Less is Less in Language Acquisition

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

Carnegie Mellon University and the Center for the Neural Basis of Cognition

February 2002

To appear in Quinlin, P. (Ed.) (in press) *Connectionist modelling of cognitive development.* Hove, UK: Psychology Press.

1 Introduction

A principal observation in the study of language acquisition is that people exposed to a language as children are more likely to achieve fluency in that language than those first exposed to it as adults, giving rise to the popular notion of a critical period for language learning (Lenneberg, 1967; Long, 1990). This is perhaps surprising since children have been found to be inferior to adults in most tests of other cognitive abilities.

A variety of explanations have been put forth to account for the benefit of early language learning. Possibly the most prevalent view is that children possess a specific "language acquisition device" that is programmatically deactivated prior to or during adolescence (Chomsky, 1965; McNeill, 1970). Important to this view is that knowledge or processes necessary for effective language learning are only available for a limited period of time. But this theory has trouble accounting for continued effects of age-of-acquisition after adolescence (Bialystok & Hakuta, 1999) and evidence that some adult second language learners are still able to reach fluency (see Birdsong, 1999).

An alternative account is provided by Newport's (1990) "less-is-more" hypothesis. Rather than attributing the early language advantage to a specific language learning device, this theory postulates that children's language acquisition may be aided rather than hindered by their limited cognitive resources. According to this view, the ability to learn a language declines over time as a result of an *increase* in cognitive abilities. The reasoning behind this suggestion is that a child's limited perception and memory may force the child to focus on smaller linguistic units which form the fundamental components of language, as opposed to memorizing larger units which are less amenable to recombination. While this is an attractive explanation, for such a theory to be plausible, the potential benefit of limited resources must be demonstrated both computationally and empirically.

The strongest evidence for Newport's theory comes from computational simulations and empirical findings of Elman (1991, 1993), Goldowsky and Newport (1993), Kareev, Lieberman, and Lev (1997), Cochran, McDonald, and Parault (1999), and Kersten and Earles (2001). In the current chapter, we consider these studies in detail and, in each case, find serious cause to doubt their intended support for the less-is-more hypothesis.

- Elman (1991, 1993) found that simple recurrent connectionist networks could learn the structure of an English-like artificial grammar only when "starting small"—when either the training corpus or the network's memory was limited initially and only gradually made more sophisticated. We show, to the contrary, that language learning by recurrent networks does not depend on starting small; in fact, such restrictions hinder acquisition as the languages are made more realistic by introducing graded semantic constraints (Rohde & Plaut, 1999).
- We discuss the simple learning task introduced by Goldowsky and Newport (1993) as a clear demonstration of the advantage of memory limitations. But we show that their filtering mechanism actually constitutes a severe impairment to learning in both a simple statistical model and a neural network model.
- Kareev, Lieberman, and Lev (1997) argued that small sample sizes, possibly resulting from weak short-term memory, have the effect of enhancing correlations between two observable variables. But we demonstrate that the chance that a learner is able to detect a correlation actually improves with sample size and that a simple prediction model indeed performs better when it relies on larger samples.
- Cochran, McDonald, and Parault (1999) taught participants ASL verbs with and without additional cognitive loads and found apparently better generalization performance for participants in the load condition. But we argue that the learning task actually provided no support for the expected generalization and that the no-load participants simply learned the more reasonable generalization much better.
- Finally, we consider the Kersten and Earles (2001) findings to provide little support for the less-is-more

hypothesis because the task learned by participants in their experiment is unlike natural language learning in some important and relevant aspects and the critical manipulation in their experiment involved staged input, rather than cognitive limitations.

In the final section, we consider some general principles of learning language-like tasks in recurrent neural networks and what the implications for human learning might be. We then briefly discuss an alternative account for the language-learning superiority of children.

2 Elman (1991, 1993)

Elman (1990, 1991) set out to provide an explicit formulation of how a general connectionist system might learn the grammatical structure of a language. Rather than comprehension or overt parsing, Elman chose to train the networks to perform word prediction. Although word prediction is a far cry from language comprehension, it can be viewed as a useful component of language processing, given that the network can make accurate predictions only by learning the structure of the grammar. Elman trained a simple recurrent network—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. He 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) and only later exposed to an increasing proportion of complex sentences.

It thus seems reasonable to conclude that staged input enabled 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. Recognizing the parallel between this finding and the less-is-more hypothesis, Elman (1993) decided to investigate a more direct test of Newport's (1990) theory. Rather than staging the input presentation, Elman initially interfered with the network's memory span and then allowed it to gradually improve. Again, he found successful learning in this memory limited condition, providing much stronger support for the hypothesis.

2.1 Rohde and Plaut (1999) Simulation 1: Progressive Input

Rohde and Plaut (1999) investigated how the need for starting small in learning a pseudo-natural language would be affected 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, Servan-Schreiber, and McClelland (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 words 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 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. The methods we used were very similar, but not identical, to those used by Elman (1991, 1993).

2.1.1 Grammar

Our pseudo-natural language was based on the grammar shown in Table 1, which generates simple noun-verb and noun-verb-noun sentences with the possibility of relative clause modification of most nouns. Relative clauses could be either subject-extracted or object-extracted. Although this language is quite simple, in comparison to natural language, it is nonetheless of interest because, in order to make accurate predictions, a network must learn to form representations of potentially complex syntactic structures and remember information, such as whether the subject was singular or plural, over lengthy embeddings. The

Table 1: The Grammar Used in Simulation 1

S	→	NP VI . | NP VT NP .
NP	→	N | N RC
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

Note: Transition probabilities are specified and additional constraints are applied on top of this framework.

Table 2: Semantic Constraints on Verb Usage

	Intransitive	Transitive	Objects
Verb	Subjects	Subjects	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	–	–

Note: Columns indicate legal subject nouns when verbs are used intransitively or transitively and legal object nouns when transitive.

grammar used by Elman was nearly identical, 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 our simulation, 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 (see Table 2). Another restriction in the language was that proper nouns could not act on themselves. Finally, constructions which repeat an intransitive verb, such as Boys who walk walk, were disallowed because of redundancy. These so-called *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.

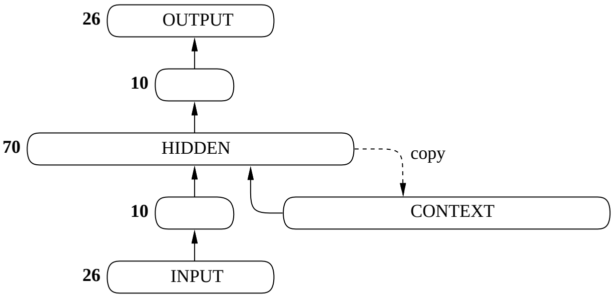

Figure 1: The architecture of the network used in the simulations. Each solid arrow represents full connectivity between layers, with numbers of units next to each layer. Hidden unit states are copied to corresponding context units (dashed arrow) after each word is processed.

2.1.2 Network Architecture

The simple recurrent network used in both Elman's simulations and in the current work is shown in Figure 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.

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 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, nonlinear (logistic, or sigmoid) function of its summed weighted input from other units. The resulting activations over the output units were then compared with their target activations, generating an error signal. In a simple recurrent network, errors are not back-propagated through time (cf. Rumelhart, Hinton, & Williams, 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 prior hidden representation copied onto the context units—to influence current performance.

Although the target output used during training was the encoding for the actual next word, a number of words were typically 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 generally result in the lowest performance error.

2.1.3 Corpora

Elman's *complex* training regimen involved training a network on a corpus of 10,000 sentences, 75% of which were "complex" in that they contained at least one relative clause. In his *simple* regimen, 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—identical to 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 a clause and its subclause, only between nouns and verbs explicitly present in each individual clause. In class B, 25% of the subclauses respected the semantic constraints of their parent clause. In such cases, the modified noun must be a semantically valid subject of the verb for a subject-relative or object of the verb for an object-relative. In class C, 50% of the subclauses respected this constraint, 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 a second simulation, we 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% respectively had 0%, 10%, 20%, and 30% modification probability for each noun.

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, a network of this size couldn't possibly "memorize" the training corpus but must learn the structure of the language.

2.1.4 Training and Testing Procedures

In the condition Elman referred to as "starting small," he trained his network for 5 epochs (complete presentations) of 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 prediction 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 Figure 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 six extra 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 the performance level possible under optimal conditions, we searched a wide range of training parameters to determine a set which consistently achieved the best performance overall.[1] 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

[1] The effects of changes to some of these parameter values—in particular, the magnitude of initial random weights—are evaluated in a second simulation.

in terms of *divergence* and network outputs were normalized using Luce ratios (Luce, 1986), also known as *softmax* constraints (see Rohde & Plaut, 1999).

Because our grammars were in standard stochastic, context-free form, it was possible to evaluate the network by comparing its predictions to the theoretically correct next-word distributions given the sentence context (Rohde, 1999). By contrast, it was not possible to generate such optimal predictions based on Elman's grammar. In order to form an approximation to optimal 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.

2.1.5 Results and Discussion

Elman did not provide numerical results for the *complex* condition, but he did report 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 mean cosine[2] of the angle between the network's prediction vectors and those of the empirical model was 0.852 (*SD* = 0.259), where 1.0 is optimal.

Figure 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 ($F_{1,150} = 53.8$, $p < .001$). 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, when the 16 simple and 16 complex training regimen networks for each grammar were paired with one another in order of increasing overall performance, there was a strong positive correlation ($r = .76$, $p < .001$) between the order of the grammars from A–E and the difference in error between the simple versus complex training regimes.[3] 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.

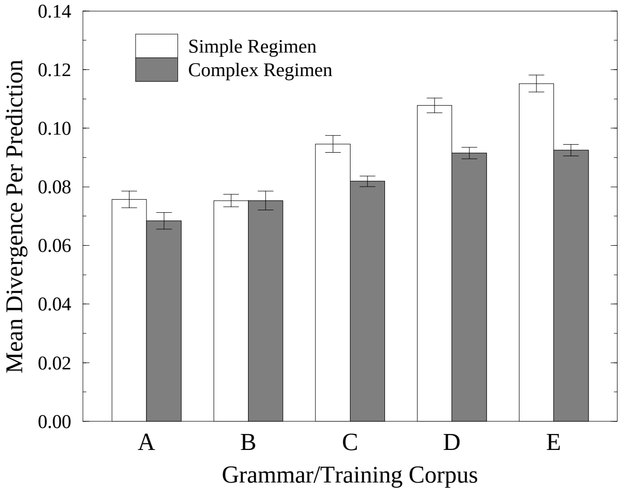

Figure 2: Mean divergence per word prediction over the 75% complex testing corpora generated from grammar classes A through E (increasing in the extent of semantic constraints) for the simple and complex training regimes. Note that lower values correspond to better performance. Means and standard errors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

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, as evidenced by additional analyses reported in Rohde and Plaut (1999). 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, starting small does not appear to be of general benefit in all language learning environments.

2.2 Rohde and Plaut (1999) Simulation 2: Replication of Elman (1993)

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

[2] The cosine of the angle between two vectors of equal dimensionality can be computed as the dot product (or sum of the pairwise products of the vector elements) divided by the product of the lengths of the two vectors.

[3] The correlation with grammar class is also significant ($r = .65$, $p <$ .001) when using the *ratio* of the simple to complex regimen error rates for each pair of networks, rather than their difference.

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 potential source of our disparate results.

2.2.1 Methods

Like Elman's grammar, grammar R uses just 12 verbs: 2 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 very closely matched Elman's corpora in terms of average sentence length.

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.

2.2.2 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, $p <$ .001). Aside from the learning parameters, one important difference between our training method and Elman's was that we added 6 extra epochs of training on the final corpus to both conditions. This extended training did not, however, disproportionately benefit the complex condition. Between epoch 20 and 25, the average divergence error under the simple regimen dropped from 0.085 to 0.061, or 28%. During the same period, the error under the complex regimen only fell 8%, from 0.051 to 0.047.[4]

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. In terms of city-block distance, these minima fall at 1.13 and 1.32 respectively—much worse than the results reported 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, $p <$ .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). Figure 3 shows the sum squared error on the testing corpus over the course of training, as a function of the range of the initial random weights. It is apparent that larger initial weights help the network break through the plateau which lies at an error value of 0.221.

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 place 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 it leads to context representations (copied from previous hidden activations) that contain little if any usable information about previous inputs. Consequently, considerably extended

[4] 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. Ending with a bit of training with a very low learning rate is particularly useful when doing online, or small batch size, learning.

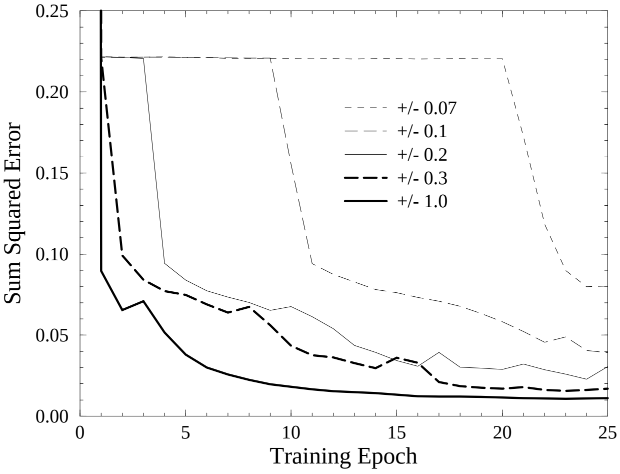

Figure 3: Sum squared error produced by the network on the testing set at each epoch of training on corpus R_{75} under the complex regimen, as a function of the range of initial random weights.

training may be required to accumulate sufficient weight changes to begin to differentiate even the simplest differences in context (see Figure 3). 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.

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, in addition, training exclusively on complex sentences. But 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.

2.3 Rohde and Plaut (1999) 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., Gallaway & Richards, 1994; Snow, 1995; Sokolov, 1993). 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 & 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. 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.

2.3.1 Methods

In his limited-memory simulation, Elman (1993) trained a network exclusively on the complex corpus,[5] which he had previously found to be unlearnable. As a model of

[5] It is unclear from the text whether Elman (1993) used the corpus with 75% or 100% complex sentences in the progressive memory experiments.

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 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 5 epochs to 12 epochs. We then trained for 5 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 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.

2.3.2 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, $p > .15$). 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, $p < .001$). 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, $p < .001$).

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 almost 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, $p < .001$).

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 7 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.

2.4 Summary

Contrary to the results of Elman (1991, 1993), Rohde and Plaut (1999) found 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. There may, of course, be other training methods which would yield even better performance. However, at the very least, it appears that the advantage of staged input 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.

Certainly there are situations in which starting with simplified inputs is necessary for effective learning of a prediction task by a recurrent network. For example, Bengio, Simard, and Frasconi (1994) (see also Lin, Horne, & Giles, 1996) report such results for tasks requiring a network to learn contingencies which span 10–60 entirely unrelated inputs. However, such tasks are quite unlike the learning of natural language. It may also be possible that starting with a high proportion of simple sentences is of significant benefit in learning other language processing tasks, such as comprehension. A child's discovery of the mapping between form and meaning will likely be facilitated if he or she experiences propositionally simple utterances whose meaning is apparent or is clarified by the accompanying actions of the parent. However, the real question in addressing the less-is-more hypothesis is whether limited cognitive capacity will substantially aid this process.

Having failed to replicate Elman's results, it seems appropriate to turn a critical eye on the other major sources of evidence for the less-is-more hypothesis. Aside from Elman's findings, four main studies have been characterized as providing support for the advantage of learning with limited resources. Goldowsky and Newport (1993) presented evidence of the noise-reducing power of random filtering in a statistical learning model of a simple morphological system. Kareev, Lieberman, and Lev (1997) offered a statistical argument in favor of the correlation-enhancing power of small samples and performed two empirical studies purported to confirm this. The other two studies are more purely empirical. Cochran, McDonald, and Parault (1999) taught participants ASL verbs with and without the presence of a simultaneous cognitive load and with practice on the full signs or on individual morphemes. Finally, Kersten and Earles (2001) taught participants a simple novel language with and without sequential input. We discuss each of the four papers here in some detail.

3 Goldowsky and Newport (1993)

Goldowsky and Newport (1993) proposed a simple learning task, and one form of learning model that might be used to solve the task. Training examples consisted of pairings of forms and meanings. A form had three parts, A, B, and C. For each part there were three possible values: A_1, A_2, A_3, B_1, B_2, etc. Meanings were also composed of three parts, M, N, and O, each with three values. There was a very simple mapping from forms to meanings: A_1, A_2, and A_3 corresponded to M_1, M_2, and M_3, respectively, B_1, B_2, and B_3 corresponded to N_1, N_2, and N_3, and so forth.[6] Thus, the form $A_2B_1C_3$ had the meaning $M_2N_1O_3$. The task was, apparently, to learn this simple underlying mapping.

Goldowsky and Newport suggested that one way to solve the task might be to gather a table with counts of all form and meaning correspondences across some observed data. If the form $A_2B_1C_3$ and the meaning $M_2N_1O_3$ were observed, the model would increment values of cells in the table corresponding to the pairing of each of the eight subsets of the form symbols with each subset of the three meaning symbols. If trained on all 27 possible examples, the model would have a value of 9 for each of the cells correctly pairing individual elements of the form to individual elements of the meaning (e.g. A_1 to M_1 and B_3 to N_3). The next largest, incorrectly paired, cells would have a value of 3 and the rest of the cells would have a value of 1.

Goldowsky and Newport suggested that there is too much noise in such a table because of the many values representing incorrect or overly complex pairings. They then introduced a filtering scheme meant to simulate the effect of poor working memory on a child's experiences. Before a form/meaning pair is entered into the table, some of its information is lost at random. Half of the time one of the three elements of the form is retained and half of the time two elements are retained. Likewise for the meaning. The authors argued that this improves learning because it produces a table with a higher signal-to-noise ratio. Therefore, they concluded, having limited memory can be helpful because it can help the learner focus on the simple, often important, details of a mapping.

But we should examine this learning situation a bit more carefully. First of all, in what sense is the signal-to-noise ratio improving as a result of filtering? The ratio between the correct, largest values in the table in the adult (unfiltered) case and the next largest competitors was 3:1. In the child (filtered) case, the expected ratio remains 3:1. Although some of the competitors will become proportionately less likely, others will not. What is eliminated by the filtering is the large number of very unlikely mappings. So the signal-to-noise ratio is improving if it is taken to be the ratio of the correct value to the sum of all other values. If taken to be the ratio of the correct value to the nearest incorrect value, there is no improvement. Furthermore, the child learner must experience many more form/meaning pairings than the adult learner before it can adequately fill its co-occurrence table.

To see the implications of these points, we need to make

[6] The mapping used in the Goldowsky and Newport (1993) paper actually included one exception, that form $A_4B_4C_4$ has meaning $M_3N_3O_3$. Because the introduction of this did not seem to strengthen their case for starting small, it is eliminated here for simplicity.

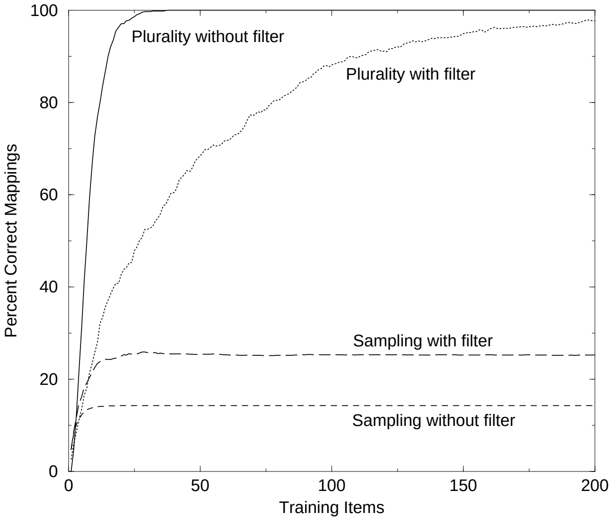

Figure 4: Learning the Goldowsky & Newport (1993) task using raw counts in a noise-free environment.

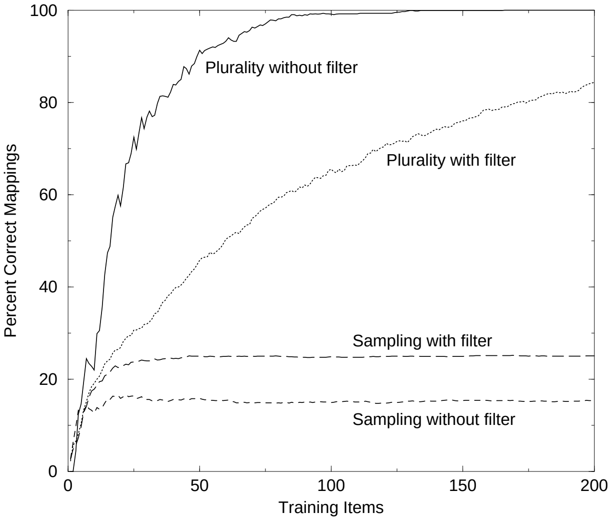

Figure 5: Learning the Goldowsky & Newport (1993) task using raw counts with random loss of 50% of the data.

the task somewhat more explicit. Goldowsky and Newport (1993) presented a model that counts statistics, but not one that actually solves the form/meaning mapping. To complete the story, we will need to generate a model that is capable of taking a form and producing its best guess for the appropriate meaning. Two potential solutions to this problem immediately come to mind. In the first, arguably simpler, method, the model looks down the column of values under the given form and chooses the meaning corresponding to the largest value. If two meanings have the same strength, the model is counted wrong. This will be referred to as the *Plurality* method.

In the second method, the model draws at random from the distribution of values, such that the probability of selecting a meaning is proportional to the value associated with that meaning. This *Sampling* method seems to be more in line with what Goldowsky and Newport implied might be going on, judging from their use of the term signal-to-noise ratio. The Plurality method only fails if the nearest competitor is as strong as the correct answer. In contrast, the Sampling method is wrong in proportion to the total strength of competitors. Both of these methods were implemented and tested experimentally with and without random filtering. The models were judged by their ability to provide the correct meaning for each of the nine forms involving a single element. The results, averaged over 100 trials in each condition, are shown in Figure 4.

As Goldowsky and Newport (1993) suggested, their filtering mechanism is indeed beneficial when used with the Sampling method, achieving a score of about 25.2% versus 14.3% without filtering. However, Sampling overall performs quite poorly. The Plurality method is much more effective. But in that case, filtering is harmful, and slows learning down considerably. Even after 200 trials, the filtered model is able to completely solve the task only about 80% of the time.

Now one might reasonably make the argument that this isn't a fair comparison. Perhaps the Plurality method is much more susceptible to noise and the benefit of the filter isn't apparent in such perfect conditions. After all, it is probably unreasonable to expect that a human learner is able to perfectly notice and store all available information. To test this possibility, a source of noise was added to the simulations. 50% of the time, the operation of incrementing a value in the table failed. Thus, half of the data was lost at random. As shown in Figure 5, this manipulation had almost no effect on the Sampling method, but did have some effect on the Plurality method. However, the Plurality method remained significantly better without the filter.

A final consideration is that the bubble diagrams used to represent the form/meaning co-occurrence table in the Goldowsky and Newport (1993) paper did not directly reflect raw co-occurrence counts. The radius of the bubbles was proportional to the ratio of the co-occurrence count to the square root of the product of the overall number of occurrences of the form and the overall number of occurrences of the meaning. This was termed the *consistency of co-occurrence*. So one might ask, how well do the two proposed models perform if they work with co-occurrence consistency values rather than raw counts. As shown in Figure 6, performance declines slightly for the Sampling method and improves slightly for the Plurality method with filtering. But overall the results are qualitatively similar.

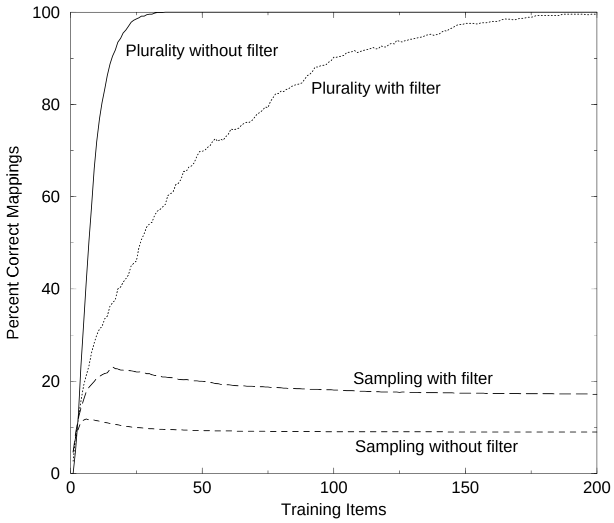

Figure 6: Learning the Goldowsky & Newport (1993) task using correlation values with no noise.

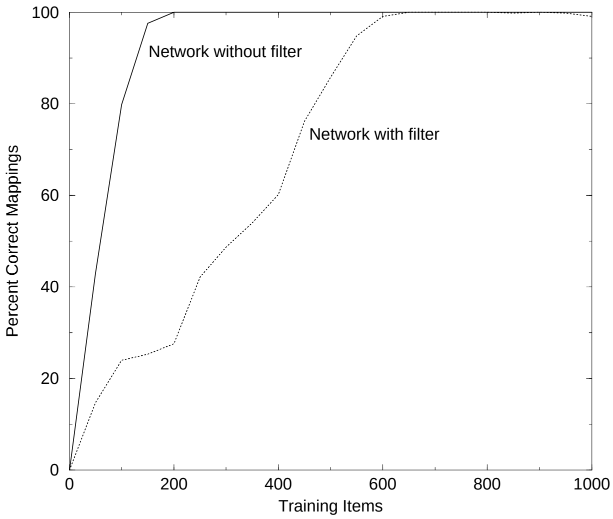

Figure 7: Learning the Goldowsky & Newport (1993) task using a single layer neural network.

Thus, with the much more effective Plurality method of determining form/meaning pairs from co-occurrence data, the filtering mechanism was a serious hindrance. But it seems that building a large table may not be at all similar to the way the human brain might solve this mapping task. Perhaps a better model is that of a connectionist network. Could such a model learn the underlying regularity and would it benefit from the same filtering method proposed by Goldowsky and Newport? To answer this question, we performed some simulation experiments.

First a simple one-layer network was constructed, with a 9-unit input layer fully connected to a 9-unit output layer. The nine input units corresponded to the nine possible elements of the form. One of the first three units was turned on to represent the *A* element, one of the second set of three units was turned on to represent the *B* element, and so forth. Similarly, the nine units in the output representation corresponded to the nine possible elements of the meaning, with three of the nine units normally having targets of 1, and the rest having targets of 0. If an element of the form was eliminated by the filtering mechanism, the corresponding three units of the input were all turned off. If an element of the meaning was eliminated, the corresponding three units of the output had no target values. The network was tested by presenting it with a single element of the form as an input. Although the network may never have been trained to perform this particular mapping, the desired response is that it will output just the corresponding element of the meaning. A response was considered correct if the activations of all nine output units were on the correct side of 0.5.

In order to argue that filtering is or is not beneficial, one cannot simply rely on performance under a single set of training parameters. It is possible that the benefit of filtering could be masked by a poor choice of parameters. Therefore, we trained networks using 32 parameter sets. Four learning rates (0.05, 0.1, 0.2, 0.4) were crossed with two momentum values (0.0, 0.9), two initial weight ranges (±0.1, ±1.0), and two weight decay values (0.0, 0.0001). Networks were trained on 1000 randomly selected examples using online learning, meaning that weight updates were performed after each example.

Performance was measured by testing the model's ability to produce the correct meaning for each of the nine isolated forms. The final performance in each condition, averaged over 50 trials, is shown in Table 3. Without filtering, the network learns best with small initial weights, some weight decay, momentum, and a large learning rate. With filtering, the network learns best with a small learning rate and no momentum. But under no conditions did filtering improve learning. Figure 7 shows the averaged learning profiles with and without filtering using training parameters with which the filtered networks performed quite well: no weight decay or momentum, initial weights ±0.1, and learning rate 0.05. Although they reach similar final performance, the networks learned much more quickly and smoothly without filtering.

One might argue that we have cheated by applying a single layer network to the task because such a network cannot learn very complex mappings, so it doesn't need filtering to learn this simple one. Admittedly, if the task were not so simple, we would have used a larger network. To test the possibility that a larger network will fail to learn the simple rule without filtering, we trained a two layer, 9-9-9, feed-forward network using the same task and parameters.

Table 3: Final performance levels with a 9-9 network under various conditions. The left value in each pair is the performance without filtering and the right value is the performance with filtering.

Weight	**Momentum**	**Initial**	**Learning Rate**							
Decay		**Weights**	**0.05**		**0.1**		**0.2**		**0.4**	
0	0	±0.1	100.0	98.9	100.0	98.4	100.0	76.7	100.0	44.9
0	0	±1.0	85.6	77.3	96.9	88.7	98.7	75.6	100.0	45.6
0	0.9	±0.1	100.0	33.3	100.0	16.7	100.0	4.4	100.0	3.3
0	0.9	±1.0	100.0	32.2	100.0	15.8	100.0	4.4	100.0	3.3
0.0001	0	±0.1	100.0	99.6	100.0	97.6	100.0	78.0	100.0	44.4
0.0001	0	±1.0	88.9	79.6	97.1	89.3	100.0	76.0	100.0	46.4
0.0001	0.9	±0.1	100.0	42.2	100.0	22.2	100.0	5.6	100.0	3.3
0.0001	0.9	±1.0	100.0	42.2	100.0	22.0	100.0	5.6	100.0	3.1

Table 4: Final performance levels with a 9-9-9 network under various conditions. The left value in each pair is the performance without filtering and the right value is the performance with filtering.

Weight	**Momentum**	**Initial**	**Learning Rate**							
Decay		**Weights**	**0.05**		**0.1**		**0.2**		**0.4**	
0	0	±0.1	0.0	1.1	42.0	2.2	92.9	8.9	99.1	26.9
0	0	±1.0	60.2	14.2	72.2	41.6	88.4	40.7	88.4	33.3
0	0.9	±0.1	98.7	24.9	93.8	14.4	81.1	6.4	19.6	2.4
0	0.9	±1.0	81.8	23.8	79.1	14.4	76.2	5.8	41.1	2.4
0.0001	0	±0.1	0.0	1.1	35.6	2.2	94.0	7.6	99.6	26.9
0.0001	0	±1.0	66.0	10.0	79.1	37.1	93.1	47.1	88.4	34.7
0.0001	0.9	±0.1	99.3	24.7	99.3	16.2	99.6	6.9	94.0	2.9
0.0001	0.9	±1.0	99.3	25.6	99.3	15.6	99.1	5.6	99.1	3.6

As shown in Table 4, the two layer network doesn't solve the task as easily as the one layer network. But under several different choices of parameters, the network is able to master the task nearly all of the time without filtering. The best performance achieved with filtering, on the other hand, was just 47.1% correct. In only two cases—with a small learning rate, small initial weights, and no momentum—did the filtered networks perform better than the unfiltered ones. But in those cases the filtered networks only reached an average performance of 1.1%.

In summary, the filtering mechanism proposed by Goldowsky and Newport (1993) for this task did not improve the performance of either an effective tabulation strategy or two neural network models. Although the random filtering mechanism sometimes isolates correct one-to-one form/meaning pairs, it more frequently destroys those pairs and isolates incorrect ones. This introduces noise that outweighs the occasional benefit and that can be detrimental to learning.

4 Kareev, Lieberman, and Lev (1997)

Kareev, Lieberman, and Lev (1997) began by reiterating a theoretical point about sampled distributions which was first raised in Kareev (1995). If a distribution over two correlated real-valued variables is sampled repeatedly, the expected median of the observed correlations in the samples increases as the size of the sample decreases. On the basis of this fact, Kareev et al. suggested that humans estimating correlations in observed events will be better at detecting those correlations if they have limited working memory, and thus presumably rely on smaller remembered samples in formulating their judgments.

In the first experiment, participants were given 128 envelopes, each containing a coin. Envelopes were either red or green and the coin inside was either marked with an X or an O. Participants opened envelopes one-by-one in random order and each time tried to predict the type of coin based on the envelope's color. The envelopes' contents were manipulated to produce true color/mark correlations ranging from -0.6 to 0.6. The eight participants in each condition were grouped based on the results of a single-trial digit-span test of working memory. Response correlation was computed for each participant using the matrix of envelope colors and mark predictions. Kareev et al. found that the low-span participants tended to have larger response correlations and to have more accurate overall predictions.

This is certainly an interesting result, but the theoretical explanation ought to be reconsidered. To begin with, the authors stressed the fact that *median* observed correlation increases as sample size decreases. That is, with a smaller sample, observers have a higher probability of encountering a correlation that is larger than the true correlation. This is mainly an artifact of the increased noise resulting from small samples. On the basis of increasing median, Kareev et al. concluded that, "The limited capacity of working memory increases the chances for early detection of a correlation.. . . [A] relationship, if it exists, is more likely to be detected, the smaller the sample" (p. 279). Thus, the authors seem to be equating median estimation with the ability to detect any correlation whatsoever. However, they do not offer an explicit account of how participants might be solving the correlation detection or coin prediction task.

The median correlation happens to be one measure computable over a series of samples.[7] But there are other measures that may be more directly applicable to the problem of detecting a correlation, such as the *mean*, and not all measures increase in magnitude with smaller samples. The *mean* correlation diminishes with decreasing sample size. But an individual participant is not encountering a series of samples, but just one sample, so the median or mean computed over multiple samples is not necessarily relevant.

So what is an appropriate model of how participants are solving the task, and how is this model affected by sample size? Signal detection theory typically assumes that human observers have a threshold above which a signal is detected. In this case, we might presume that the signal is the perceived correlation between envelope color and coin type, and that the correlation, whether positive or negative, is detectable if its magnitude is above a participant's threshold. If participants are basing their responses in the coin prediction task on a signal detection procedure involving a fixed threshold, we must ask what is the probability that a sample of size N from a distribution with true correlation C has an observed correlation greater than a given threshold?

It seems reasonable to suppose that the typical human threshold for detecting correlations in small samples probably falls between 0.05 and 0.2, although it presumably varies based on task demands. Figure 8 shows the probability that a small sample has an observed correlation above 0.1 as a function of the size of the sample and the strength of the true correlation. The data in this experiment involved pairs of real-valued random variables. A desired correlation, C, was achieved by generating the values as follows:

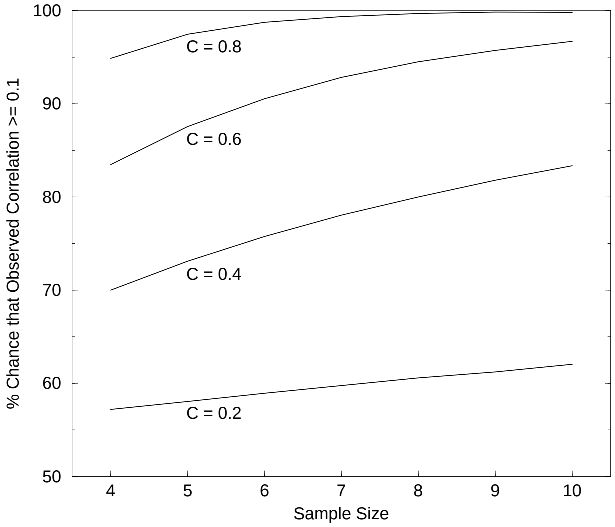

Figure 8: The probability that the observed correlation value is greater than 0.1 (and thus presumably detectable) as a function of sample size and true correlation (C).

$$a = \text{rand}()$$
$$b = Ca + \sqrt{1 - C^2}\ \text{rand}()$$

where rand() produces a random value uniformly distributed in the range [-1,1]. 1 million trials were conducted for each pairing of sample size and correlation.

Clearly, for the range of parameters covered, the chance that the observed correlation is greater than threshold increases monotonically with sample size. Larger samples lead to a greater chance of detecting a correlation. One may disagree with the arbitrary choice of 0.1 for the detection threshold, but the same penalty for small samples is seen with a value of 0.2, provided the true correlation is greater than 0.2, and the effect becomes even stronger with thresholds below 0.1. Thus, the fact that the median observed correlation increases with small sample sizes does not bear on what is arguably a reasonable model of human correlation detection.

Another important issue is that the sampling distribution measures discussed by Kareev et al. were for pairs of real-valued variables, but the experiments they conducted involved binary variables. Do the same principles apply to small samples of binary data? Figure 9 shows the median observed correlation in small samples of binary data, as a function of the sample size and the true correlation. Although median correlation decreases as a function of sample size for real-valued data, median correlation doesn't seem to vary in any systematic way as a function of sample size for binary data. There is simply more variabil-

[7] The term *sample* is used here to refer to a set of observations, or examples, not just a single observation.

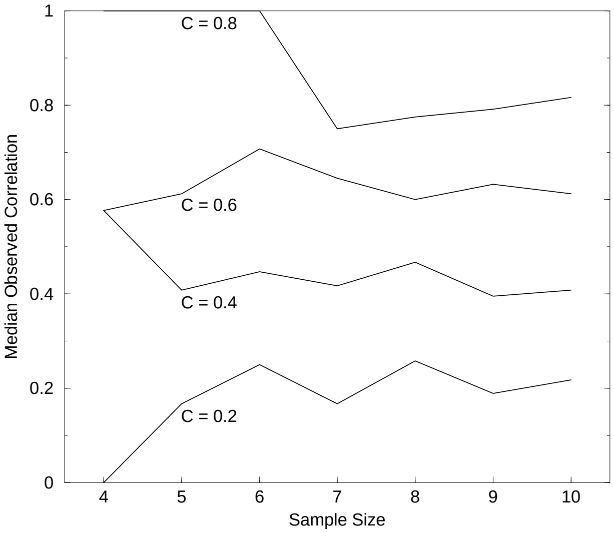

Figure 9: The median observed correlation in small samples of binary data, as a function of sample size and true correlation (*C*).

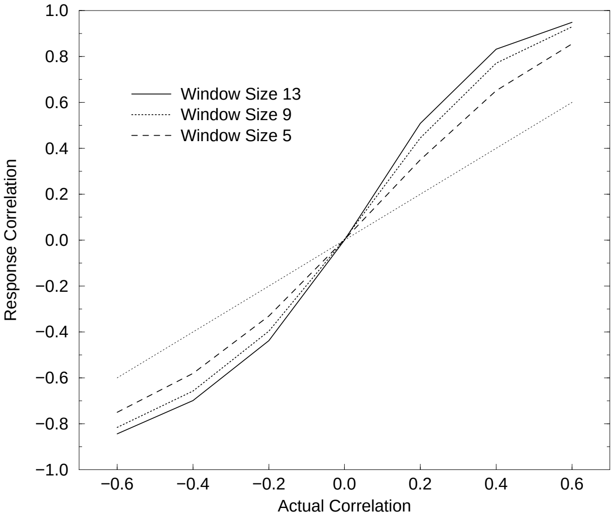

Figure 10: The correlation between envelope color and the models' predictions of coin marking as a function of the actual correlation and the model's memory window size.

ity in the small samples. But again, median correlation value is not necessarily indicative of the ease of detection. As with real-valued data, the probability that an observed correlation is greater than some small threshold tends to increase with larger samples of binary data.

But it may be possible that these statistical measures don't accurately reflect the power of small samples in a practical context. Therefore, we designed a simple model to perform the envelope/coin task using varying levels of working memory. The model was intended to reflect the manner in which Kareev et al. seem to imply humans might be solving this task. The model simply remembers the contents of the last *N* cards of each color and chooses the coin that was more frequent in that sample. If the coins were equally frequent in the sample, the choice is random. The model was run with three sample sizes, 5, 9, and 13, meant to reflect small, medium, and large working memory capacity and was run 1000 times on each of the 14 distributional conditions used by Kareev, Lieberman, and Lev (1997). 7 of these conditions were symmetric in that they used an equal number of X's and O's and 7 did not satisfy this constraint and were termed asymmetric. Each symmetric condition had a corresponding asymmetric one with approximately the same envelope/coin correlation. The correlation between the models' predictions and the envelope color was computed in the same way as for the experimental participants.

Figure 10 shows the prediction correlation values as a function of actual correlation for the three working memory levels, with results in the corresponding symmetric and asymmetric conditions averaged. The identity baseline is provided as a reference, but note that optimal performance in this task has nothing to do with matching the actual correlation values. An optimal predictor will always predict the more likely coin, whether the actual correlation is 0.1 or 0.9. Contrary to Kareev et al.'s prediction, the larger sample size results in larger response correlations, not smaller ones. Figure 11 gives the prediction accuracy as a function of correlation and window size. Although the difference is fairly small, larger window sizes consistently outperformed the smaller ones.

Therefore, although the results of the first experiment in Kareev, Lieberman, and Lev (1997) are rather interesting and deserve replication and explanation, these results cannot be attributed to the effects of small samples on perceived correlation. The probability of observing a correlation stronger than a relatively sensitive detection threshold is lower with small sample sizes and the median observed correlation value with binary data does not change systematically with sample size. A simple prediction model that relies on samples of varying size performs better with larger samples. While it is true that this model does not appear to fully capture human performance in this task, the relevant point is that the effects of small sample sizes on perceived correlation do not adequately explain the empirical findings.

The second experiment reported by Kareev, Lieberman, and Lev (1997) also does not seem to fully support their theory. In this case, participants were not blocked by digit span but were given samples of varying size upon which to base a prediction. The samples were either fully visible throughout the process or were presented sequentially

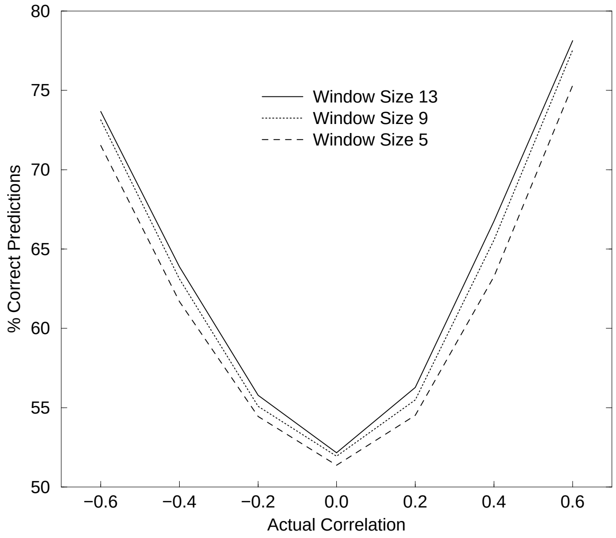

Figure 11: The prediction accuracy as a function of the actual correlation and the model's memory window size.

and were unavailable in formulating the prediction. In this case, the variables were real-valued, rather than binary. The results indicated that when samples were absent, there was better performance with the small samples than with the medium or large ones. But when the samples were present, performance increased with sample size. This latter result is inconsistent with the prediction that small samples should statistically magnify correlations. If that were true, larger samples would lead to worse performance, especially if the samples are present. The fact that participants viewing sequential samples performed better with smaller ones is indeed interesting, but cannot be explained by a statistical property of sample size itself.

5 Cochran, McDonald, and Parault (1999)

Much of the empirical support for the less-is-more hypothesis derives from the study of American Sign Language (ASL). Newport (1990) observed that late learners of ASL tend to make more morphological errors in the production of verbs than do early learners. While interesting, it is not clear to what this finding should be attributed. The problems incurred by late learners could be due to deactivation of a language acquisition device, greater cognitive capacity, different types or degrees of exposure, or a variety of other factors. Cochran, McDonald, and Parault (1999) sought to provide empirical evidence supporting the idea that cognitive limitations can actually lead to better learning of ASL verbs. They conducted three experiments in which participants unfamiliar with ASL were taught some sentences and then tested in their ability to produce either the same or novel ASL sentences.

In the first two experiments, participants were taught 16 verbs. Each verb was encountered in the context of a single sentence, in which either the subject was "I" and the object was "you", or vice-versa. Six of the verbs used *congruent agreement*, in which the direction of the sign was from the verb's subject (either the signer or the addressee) to the verb's object. Two of the verbs used *incongruent agreement*, in which the direction of the sign was from object to subject. Four *nonagreement* verbs required a static direction of motion, which was either always away from or always toward the signer. The last four verbs had a direction of motion aligned vertically, either up or down.

Participants were exposed to each verb in a single context, with half of the verbs in each condition using the subject "I" and half using the subject "you". The 16 study sentences were observed three times in the first experiment and eight times in the second experiment. In order to place a load on working memory, half of the participants performed a tone-counting task during training. This was known as the *load* condition. Participants were then tested on the 16 familiar sentences as well as the 16 novel sentences created by reversing the subject and object.

Cochran, McDonald, and Parault (1999) found that participants in the no-load condition produced the familiar sentences better overall and performed better on familiar and novel non-agreement verbs. However, participants in the no-load condition did not perform as well on the agreement verbs in novel sentences. They were much more likely to produce the sign in the same direction that they learned it, rather than reversing the direction in the new context. This was taken as evidence that "adults learning under normal conditions were failing to learn the internal structure of the language and were therefore limited in their ability to generalize to new contexts" (p. 30).

However, an alternative reading of the data is that participants in the load condition were simply not learning as well and performed more randomly during test. Not only did load participants have more movements in the correct direction, they produced more verbs with no movement or, in the first experiment, with movement outside the axis between the signer and addressee. The fact that load condition participants happened to use the correct movement more often in novel conditions can be attributed to their generally more noisy behavior, rather than their having learned to generalize to novel conditions.

The main problem with these experiments is that participants are expected to learn that the movement of certain verbs should agree with sentence context when there was no basis for such a generalization in the examples to which the participants had been exposed. Each verb was seen in just one context, with just one direction of motion, and only six of the 16 verbs underwent congruent agreement. The evidence to which the participants were exposed fully supports the simpler hypothesis: that direction of motion is an intrinsic, non-inflected part of the sign for a verb. In fact, this is the correct rule for half of the verbs used in the experiment. Given the lack of any evidence to the contrary, it seems much more reasonable for participants to surmise that ASL permits no agreement, than to surmise that some verbs have agreement, some have incongruent agreement, and some have no agreement. The results in these experiments are consistent with the hypothesis that participants in the no-load condition learned this very reasonable rule much better than did participants in the load condition.

A true test of generalization ability must provide the learner with some support for the validity of the expected generalization. Had participants experienced some agreement verbs used with different motions in different circumstances, they would have some basis for expecting that agreement plays a role in ASL. A second factor biasing the participants against formulating the desired generalization was that, unlike in ASL, pronouns were explicitly produced in all training sentences. Languages with strong verb inflection, such as Spanish, often drop first- and second-person pronouns, because they convey redundant information. Because such pronoun drop was not a feature of the training sentences, learners are more likely to assume that pronominal information is not redundantly conveyed in the verb form. In summary, the first two experiments of this study essentially found that participants trained to perform one reasonable generalization did poorly when tested on a different, more complex, generalization.

The third experiment conducted by Cochran, McDonald, and Parault (1999) tested the learning of ASL motion verbs, comparing participants who were taught to mimic whole signs to those who were taught to mimic just one part of each sign, either the form or the motion, at a time. During training, signs for a certain type of actor moving in a certain way were paired with a hand movement indicating the path of motion. For some verbs, the motion sign is produced at the same time as the verb, but for other verbs they are produced in sequence. During testing, all verbs were paired with all path signs.

Overall there was no difference in performance on the studied or the novel signs between the "whole" and "part" learners. There was an unexplained tradeoff, in that whole learners did better if the parts of the new sign were to be performed sequentially and worse if they were to be performed simultaneously. The only other difference was the marginally significant tendency for whole-practice participants to produce more frozen signs,[8] which could be a cause or effect of the other difference. If anything, this study seems to provide strong evidence that learning individual parts of signs is not, overall, of significant benefit. Although whole-sign learners produced more frozen signs, they performed better in other respects, balancing the overall performance. Somewhat disturbingly, however, more participants were thrown out for inadequate performance or unscorable data from the part-learning group. One person in the whole-sign condition was thrown out for unscoreable data and 9 people in the part-sign condition were replaced, three for bad performance and two for unscoreable data. Across the three experiments, three participants were discarded from the no-load and whole-sign conditions for performance or scoreability reasons, compared with 12 participants in the load and part-sign conditions. In experiments of this sort involving a direct comparison between training methods, eliminating participants for performance reasons during training has the clear potential to bias the average testing performance. If participants must be removed from one condition for performance reasons, an equal number of the worst performers in the other conditions should be removed as well, although this still may not fully eliminate the bias.

6 Kersten and Earles (2001)

Kersten and Earles (2001) conducted three language learning experiments which compared learning in a staged input condition to learning in a full-sentence condition. In each experiment, participants viewed events in which one bug-like object moved towards or away from another, stationary, bug-like object. In the full-sentence condition, each event was paired with the auditory presentation of a three-word sentence. The first word corresponded to the appearance of the moving bug and ended in "–ju". The second word described the manner of motion—either walking with legs together or alternating—and ended in "–gop".[9] The third word described the direction of walking—towards or away from the stationary bug—and ended in "–tig".

In the first two experiments, half of the participants heard complete sentences for the whole training period. The other participants initially heard just the first (object) word for a third of the trials, then the first two words, and finally all three words. In the testing period, participants were shown two events that varied on a single attribute and heard either an isolated word (corresponding to the manipulated attribute) or a sentence. They were to identify the event that correctly matched the word or sentence.

The most important finding in these experiments was significantly better performance, overall, for participants

[8] A frozen sign was a new sign that contained an unnecessary part of a previously studied sign.

[9] In the first experiment, some participants heard object-manner-path word order and others heard object-path-manner.

in the staged input condition. Kersten and Earles interpreted this as evidence in favor of the less-is-more hypothesis. However, one should exercise some caution in drawing conclusions from these experiments. Although there was an overall advantage for starting small, if one tests performance on object words, manner words, and path words independently, the effect is only significant for object words. Thus, the results are consistent with the hypothesis that starting small was only beneficial in learning the meanings of the object words, i.e., those words trained in isolation for the first third of the trials.

Kersten and Earles sought to rule out a slightly different, but equally viable, hypothesis—that the effect relies on the fact that the object words, as opposed to manner or path, were learned first. Therefore, in the third experiment, participants in the staged condition first heard the last (path) word, then the last two words (manner-path), and finally all three words. Again there was a significant overall advantage for the staged input condition. In this case, path words were learned better than object and manner words in both conditions. Although the overall advantage for the starting small condition reached significance, none of the tests isolating the three word types were significant. These results therefore do not rule out the hypothesis that participants in the staged input condition were only better on the words trained in isolation. Nevertheless, it is possible that these effects would reach significance with more participants.

The third experiment also added a test of the participants' sensitivity to morphology. Novel words were created by pairing an unfamiliar stem with one of the three familiar word endings (–ju, –gop, or –tig). Each word was first paired with an event that was novel in all three important dimensions. Participants were then shown a second event that differed from the first in a single dimension and were instructed to respond "Yes" if the second event was also an example of the new word. In other words, participants responded "Yes" if the two events *didn't* differ on the feature associated with the word ending. Kersten and Earles again found a significant advantage for the starting small condition.

However, there is some reason to question the results of this experiment. With the path-word ending, there was clearly no difference between the two conditions. In three of the four other conditions, participants performed below chance levels, significantly so in one of them. The finding of significantly below chance performance leads one to suspect that participants may have been confused by the task and that some participants may have incorrectly been responding "Yes" if the events did differ on the feature associated with the word ending.

Even if we accept that there was an across-the-board advantage for the staged input condition in these experiments, we should be cautious in generalizing to natural language learning. The language used in this study was missing a number of important features of natural language. Word order and morphology were entirely redundant and, more importantly, conveyed no meaning. Words always appeared in the same position in every sentence and were always paired with the same ending. In this simple language, there wasn't a productive syntax or morphology, just a conventional word order. Participants were thus free to use strategies such as ignoring word order and morphological information, much as they learned to ignore meaningless details of the events.

Participants in the full sentence condition were therefore at a potential disadvantage. Any effective, general learning mechanism in a similar situation would devote time and resources to testing the information carried in all aspects of the events and sentences, including morphology and word order. In this case, those features happened to convey no additional information beyond that provided by the word stems themselves, placing participants who paid attention to word order and morphology at a disadvantage. However, these factors play critical roles in shaping the meaning of natural language sentences, and devoting time and resources to learning them is useful, and even necessary. The staged input learner, on the other hand, will have traded off exposure to syntax for more exposure to individual words and their meanings, which is not clearly advantageous. A stronger test of the importance of staged input would be to measure comprehension or production of whole, novel sentences in a language with some aspects of meaning carried exclusively by syntax and morphology.

Perhaps tellingly, some studies cited by Kersten and Earles comparing children learning French in immersive programs with and without prior exposure to more traditional, elementary French-as-a-second-language courses found either no difference or an advantage for children in the purely immersive programs (Shapson & Day, 1982; Day & Shapson, 1988; Genesee, 1981). Although these studies may not have adequately controlled for age of exposure, intelligence, or motivational factors, it certainly is suggestive that staged input may be less effective than immersion in learning natural languages.

A final point of criticism of the Kersten and Earles (2001) paper is their desire to equate the effects of staged input with those of internal memory limitations. There is little reason to believe that these two factors will have similar effects. Teaching the meanings of isolated words is bound to be helpful, provided that it is only a supplement to exposure to complete language, is relatively noise free, and makes up a relatively small percentage of linguistic experience. However, memory limitations do not result in the same simple pairing of words and their meanings. At best, memory limitations have the effect of pairing isolated words or phrases to noisy, randomly sampled portions of a complex meaning. The actual part of the complex meaning contributed by the isolated word may be partially or completely lost and some extraneous information may be retained. Learning the correct pairings of words to meanings is no easier in this case than when faced with the full, complex meaning.

A more appropriate, though still not entirely sufficient, test of the benefit of memory limitations in the context of Kersten and Earles's design would be to test randomly selected words in the isolated word condition, rather than always the first or last word of the sentence. These should be paired with scenes with randomly selected details, such as the identity of the moving object or the location of the stationary object, obscured. Furthermore, tests should not be performed on familiar sentences but on novel ones, as the potential problem in starting with complete sentences is that adults will memorize them as wholes and will not generalize well to novel ones. It would be quite interesting if initial training of this form, which is more like the presumed effect of poor attention or working memory, was beneficial in the comprehension or production of novel sentences.

The actual claim of Newport's less-is-more hypothesis does not concern staged input. It is that memory or other internal limitations are the key factor in enabling children to learn language more effectively. Evidence for or against the benefit of staged input should be clearly distinguished from evidence concerning the effect of internal cognitive impairments.

7 General Discussion

We believe that studying the way in which connectionist networks learn languages is particularly helpful in building an understanding of human language acquisition. 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 using its 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 context to hidden connections involved in retaining that information will be reinforced, leading the information to 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 this type of 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 representations 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).[10]

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 nonstationary probability distribution over inputs. In extreme form, such nonstationarity can lead to so-called *catastrophic interference*, in which training exclusively on a new task can dramatically impair

[10] It should be pointed out that the bias towards learning short- before long-range dependencies is not specific to simple recurrent networks; backpropagation-through-time and 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 when the mapping is simple (see Bengio et al., 1994; Lin et al., 1996).

performance on a previously learned task that is similar to but inconsistent with the new task (see, e.g., McClelland, McNaughton, & O'Reilly, 1995; McCloskey & Cohen, 1989).

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 (2001) 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, 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).

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, Gleitman, & Gleitman, 1977; 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. The simple recurrent network simulations have focused on 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 for extended periods in the input to either a child or adult learner.

But the main implication of the less-is-more hypothesis is not that staged input is necessary, but that the child's superior language learning ability is a consequence of the child's limitations. This might be interpreted in a variety of ways. Goldowsky and Newport (1993), Elman (1993), Kareev, Lieberman, and Lev (1997), and Cochran, McDonald, and Parault (1999) suggest that the power of reduced memory is that it leads to information loss which can be beneficial in highlighting simple contingencies in the environment. This, it is suggested, encourages analytical processing over rote memorization. We have argued, to the contrary, that in a range of learning procedures, from simple decision making models to recurrent connectionist networks, such random information loss is of no benefit and may be harmful. Although it sometimes has the effect of isolating meaningful analytical units, it more often destroys those units or creates false contigencies.

Another take on the less-is-more hypothesis is that a learning system can benefit by being differentially sensitive to local information or simple input/output relationships. This we do not deny. In fact, it seems difficult to conceive of an effective learning procedure that is not better able to learn simple relationships. A related argument is that when the mapping to be learned is componential, a learning procedure specialized for learning such mappings, as opposed to one specialized for rote memorization, is to be preferred. This, too, we support. However, we suggest that neural networks—and, by possible implication, the human brain—are naturally better at learning simple or local contingencies and regular, rather than arbitrary, mappings. But this is true of learning in experienced networks or adults, just as it is true of learning in randomized networks or children. The general architecture of the system is the key factor that enables learning of componentiality, not the child's limited working memory.

Simulating poor working memory by periodically disrupting a network's feedback during the early stages of learning has relatively little effect because, at that point, the network has not yet learned to use its memory effectively. As long as memory is interfered with less as the network develops, there will continue to be little impact on learning. In a sense, early interference with the network's memory is superfluous because the untrained network is naturally memory limited. One might say that is the very point of the less-is-more argument, but it is missing a vital component. While we accept that children have limited cognitive abilities, we don't see these limitations as a source of substantial learning advantage to the child. Both are symptoms of the fact that the child's brain is in an early stage in development at which its resources are largely uncommitted, giving it great flexibility in adapting to the particular tasks to which it is applied.

7.1 Late Exposure and Second Languages

Elman's (1991, 1993) computational findings of the importance of starting small in language acquisition, as well as the other studies reviewed here, 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 & Newport, 1989; Newport, 1990; Goldowsky & 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 normally 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, Munro, & MacKay, 1995). The aspects of an adult's second language that are most difficult may be those that directly conflict with the learned properties of the native language. For example, learning the inflectional morphology of English may be particularly difficult for adult speakers of an isolating language, such as Chinese, which does not inflect number or tense.

By contrast to the adult, the child ultimately achieves a higher level of performance on a first or second language because his or her resources are initially uncommitted, allowing neurons to be more easily recruited and the response characteristics of already participating neurons to be altered. Additionally, the child is less hindered by interference from prior learned representations. This idea, which accords with Quartz and Sejnowski's (1997) theory of *neural constructivism*, is certainly not a new one, but is one that seems to remain largely ignored (although see Marchman, 1993; McClelland, 2001). 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 are not themselves 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 & 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 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 he or she begins 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 these abilities rely 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 & 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.

7.2 Conclusion

We seem to be in agreement with most proponents of the less-is-more hypothesis in our belief that the proper account of human language learning need not invoke the existence of innate language-specific learning devices. However, we depart from them in our skepticism that limited cognitive resources are themselves of critical importance in the ultimate attainment of linguistic fluency. The simulations reported here, principally those inspired by Elman's language-learning work, call into question the proposal that staged input or limited cognitive resources are necessary, or even beneficial, for learning. We believe that the cognitive limitations of children are only advantageous for language acquisition to the extent that they are symptomatic of a system that is unorganized and inexperienced but possesses great flexibility and potential for future adaptation, growth and specialization.

Acknowledgements

This research was supported by NIMH Program Project Grant MH47566 (J. McClelland, PI), and by an NSF Graduate Fellowship to the first author. Correspondence regarding this article may be sent either to Douglas Rohde (dr@cs.cmu.edu), School of Computer Science, Carnegie Mellon University, 5000 Fifth Avenue, Pittsburgh, PA 15213–3890, USA, or to David Plaut (plaut@cmu.edu), Mellon Institute 115–CNBC, 4400 Fifth Avenue, Pittsburgh, PA 15213–2683, USA.

References

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

Bialystok, E., & Hakuta, K. (1999). Confounded age: Linguistic and cognitive factors in age differences for second language acquisition. In D. P. Birdsong (Ed.), *Second language acquisition and the critical period hypothesis* (pp. 161–181). Mahwah, NJ: Erlbaum.

Birdsong, D. (1999). Introduction: Whys and why nots of the critical period hypothesis for second language acquisition. In D. P. Birdsong (Ed.), *Second language acquisition and the critical period hypothesis* (pp. 1–22). Mahwah, NJ: Erlbaum.

Chomsky, N. (1965). *Aspects of the theory of syntax.* Cambridge, MA: MIT Press.

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

Cochran, B. P., McDonald, J. L., & Parault, S. J. (1999). Too smart for their own good: The disadvantage of a superior processing capacity for adult language learners. *Journal of Memory and Language*, *41*, 30–58.

Day, E. M., & Shapson, S. (1988). A comparison study of early and late French immersion programs in British Columbia. *Canadian Journal of Education*, *13*, 290–305.

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 important 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.* Cambridge, MA: MIT Press.

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

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.

Gallaway, C., & Richards, B. J. (Eds.). (1994). *Input and interaction in language acquisition.* London: Cambridge University Press.

Genesee, F. (1981). A comparison study of early and late second language learning. *Canadian Journal of Behavioral Sciences*, *13*, 115–128.

Goldowsky, B. N., & Newport, E. L. (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* (pp. 124–138). Stanford, CA: Center for the Study of Language and Information.

Johnson, J. S., & Newport, E. L. (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.

Kareev, Y. (1995). Through a narrow window: Working memory capacity and the detection of covariation. *Cognition*, *56*, 263–269.

Kareev, Y., Lieberman, I., & Lev, M. (1997). Through a narrow window: Sample size and the perception of correlation. *Journal of Experimental Psychology*, *126*(3), 278–287.

Kersten, A. W., & Earles, J. L. (2001). Less really is more for adults learning a miniature artificial language. *Journal of Memory and Language*, *44*, 250–273.

Kohonen, T. (1984). *Self-organization and associative memory.* New York: Springer-Verlag.

Lenneberg, E. H. (1967). *Biological foundations of language.* NY: Wiley.

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. Nos. CS-TR-3626, UMIACS-TR-96-28). College Park, MD: University of Maryland.

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

Luce, D. R. (1986). *Response times.* New York: Oxford.

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

McClelland, J. L. (2001). Failures to learn and their remediation: A competitive, Hebbian approach. In J. L. McClelland & R. S. Siegler (Eds.), *Mechanisms of cognitive development: Behavioral and neural perspectives.* Mahwah, NJ: Lawrence Erlbaum Associates.

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.

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

McNeill, D. (1970). *The acquisition of language: The study of developmental psycholinguistics.* New York: Harper & Row.

Merzenich, M. M., & Jenkins, W. M. (1995). Cortical plasticity, learning and learning dysfunction. In B. Julesz & I. Kovacs (Eds.), *Maturational windows and adult cortical plasticity* (pp. 247–272). Reading, MA: Addison-Wesley.

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.

Newport, E. L. (1990). Maturational constraints on language learning. *Cognitive Science*, *34*, 11–28.

Newport, E. L., Gleitman, H., & Gleitman, L. R. (1977). Mother, i'd rather do it myself: Some effects and non-effects of maternal speech style. In C. E. Snow & C. A. Ferguson (Eds.), *Talking to children: Language input and acquisition* (pp. 109–149). Cambridge, England: Cambridge University Press.

Pine, J. M. (1994). The language of primary caregivers. In C. Gallaway & B. J. Richards (Eds.), *Input and interaction in language acquisition* (pp. 38–55). London: Cambridge University Press.

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

Richards, B. J. (1994). Child-directed speech and influences on language acquisition: Methodology and interpretation. In C. Gallaway & B. J. Richards (Eds.), *Input and interaction in language acquisition* (pp. 74–106). London: Cambridge University Press.

Rohde, D. L. T. (1999). *The Simple Language Generator: Encoding complex languages with simple grammars* (Tech. Rep. No. CMU-CS-99-123). Pittsburgh, PA: Carnegie Mellon University, Department of Computer Science.

Rohde, D. L. T., & Plaut, D. C. (1999). Language acquisition in the absence of explicit negative evidence: How important is starting small? *Cognition*, *72*(1), 67–109.

Rumelhart, D. E., Durbin, R., Golden, R., & Chauvin, Y. (1995). Backpropagation: The basic theory. In Y. Chauvin & D. Rumelhart (Eds.), *Back-propagation: Theory, architectures, and applications* (pp. 1–34). Hillsdale, NJ: Erlbaum.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning internal representations by error propagation. In D. E. Rumelhart, J. L. McClelland, & the PDP Research Group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition. Volume 1: Foundations* (pp. 318–362). Cambridge, MA: MIT Press.

Shapson, S. M., & Day, E. M. (1982). A comparison of three late immersion programs. *Alberta Journal of Educational Research*, *28*, 135–148.

Snow, C. E. (1994). Beginning from baby talk: Twenty years of research on input and interaction. In C. Gallaway & B. J. Richards (Eds.), *Input and interaction in language acquisition* (pp. 3–12). London: Cambridge University Press.

Snow, C. E. (1995). Issues in the study of input: Finetuning, universality, individual and developmental differences, and necessary causes. In P. Fletcher & B. MacWhinney (Eds.), *The handbook of child language* (pp. 180–193). Oxford: Blackwell.

Sokolov, J. L. (1993). A local contingency analysis of the finetuning hypothesis. *Developmental Psychology*, *29*, 1008–1023.

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.