficial, for language acquisition. However, perhaps the most important aspect of Elman's work is reinforced by ours – that connectionist systems can learn the structure of a language in the absence of explicit negative evidence. We claim that prediction is the principal mechanism by

which the human cognitive system is able to take advantage of implicit negative evidence. Our work suggests that learning the structure of natural language may be possible despite a lack of explicit negative feedback, despite experiencing unsimplified grammatical structures, and in the absence of detailed, innate language-acquisition mechanisms.

Acknowledgements

This research was supported by NIMH Program Project Grant MH47566 (J. McClelland, PI), and by an NSF Graduate Fellowship to D.L.T. Rohde. Preliminary versions of some of the simulations in the current article were reported in Rohde and Plaut (1997). We thank Marlene Behrmann, Jeff Elman, Brian MacWhinney, Jay McClelland, the CMU PDP research group, and two anonymous reviewers for helpful comments and/or discussions.

References

Allen, J., Seidenberg, M.S., 1999. The emergence of grammaticality in connectionist networks. In: MacWhinney, B. (Ed.), Emergentist approaches to language, Proceedings of the 28th Carnegie Symposium on Cognition. Lawrence Erlbaum Associates, Hillsdale, NJ.

Angluin, D., 1988. Identifying languages from stochastic examples (Tech. Rep. YALEU/DCS/RR-614). Yale University, Department of Computer Science, New Haven, CT.

Baker, C.L., 1979. Syntactic theory and the projection problem. Linguistic Inquiry 10, 533–581.

Baker, C.L., McCarthy, J.J., 1981. The Logical Problem of Language Acquisition. The MIT Press, Cambridge, MA.

Bengio, Y., Simard, P., Frasconi, P., 1994. Learning long-term dependencies with gradient descent is difficult. IEEE Transactions on Neural Networks 5, 157–166.

Berwick, R.C., 1985. The acquisition of syntactic knowledge. The MIT Press, Cambridge, MA.

Berwick, R.C., 1987. Parsability and learnability. In: MacWhinney, B. (Ed.), Mechanisms of Language Acquisition. Lawrence Erlbaum Associates, Hillsdale, NJ, pp. 345–366.

Bowerman, M., 1987. Commentary: mechanisms of language acquisition. In: MacWhinney, B. (Ed.), Mechanisms of Language Acquisition. Lawrence Erlbaum Associates, Hillsdale, NJ, pp. 443–466.

Bowerman, M., 1988. The 'no negative evidence' problem: How do children avoid constructing an overly general grammar? In: Hawkins, J. (Ed.), Explaining Language Universals. Basil Blackwell, Oxford, pp. 73–101.

Chomsky, N., 1957. Syntactic structures. Mouton, The Hague.

Chomsky, N., 1981. Lectures on government and binding. Foris Publications, Dordrecht, Holland.

Cleeremans, A., Servan-Schreiber, D., McClelland, J., 1989. Finite state automata and simple recurrent networks. Neural Computation 1, 372–381.

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

Elman, J.L., 1990. Finding structure in time. Cognitive Science 14, 179–211.

Elman, J.L., 1991. Distributed representations, simple recurrent networks, and grammatical structure. Machine Learning 7, 195–225.

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

Elman, J.L., Bates, E.A., Johnson, M.H., Karmiloff-Smith, A., Parisi, D., Plunkett, K., 1996. Rethinking innateness: a connectionist perspective on development. MIT Press, Cambridge, MA.

Flege, J.E., 1992. Speech learning in a second language. In: Ferguson, C.A., Menn, L., Stoel-Gammon, C. (Eds.), Phonological Development: Models, Research, Implications. York Press, Timonium, MD, pp. 565–604.

Flege, J.E., Munro, M.J., MacKay, I.R.A., 1995. Factors affecting strength of perceived foreign accent in a second language. Journal of the Acoustical Society of America 97, 3125–3134.

Fodor, J.D., Crain, S., 1987. Simplicity and generality of rules in language acquisition. In: MacWhinney, B. (Ed.) Mechanisms of Language Acquisition. Lawrence Erlbaum Associates, Hillsdale, NJ, pp. 35–63.

Gallaway, C., Richards, B. (Eds.), 1994. Input and Interaction in Language Acquisition. Cambridge University Press, London.

Gold, E.M., 1967. Language identification in the limit. Information and Control 10, 447–474.

Goldowsky, B.N., Newport, E.J., 1993. Modeling the effects of processing limitations on the acquisition of morphology: the less is more hypothesis. In: E. Clark (Ed.), The Proceedings of the 24th Annual Child Language Research Forum. Center for the Study of Language and Information, Stanford, CA, pp. 124–138.

Gropen, J., Pinker, S., Hollander, M., Goldberg, R., 1991. Syntax and semantics in the acquisition of locative verbs. Journal of Child Language 18, 115–151.

Hinton, G.E., Shallice, T., 1991. Lesioning an attractor network: investigations of acquired dyslexia. Psychological Review 98, 74–95.

Hopcroft, J.E., Ullman, J.D., 1979. Introduction to Automata Theory, Languages, and Computation. Addison-Wesley, Reading, MA.

Horning, J.J., 1969. A study of grammatical inference. PhD thesis, Stanford University.

Huang, X.D., Ariki, Y., Jack, M.A., 1990. Hidden Markov Models for speech recognition. Edinburgh University Press, Edinburgh.

Johnson, J.S., Newport, E.J., 1989. Critical period effects in second language learning: the influence of maturational state on the acquisition of English as a second language. Cognitive Psychology 21, 60–99.

Jordan, M.I.., 1992. Constrained supervised learning. Journal of Mathematical Psychology 36, 396–425.

Jordan, M.I., Rumelhart, R.A., 1992. Forward models: supervised learning with a distal teacher. Cognitive Science 16, 307–354.

Joyce, J., 1996. When/why/of what is less more? Masters thesis, Centre for Cognitive Science, University of Edinburgh.

Kapur, S., Bilardi, G., 1992. Language learning from stochastic input. In Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory. Pittsburgh, PA, pp. 303–310.

Kawamoto, A.H., Farrar, W.T., Kello, C.T., 1994. When two meanings are better than one: modeling the ambiguity advantage using a recurrent distributed network. Journal of Experimental Psychology: Human Perception and Performance 20, 1233-1247.

Kohonen, T., 1984. Self-Organization and Associative Memory. Springer, New York.

Lin, T., Horne, B.G., Giles, C.L., 1996. How embedded memory in recurrent neural network architectures helps learning long-term temporal dependencies (Tech. Rep. CS-TR-3626, UMIACS-TR-96-28). University of Maryland, College Park, MD.

Logan, S.E., Lively, J.S., Pisoni, D.B., 1991. Training Japanese listeners to identify English /r/ and /l/: a first report. Journal of the Acoustical Society of America 89, 874–886.

Long, M.H., 1990. Maturational constraints on language development. Studies in Second Language Acquisition 12, 251–285.

Luce, D.R., 1986. Response Times. Oxford University Press, New York.

MacWhinney, B., 1991. The CHILDES Database. Discovery Systems, Dublin, OH.

MacWhinney, B., Snow, C., 1985. The Child Language Data Exchange System. Journal of Child Language 12, 271–295.

MacWhinney, B., 1993. The (il)logical problem of language acquisition. Proceedings of the 15th Annual Conference of the Cognitive Science Society. Lawrence Erlbaum Associates, Mahwah, NJ, pp. 61–70.

Marchman, V.A., 1993. Constraints on plasticity in a connectionist model of the English past tense. Journal of Cognitive Neuroscience 5, 215–234.

Marcus, G.F., 1993. Negative evidence in language acquisition. Cognition 46, 53–85.

Marcus, G.F., Pinker, S., Ullman, M., Hollander, M., Rosen, T.J., Xu, F., 1992. Overregularization in Language Acquisition. Monographs of the Society for Research in Child Development 57(4), Serial No. 228.

Marslen-Wilson, W., 1987. Functional parallelism in spoken word-recognition. Cognition 25, 71–102.

McClelland, J.L., 1999. Failures to learn and their remediation: a competitive, Hebbian approach. In: McClelland, J.L., Siegler, R.S. (Eds.), Mechanisms of Cognitive Development: Behavioral and Neural Perspectives. Erlbaum, Mahwah, NJ, in press.

McClelland, J.L., 1994. The interaction of nature and nurture in development: a parallel distributed processing perspective. In: Bertelson, P., Eelen, P., d'Ydewalle, G. (Eds.), International Perspectives on Psychological Science, Vol. 1: Leading Themes. Erlbaum, Hillsdale, NJ, pp. 57–88.

McClelland, J.L., McNaughton, B.L., O'Reilly, R.C., 1995. Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. Psychological Review 102, 419–457.

McClelland, J.L., St. John, M., Taraban, R., 1989. Sentence comprehension: a parallel distributed processing approach. Language and Cognitive Processes 4, 287–335.

McCloskey, M., Cohen, N.J., 1989. Catastrophic interference in connectionist networks: the sequential learning problem. In: Bower, G.H. (Ed.), The Psychology of Learning and Motivation. Academic Press, New York, pp. 109–165.

Merzenich, M.M., Jenkins, W.M., 1995. Cortical plasticity, learning and learning dysfunction. In: Julesz, B., Kovacs, I. (Eds.), Maturational Windows and Adult Cortical Plasticity. Addison-Wesley, Reading, MA, pp. 247–272.

Merzenich, M.M., Jenkins, W.M., Johnson, P., Schreiner, C., Miller, S.L., Tallal, P., 1996. Temporal processing deficits of language-learning impaired children ameliorated by training. Science 271, 77–81.

Miller, G.A., Chomsky, N., 1963. Finitary models of language users. In: Luce, R.D., Bush, R.B., Galanter, E. (Eds.), Handbook of Mathematical Psychology, Vol. II. Wiley, New York, pp. 419–491.

Morgan, J.L., Bonamo, K.M., Travis, L.L., 1995. Negative evidence on negative evidence. Developmental Psychology 31, 180–197.

Morgan, J.L., Travis, L.L., 1989. Limits on negative information in language input. Journal of Child Language 16, 531–552.

Newport, E.L., 1990. Maturational constraints on language learning. Cognitive Science 14, 11–28.

Newport, E.L., Gleitman H., Gleitman, L.R., 1977. Mother, I'd rather do it myself: some effects and noneffects of maternal speech style. In: Snow, C.E., Ferguson, C.A. (Eds.), Talking to Children: Language Input and Acquisition. Cambridge University Press, Cambridge, UK, pp. 109–149.

Osherson, D., Stob, M., Weinstein, S., 1986. Systems That Learn. MIT Press, Cambridge, MA.

Pine, J.M., 1994. The language of primary caregivers. In: Gallaway, C., Richards, B.J. (Eds.), Input and Interaction in Language Acquisition. Cambridge University Press, Cambridge, UK, pp. 38–55.

Pinker, S., 1984. Language Learnability and Language Development. Harvard University Press, Cambridge, MA.

Quartz, S.R., Sejnowski, T.J., 1997. The neural basis of cognitive development: a constructivist manifesto. Behavioral and Brain Sciences 20, 537–596.

Ratcliff, R., 1990. Connectionist models of recognition memory: constraints imposed by learning and forgetting functions. Psychological Review 97, 285–308.

Richards, B.J., 1994. Child-directed speech and influences on language acquisition: methodology and interpretation. In: Gallaway, C., Richards, B.J. (Eds.), Input and Interaction in Language Acquisition. Cambridge University Press, Cambridge, UK, pp. 74–106.

Rohde, D.L.T., 1999. The Simple Language Generator: Encoding complex languages with simple grammars (Tech. Rep. CMU-CS-99-123). Carnegie Mellon University, Pittsburgh, PA.

Rohde, D.L.T., Plaut, D.C., 1997. Simple recurrent networks and natural language: how important is starting small? In Proceedings of the 19th Annual Conference of the Cognitive Science Society. Erlbaum, Hillsdale, NJ.

Rumelhart, D.E., Hinton, G.E., Williams, R.J., 1986. Learning internal representations by error propagation. In: Rumelhart, D.E., McClelland, J.L. and the PDP Research Group (Eds.), Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1: Foundations. MIT Press, Cambridge, MA, pp. 318–362.

Rumelhart, D.E., Durbin, R., Golden, R., Chauvin, Y., 1995. Backpropagation: The basic theory. In: Chauvin, Y., Rumelhart, D.E. (Eds.), Back-Propagation: Theory, Architectures, and Applications. Erlbaum, Hillsdale, NJ, pp. 1–34.

Seidenberg, M.S., 1997. Language acquisition and use: learning and applying probabilistic constraints. Science 275, 1599–1603.

Seidenberg, M.S., MacDonald, M.C., 1999. A probabilistic constraints approach to language acquisition and processing. Cognitive Science, in press.

Snow, C.E., 1994. Beginning from baby talk: twenty years of research on input and interaction. In: Gallaway, C., Richards, B.J. (Eds.), Input and Interaction in Language Acquisition. Cambridge University Press, Cambridge, UK, pp. 3–12.

Snow, C.E., 1995. Issues in the study of input: finetuning, universality, individual and developmental differences, and necessary causes. In: Fletcher, P., MacWhinney, B. (Eds.), The Handbook of Child Language. Blackwell, Oxford.

Snow, C.E., Ferguson, C.A. (Eds.), Talking to Children: Language Input and Acquisition. Cambridge University Press, Cambridge, UK.

Sokolov, J.L., 1993. A local contingency analysis of the fine-tuning hypothesis. Developmental Psychology 29, 1008–1023.

Sokolov, J.L., Snow, C.E., 1994. The changing role of negative evidence in theories of language development. In: Gallaway, C., Richards, B.J. (Eds.), Input and Interaction in Language Acquisition. Cambridge University Press, Cambridge, pp. 38–55.

Tallal, P., Miller, S.L., Bedi, G., Byma, G., Wang, X., Nagaraja, S.S., Schreiner, C., Jenkins, W.M., Merzenich, M.M., 1996. Language comprehension in language-learning impaired children improved with acoustically modified speech. Science 271, 81–84.

Thomas, A., McClelland, J.L., 1997. How plasticity can prevent adaptation: Induction and remediation of perceptual consequences of early experience (Abstract 97.2). Society for Neuroscience Abstracts 23, 234.

Wexler, K., Cullicover, P., 1980. Formal Principles of Language Acquisition. MIT Press, Cambridge, MA.