Machine Learning, 7, 161-193 (1991)
© 1991 Kluwer Academic Publishers, Boston. Manufactured in The Netherlands.

Graded State Machines: The Representation of Temporal Contingencies in Simple Recurrent Networks

DAVID SERVAN-SCHREIBER, AXEL CLEEREMANS AND JAMES L. McCLELLAND
School of Computer Science and Department of Psychology, Carnegie Mellon University

Abstract. We explore a network architecture introduced by Elman (1990) for predicting successive elements of a sequence. The network uses the pattern of activation over a set of hidden units from time-step t-1, together with element t, to predict element t + 1. When the network is trained with strings from a particular finite-state grammar, it can learn to be a perfect finite-state recognizer for the grammar. When the net has a minimal number of hidden units, patterns on the hidden units come to correspond to the nodes of the grammar; however, this correspondence is not necessary for the network to act as a perfect finite-state recognizer. Next, we provide a detailed analysis of how the network acquires its internal representations. We show that the network progressively encodes more and more temporal context by means of a probability analysis. Finally, we explore the conditions under which the network can carry information about distant sequential contingencies across intervening elements to distant elements. Such information is maintained with relative ease if it is relevant at each intermediate step; it tends to be lost when intervening elements do not depend on it. At first glance this may suggest that such networks are not relevant to natural language, in which dependencies may span indefinite distances. However, embeddings in natural language are not completely independent of earlier information. The final simulation shows that long distance sequential contingencies can be encoded by the network even if only subtle statistical properties of embedded strings depend on the early information. The network encodes long-distance dependencies by *shading* internal representations that are responsible for processing common embeddings in otherwise different sequences. This ability to represent simultaneously similarities and differences between several sequences relies on the graded nature of representations used by the network, which contrast with the finite states of traditional automata. For this reason, the network and other similar architectures may be called *Graded State Machines.*

Keywords. Graded state machines, finite state automata, recurrent networks, temporal contingencies, prediction task

1. Introduction

As language abundantly illustrates, the meaning of individual events in a stream—such as words in a sentence—is often determined by preceding events in the sequence, which provide a context. The word 'ball' is interpreted differently in "The countess threw the ball" and in "The pitcher threw the ball." Similarly, goal-directed behavior and planning are characterized by coordination of behaviors over long sequences of input-output pairings, again implying that goals and plans act as a context for the interpretation and generation of individual events.

The similarity-based style of processing in connectionist models provides natural primitives to implement the role of context in the selection of meaning and actions. However, most connectionist models of sequence processing present all cues of a sequence in parallel and often assume a fixed length for the sequence (e.g., Cottrell, 1985; Fanty, 1985; Selman, 1985; Sejnowski & Rosenberg, 1987; Hanson and Kegl, 1987). Typically, these models use a pool of input units for the event present at time **t**, another pool for event **t + 1**, and so on, in what is often called a 'moving window' paradigm. As Elman (1990) points out, such implementations are not psychologically satisfying, and they are also computationally wasteful since some unused pools of units must be kept available for the rare occasions when the longest sequences are presented.

Some connectionist architectures have specifically addressed the problem of learning and representing the information contained in sequences in more elegant ways. Jordan (1986) described a network in which the output associated to each state was fed back and blended with the input representing the next sate over a set of 'state units' (Figure 1).

After several steps of processing, the pattern present on the input units is characteristic of the particular sequence of states that the network has traversed. With sequences of increasing length, the network has more difficulty discriminating on the basis of the first cues presented, but the architecture does not rigidly constrain the length of input sequences. However, while such a network learns *how to use* the representation of successive states, it does not discover a representation for the sequence.

Elman (1990) has introduced an architecture—which we call a simple recurrent network (SRN)—that has the potential to master an infinite corpus of sequences with the limited means of a learning procedure that is *completely local in time* (Figure 2). In the SRN, the hidden unit layer is allowed to feed back on itself, so that the intermediate results of processing at time **t − 1** can influence the intermediate results of processing at time **t**. In practice, the simple recurrent network is implemented by copying the pattern of activation on the hidden units onto a set of 'context units' which feed into the hidden layer along with the input units. These context units are comparable to Jordan's state units.

In Elman's simple recurrent networks, the set of context units provides the system with memory in the form of a trace of processing at the previous time slice. As Rumelhart, Hinton and Williams (1986) have pointed out, the pattern of activation on the hidden units

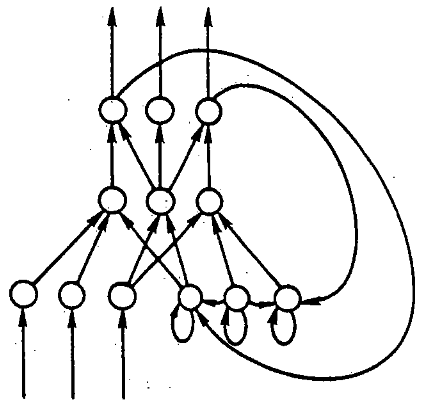

Figure 1. The Jordan (1986) Sequential Network.

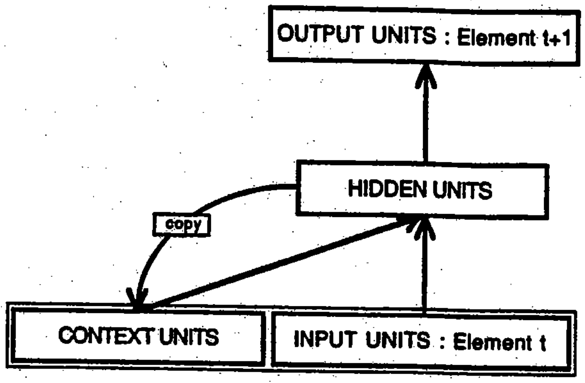

Figure 2. The Simple Recurrent Network. Each box represents a pool of units and each forward arrow represents a complete set of trainable connections from each sending unit to each receiving unit in the next pool. The backward arrow from the hidden layer to the context layer denotes a copy operation.

corresponds to an 'encoding' or 'internal representation' of the input pattern. By the nature of back-propagation, such representations correspond to the input pattern partially processed into features relevant to the task (e.g., Hinton, McClelland & Rumelhart, 1986). In the recurrent networks, internal representations encode not only the prior event but also relevant aspects of the representation that was constructed in predicting the prior event from its predecessor. When fed back as input, these representations could provide information that allows the network to maintain prediction-relevant features of an entire sequence.

In this study, we show that the SRN can learn to mimic closely a finite state automaton (FSA), both in its behavior and in its state representations. In particular, we show that it can learn to process an *infinite* corpus of strings based on experience with a *finite* set of training exemplars. We then explore the capacity of this architecture to recognize and use non-local contingencies between elements of a sequence that cannot be represented conveniently in a traditional finite state automaton. We show that the SRN encodes long-distance dependencies by *shading* internal representations that are responsible for processing common embeddings in otherwise different sequences. This ability to represent simultaneously similarities and differences between sequences in the same state of activation relies on the graded nature of representations used by the network, which contrast with the finite states of traditional automata. For this reason, we suggest that the SRN and other similar architectures may be exemplars of a new class of automata, one that we may call *Graded State Machines*.

2. Learning a finite state grammar

2.1. Material and task

In our first experiment, we asked whether the network cold learn the contingencies implied by a small finite state grammar. As in all of the following explorations, the network is assigned the task of predicting successive elements of a sequence. This task is interesting because it allows us to examine precisely how the network extracts information about whole sequences without actually seeing more than two elements at a time. In addition, it is possible to manipulate precisely the nature of these sequences by constructing different training and testing sets of strings that require integration of more or less temporal information. The stimulus set thus needs to exhibit various interesting features with regard to the potentialities of the architecture (i.e., the sequences must be of different lengths, their elements should be more or less predictable in different contexts, loops and subloops should be allowed, etc.).

Reber (1976) used a small finite-state grammar in an artificial grammar learning experiment that is well suited to our purposes (Figure 3). Finite-state grammars consist of nodes connected by labeled arcs. A grammatical string is generated by entering the network through the 'start' node and by moving from node to node until the 'end' node is reached. Each transition from one node to another produces the letter corresponding to the label of the arc linking these two nodes. Examples of strings that can be generated by the above grammar are: 'TXS', 'PTVV', 'TSXXTVPS'.

The difficulty in mastering the prediction task when letters of a string are presented individually is that two instances of the same letter may lead to different nodes and therefore different predictions about its successors. In order to perform the task adequately, it is thus necessary for the network to encode more than just the identity of the current letter.

2.2. Network architecture

As illustrated in Figure 4, the network has a three-layer architecture. The input layer consists of two pools of units. The first pool is called the context pool, and its units are used to represent the temporal context by holding a copy of the hidden units' activation level at the previous time slice (note that this is strictly equivalent to a fully connected feedback loop on the hidden layer). The second pool of input units represents the current element

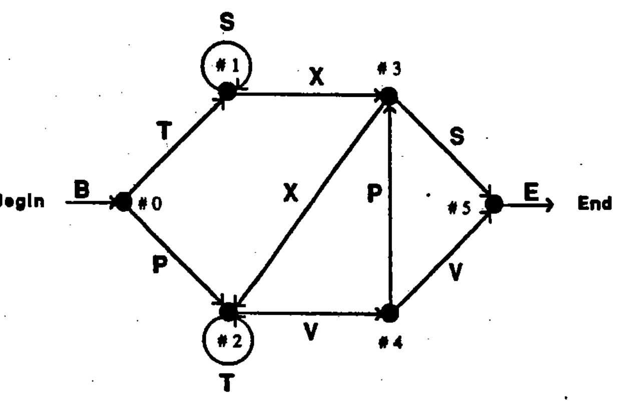

Figure 3. The finite-state grammar used by Reber (1976).

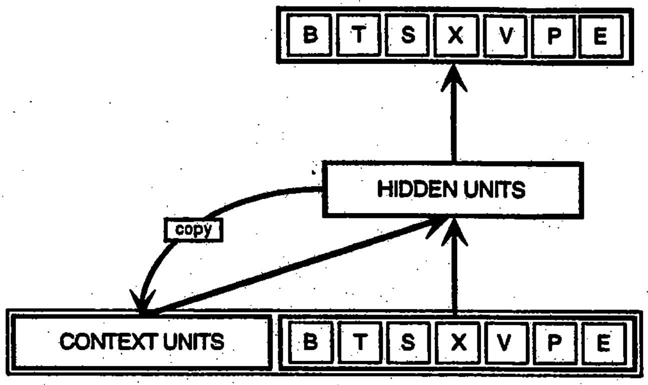

Figure 4. General architecture of the network.

of the string. On each trial, the network is presented with an element of the string, and is supposed to produce the next element on the output layer. In both the input and the output layers, letters are represented by the activation of a single unit. Five units therefore code for the five different possible letters in each of these two layers. In addition, two units code for *begin* and *end* bits. These two bits are needed so that the network can be trained to predict the first element and the end of a string (although only one *transition* bit is strictly necessary). In this first experiment, the number of hidden units was set to 3. Other values will be reported as appropriate.

2.3. *Coding of the strings*

A string of **n** letters is coded as a series of **n + 1** training patterns. Each pattern consists of two input vectors and one target vector. The target vector is a seven-bit vector representing element **t + 1** of the string. The two input vectors are:

- A three-bit vector representing the activation of the hidden units at time **t − 1**, and
- A seven-bit vector representing element **t** of the string.

2.4. *Training*

On each of 60,000 training trials, a string was generated from the grammar, starting with the 'B.' Successive arcs were then selected randomly from the two possible continuations, with a probability of 0.5. Each letter was then presented sequentially to the network. The activations of the context units were reset to 0 at the beginning of each string. After each letter, the error between the network's prediction and the *actual successor* specified by the string was computed and back-propagated. The 60,000 randomly generated strings ranged from 3 to 30 letters (mean: 7, sd: 3.3)[1].

2.5. Performance

Figure 5 shows the state of activation of all the units in the network, after training, when the start symbol is presented (here the letter 'B'—for begin). Activation of the output units indicate that the network is predicting two possible successors, the letters 'P' and 'T.' Note that the best possible prediction always activates two letters on the output layer except when the end of the string is predicted. Since during training 'P' and 'T' followed the start symbol equally often, each is activated partially in order to minimize error. Figure 6 shows the state of the network at the next time step in the string 'BTXXVV.' The pattern of activation on the context units is now a copy of the pattern generated previously on the hidden layer. The two successors predicted are 'X' and 'S.'

The next two figures illustrate how the network is able to generate different predictions when presented with two instances of the same letter on the input layer in different contexts. In Figure 7a, when the letter 'X' immediately follows 'T,' the network predicts again 'S' and 'X' appropriately. However, as Figure 7b shows, when a second 'X' follows, the prediction changes radically as the network now expects 'T' or 'V.' Note that if the network were not provided with a copy of the previous pattern of activation on the hidden layer, it would activate the four possible successors of the letter 'X' in both cases.

```
String     B t x x v v
           B  T  S  P  X  V  E
Output     00 53 00 38 01 02 00
Hidden           01 00 10
Context          00 00 00
           B  T  S  P  X  V  E
Input     100 00 00 00 00 00 00
```

Figure 5. State of the network after presentation of the 'Begin' symbol (following training). Activation values are internally in the range 0 to 1.0 and are displayed on a scale from 0 to 100. The capitalized bold letter indicates which letter is currently being presented on the input layer.

```
String     b T x x v v
           B  T  S  P  X  V  E
Output     00 01 39 00 56 00 00
Hidden           84 00 28
Context          01 00 10
           B  T  S  P  X  V  E
Input      00100 00 00 00 00 00
```

Figure 6. State of the network after presentation of an initial 'T.' Note that the activation pattern on the context layer is identical to the activation pattern on the hidden layer at the previous time step.

```
String     b t X x v v
           B  T  S  P  X  V  E
Output     00 04 44 00 37 07 00
Hidden           74 00 93
Context          84 00 28
           B  T  S  P  X  V  E
Input      00 00 00 00 100 00 00
```

a

```
String     b t x X v v
           B  T  S  P  X  V  E
Output     00 50 01 01 00 55 00
Hidden           06 09 99
Context          74 00 93
           B  T  S  P  X  V  E
Input      00 00 00 00 100 00 00
```

b

Figure 7. a) State of the network after presentation of the first 'X.' b) State of the network after presentation of the second 'X.'

In order to test whether the network would generate similarly good predictions after every letter of any grammatical string, we tested its behavior on 20,000 strings derived randomly from the grammar. A prediction was considered accurate if, for every letter in a given string, activation of its successor was above 0.3. If this criterion was not met, presentation of the string was stopped and the string was considered 'rejected.' With this criterion, the network correctly 'accepted' all of the 20,000 strings presented.

We also verified that the network did not accept ungrammatical strings. We presented the network with 130,000 strings generated from the same pool of letters but in a random manner—i.e., mostly 'non-grammatical.' During this test, the network is first presented with the 'B' and one of the five letters or 'E' is then selected at random as a successor. If that letter is predicted by the network as a legal successor (i.e., activation is above 0.3 for the corresponding unit), it is then presented to the input layer on the next time step, and another letter is drawn at random as its successor. This procedure is repeated as long as each letter is predicted as a legal successor until 'E' is selected as the next letter. The procedure is interrupted as soon as the actual successor generated by the random procedure is *not* prediced by the network, and the string of letters is then considered 'rejected.' As in the previous test, the string is considered 'accepted' if all its letters have been predicted as possible continuations up to 'E.' Of the 130,000 strings, 0.2% (260) happened to be grammatical, and 99.7% were non-grammatical. The network performed flawlessly, accepting all the grammatical strings and rejecting all the others. In other words, for all non-grammatical strings, when the first non-grammatical letter was presented to the network its activation on the output layer at the previous step was less than 0.3 (i.e., it was *not* predicted as a successor of the previous—grammatically acceptable—letter).

Finally, we presented the network with several extremely long strings such as:

'BTSSSSSSSSSSSSSSSSSSSSSSXXVPXVPXVPXVPXVPXVPXVPXVPXVPXTT
TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTVPXVPXVPXVPXVPXVPS'

and observed that, at every step, the network correctly predicted both legal successors and no others.

Note that it is possible for a network with more hidden units to reach this performance criterion with much less training. For example, a network with 15 hidden units reached criterion after 20,000 strings were presented. However, activation values on the output layer are not as clearly contrasted when training is less extensive. Also, the selection of a threshold of 0.3 is not completely arbitrary. The activation of output units is related to the frequency with which a particular letter appears as the successor of a given sequence. In the training set used here, this probability is 0.5. The activation of a legal successor would then be expected to be 0.5.[2] However, because of the use of a momentum term in the back propagation learning procedure, the activation of correct output units following training was occasionally below 0.5—sometimes as low as 0.3.

2.6. *Analysis of internal representations*

Obviously, in order to perform accurately, the network takes advantage of the representations that have developed on the hidden units which are copied back onto the context layer. At any point in the sequence, these patterns must somehow encode the position of the current input in the grammar on which the network was trained. One approach to understanding how the network uses these patterns of activation is to perform a cluster analysis. We recorded the patterns of activation on the hidden units following the presentation of each letter in a small random set of grammatical strings. The matrix of Euclidean distances between each pair of vectors of activation served as input to a cluster analysis program.[3] The graphical result of this analysis is presented in Figure 8A. Each leaf in the tree corresponds to a particular string, and the capitalized letter in that string indicates which letter has just been presented. For example, if the leaf is identified as 'pvPs,' 'P' is the current letter and its predecessors were 'P' and 'V' (the correct prediction would thus be 'X' or 'S').

From the figure, it is clear that activation patterns are grouped according to the different nodes in the finite state grammar; all the patterns that produce a similar prediction are grouped together, independently of the current letter. This grouping by similar predictions is apparent in Figure 8b, which represents an enlargement of the bottom cluster (cluster 5) of Figure 8a. One can see that this cluster groups patterns that result in the activation of the "End" unit: All the strings corresponding to these patterns end in 'V' or 'S' and lead to node #5 of the grammar, out of which "End" is the only possible successor. Therefore, when one of the hidden layer patterns is copied back onto the context layer, the network is provided with information about the *current node*. That information is combined with input representing the *current letter* to produce a pattern on the hidden layer that is a representation of the *next node*. To a degree of approximation, the recurrent network behaves exactly like the finite state automaton defined by the grammar. It does not use a stack or registers to provide contextual information but relies instead on simple state transitions, just like a finite state machine. Indeed, the network's perfect performance on randomly generated grammatical and non-grammatical strings shows that it can be used as a finite state recognizer.

However, a closer look at the cluster analysis reveals that within a cluster corresponding to a particular node, patterns are further divided according to the path traversed before that node. For example, an examination of Figure 8b reveals that patterns ending by 'VV,' 'PS' and 'SXS' endings have been grouped separately by the analysis: they are more similar to each other than to the abstract prototypical pattern that would characterize the corresponding "node."[4] We can illustrate the behavior of the network with a specific example. When the first letter of the string 'BTX' is presented, the initial pattern on context units corresponds to node 0. This pattern together with the letter 'T' generates a hidden layer pattern corresponding to node 1. When that pattern is copied onto the context layer and the letter 'X' is presented, a new pattern corresponding to node 3 is produced on the hidden layer, and this pattern is in turn copied on the context units. If the network behaved *exactly* like a finite state automaton, the exact same patterns would be used during processing of the other strings 'BTSX' and 'BTSSX.' That behavior would be adequately captured by the transition network shown in Figure 9. However, since the cluster analysis shows that slightly different patterns are produced by the substrings 'BT,' 'BTS' and 'BTSS,' Figure 10 is a more accurate description of the network's state transitions. As states 1, 1' and 1" on the one hand and 3, 3' and 3" on the other are nevertheless very similar to each other, the finite state machine that the network implements can be said to approximate the idealization of a finite state automaton corresponding exactly to the grammar underlying the exemplars on which it has been trained.

However, we should point out that the close correspondence between representations and function obtained for the recurrent network with three hidden units is rather the exception than the rule. With only three hidden units, representational resources are so scarce that backpropagation forces the network to develop representations that yield a prediction on the basis of the current node alone, ignoring contributions from the path. This situation precludes the development of different—redundant—representations for a particular node that typically occurs with larger numbers of hidden units. When redundant representations do develop, the network's behavior still converges to the theoretical finite state automaton—in the sense that it can still be used as a perfect finite state recognizer for strings generated from the corresponding grammar—but internal representations do not correspond to that idealization. Figure 11 shows the cluster analysis obtained from a network with 15 hidden units after training on the same task. Only nodes 4 and 5 of the grammar seem to be

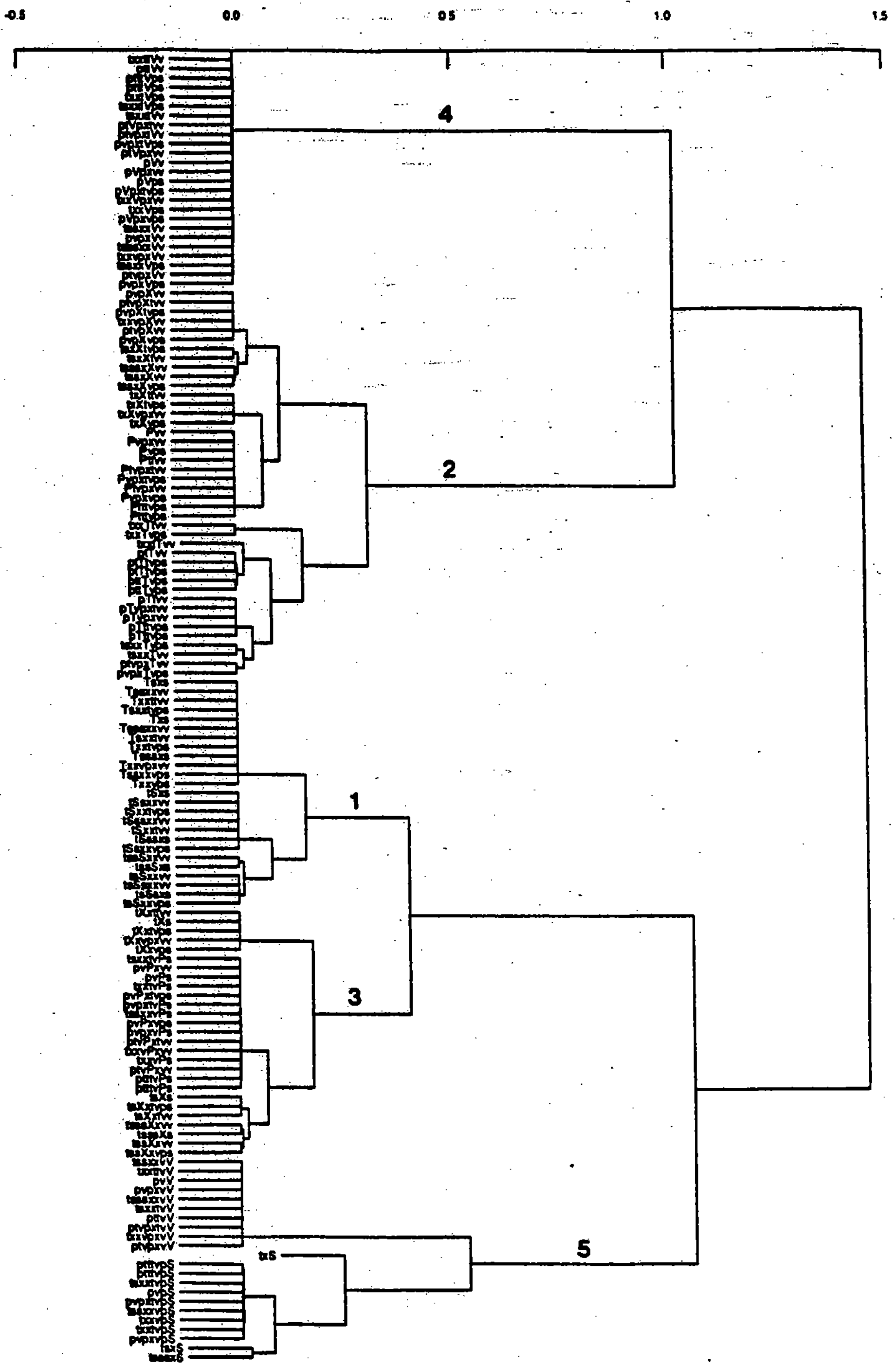

Figure 8A. Hierarchical Cluster Analysis of the H.U. activation patterns after 200,000 presentations from strings generated at random according to the Reber grammar (Three hidden units).

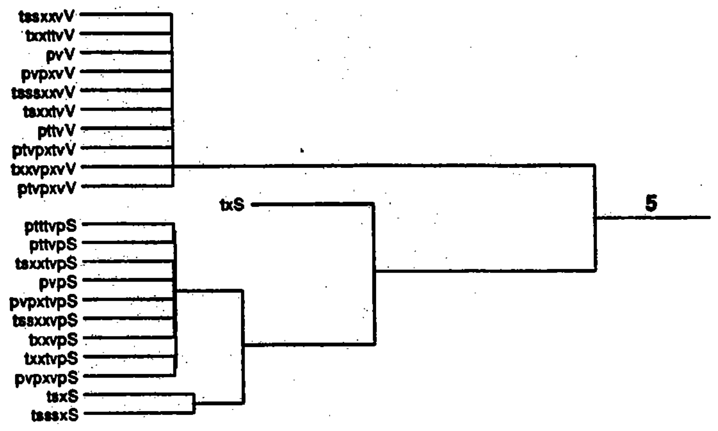

Figure 8B. An enlarged portion of Figure 8A, representing the bottom cluster (cluster 5). The proportions of the original figure have not necessarily been respected in this enlargement.

$$\begin{tikzcd}
 & \#1 \arrow[loop above, "S"] \arrow[r, "X"] & \#3 \\
\#0 \arrow[ur, "T"] & &
\end{tikzcd}$$

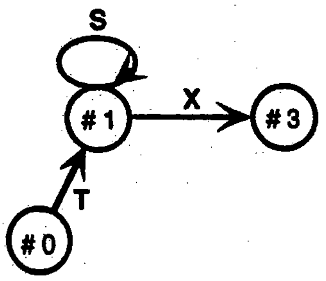

Figure 9. A transition network corresponding to the upper-left part of Reber's finite-state grammar.

$$\begin{tikzcd}
 & & & & \\
 & & \#1'' \arrow[ur] \arrow[r, "X"] & \#3'' & \\
 & \#1' \arrow[ur, "S"] \arrow[r, "X"] & \#3' & & \\
 \#1 \arrow[ur, "S"] \arrow[r, "X"] & \#3 & & & \\
\#0 \arrow[u, "T"] & & & &
\end{tikzcd}$$

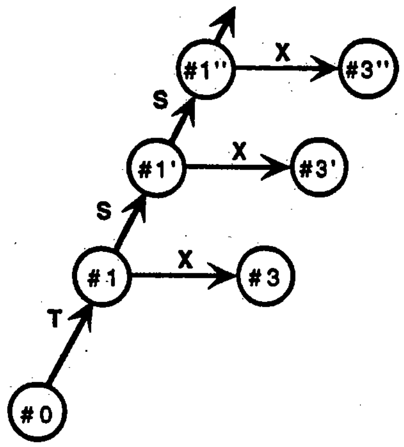

Figure 10. A transition network illustrating the network's true behavior.

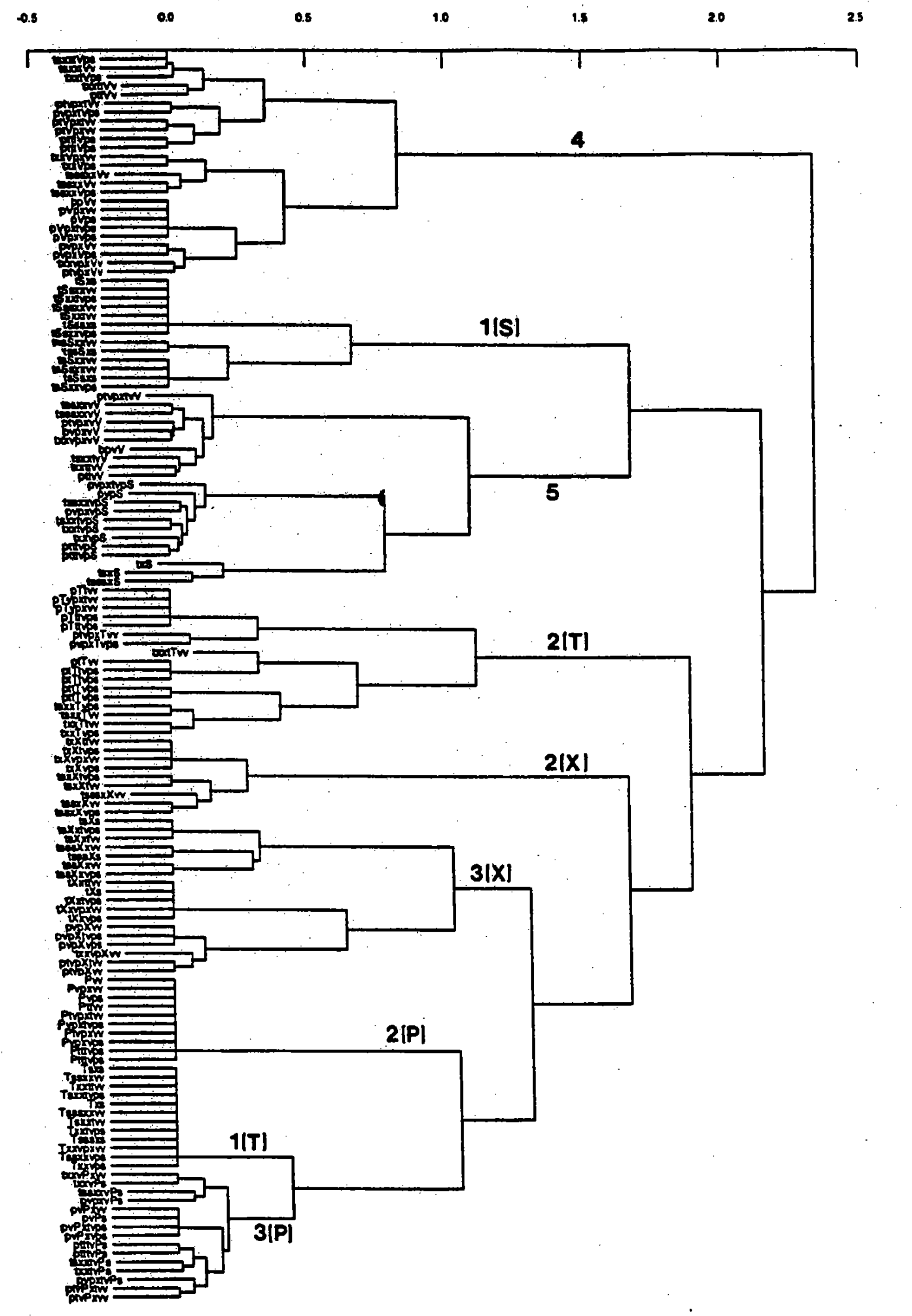

Figure 11. Hierarchical Cluster Analysis of the H.U. activation patterns after 200,000 presentations from strings generated at random according to the Reber grammar (Fifteen hidden units).

represented by a unique 'prototype' on the hidden layer. Clusters corresponding to nodes 1, 2 and 3 are divided according to the preceding arc. Information about arcs is not relevant to the prediction task and the different clusters corresponding to the a single node play a redundant role.

Finally, preventing the development of redundant representations may also produce adverse effects. For example, in the Reber grammar, predictions following nodes 1 and 3 are identical ('X or S'). With some random sets of weights and training sequences, networks with only three hidden units occasionally develop almost identical representations for nodes 1 and 3, and are therefore unable to differentiate the first from the second 'X' in a string.

In the next section we examine a different type of training environment, one in which information about the path traversed becomes relevant to the prediction task.

3. Discovering and using path information

The previous section has shown that simple recurrent networks can learn to encode the nodes of the grammar used to generate strings in the training set. However, this training material does not require information about arcs or sequences of arcs—the 'path'—to be maintained. How does the network's performance adjust as the training material involves more complex and subtle temporal contingencies? We examine this question in the following section, using a training set that places many additional constraints on the prediction task.

3.1. Material

The set of strings that can be generated from the grammar is finite for a given length. For lengths 3 to 8, this amounts to 43 grammatical strings. The 21 strings shown in Figure 12 were selected and served as training set. The remaining 22 strings can be used to test generalization.

The selected set of strings has a number of interesting properties with regard to exploring the network's performance on subtle temporal contingencies:

- As in the previous task, identical letters occur at different points in each string, and lead to different predictions about the identity of the successor. No stable prediction is therefore associated with any particular letter, and it is thus necessary to encode the position, or the node of the grammar.

TSXS	TXS	TSSSXS
TSSXXVV	TSSSXXVV	TXXVPXVV
TXXTTVV	TSXXTVV	TSSXXVPS
TSXXTVPS	TXXTVPS	TXXVPS
PVV	PTTVV	PTVPXVV
PVPXVV	PTVPXTVV	PVPXVPS
PTVPS	PVPXTVPS	PTTTVPS

Figure 12. The 21 grammatical strings of length 3 to 8.

- In this limited training set, length places additional constraints on the encoding because the possible predictions associated with a particular node in the grammar are dependent on the length of the sequence. The set of possible letters that follow a particular node depends on how many letters have already been presented. For example, following the sequence 'TXX' both 'T' and 'V' are legal successors. However, following the sequence 'TXXVPX,' 'X' is the sixth letter and only 'V' would be a legal successor. This information must therefore also be somehow represented during processing.
- Subpatterns occurring in the strings are not all associated with their possible successors equally often. Accurate predictions therefore require that information about the identity of the letters that have already been presented be maintained in the system, i.e., the system must be sensitive to the frequency distribution of subpatterns in the training set. This amounts to encoding the full path that has been traversed in the grammar.

These features of the limited training set obviously make the prediction task much more complex than in the previous simulation.

3.2. Network architecture

The same general network architecture was used for this set of simulations. The number of hidden units was arbitrarily set to 15.

3.3. Performance

The network was trained on the 21 different sequences (a total of 130 patterns) until the total sum squared error (*tss*) reached a plateau with no further improvements. This point was reached after 2000 epochs and *tss* was 50. Note that *tss* cannot be driven much below this value, since most partial sequences of letters are compatible with 2 different successors. At this point, the network correctly predicts the possible successors of each letter, and distinguishes between different occurrences of the same letter—like it did in the simulation described previously. However, the network's performance makes it obvious that many additional constraints specific to the limited training set have been encoded. Figure 13a shows that the network expects a 'T' or a 'V' after a first presentation of the second 'X' in the grammar.

Contrast these predictions with those illustrated in Figure 13b, which shows the state of the network after a *second* presentation of the second 'X': Although the same node in the grammar has been reached, and 'T' and 'V' are again possible alternatives, the network now predicts only 'V.'

Thus, the network has successfully learned that an 'X' occurring late in the sequence is never followed by a 'T'—a fact which derives directly from the maximum length constraint of 8 letters.

It could be argued that that network simply learned that when 'X' is preceded by 'P' it cannot be followed by 'T', and thus relies only on the preceding letter to make that distinction. However, the story is more complicated than this.

```
String     b t x X v p x v v
           B  T  S  P  X  V  E
Output     00 49 00 00 00 50 00
Hidden     27 89 02 16 99 43 01 06 04 18 99 81 95 18 01
Context    01 18 00 41 95 01 60 59 05 06 84 99 19 05 00
           B  T  S  P  X  V  E
Input      00 00 00 00 100 00 00
```

a

```
String     b t x x v p X v v
           B  T  S  P  X  V  E
Output     00 03 00 00 00 95 00
Hidden     00 85 03 85 31 00 72 19 31 03 93 99 61 05 00
Context    01 07 05 90 93 04 00 10 71 40 99 16 90 05 82
           B  T  S  P  X  V  E
Input      00 00 00 00 100 00 00
```

b

Figure 13. a) State of the network after presentation of the second 'X.' b) State of the network after a second presentation of the second 'X.'

In the following two cases, the network is presented with the first occurrence of the letter 'V.' In the first case, 'V' is preceded by the sequence 'tssxx,' while in the second case, it is preceded by 'tsssxx.' The difference of a single 'S' in the sequence—which occurred 5 presentations before—results in markedly different predictions when 'V' is presented (Figures 14a and 14b).

The difference in predictions can be traced again to the length constraint imposed on the strings in the limited training set. In the second case, the string spans a total of 7 letters when 'V' is presented, and the only alternative compatible with the length constraint is a second 'V' and the end of the string. This is not true in the first case, in which both 'VV' and 'VPS' are possible endings.

Thus, it seems that the representation developed on the context units encodes more than the immediate context—the pattern of activation could include a full representation of the path traversed so far. Alternatively, it could be hypothesized that the context units encode only the preceding letter and a counter of how many letters have been presented.

```
String     b t s s x x V v
           B  T  S  P  X  V  E
Output     00 00 00 54 00 48 00
Hidden     44 98 30 84 99 82 00 47 00 09 41 98 13 02 00
Context    89 90 01 01 99 70 01 03 02 10 99 95 85 21 00
           B  T  S  P  X  V  E
Input      00 00 00 00 00100 00
```

a

```
String     b t s s s x x V v
           B  T  S  P  X  V  E
Output     00 00 00 02 00 97 00
Hidden     56 99 48 93 99 85 00 22 00 10 77 97 30 03 00
Context    54 67 01 04 99 59 07 09 01 06 98 97 72 16 00
           B  T  S  P  X  V  E
Input      00 00 00 00 00100 00
```

b

Figure 14. Two presentations of the first ‘V_1’ with slightly different paths.

In order to understand better the kind of representations that encode sequential context we performed a cluster analysis on all the hidden unit patterns evoked by each sequence. Each letter of each sequence was presented to the network and the corresponding pattern of activation on the hidden layer was recorded. The Euclidean distance between each pair of patterns was computed and the matrix of all distances was provided as input to a cluster analysis program.

The resulting analysis is shown in Figure 15. We labeled the arcs according to the letter being presented (the ‘current letter’) and its position in the Reber grammar. Thus ‘V_1’ refers to the first ‘V’ in the grammar and ‘V_2’ to the second ‘V’ which immediately precedes the end of the string. ‘Early’ and ‘Late’ refer to whether the letter occurred early or late in the sequence (for example in ‘PT. .’ ‘T_2’ occurs early; in ‘PVPXT. .’ it occurs late). Finally, in the left margin we indicated what predictions the corresponding patterns yield on the output layer (e.g., the hidden unit pattern generated by ‘B’ predicts ‘T’ or ‘P’).

From the figure, it can be seen that the patterns are grouped according to three distinct principles: (1) according to similar predictions, (2) according to similar letters presented

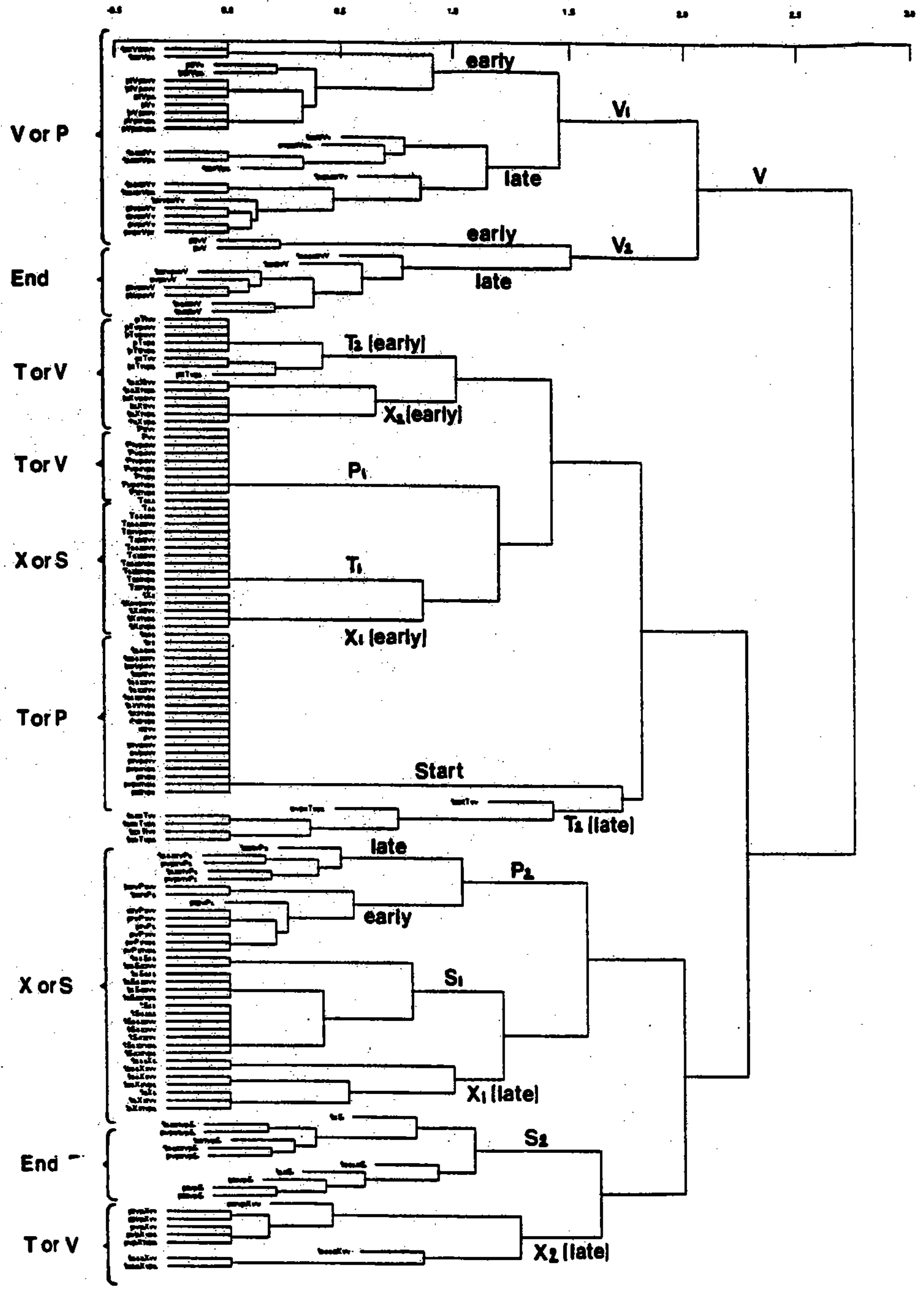

Figure 15. Hierarchical cluster analysis of the H.U. activation patters after 2000 epochs of training on the set of 21 strings.

on the input units, and (3) according to similar paths. These factors do not necessarily overlap since several occurrences of the same letter in a sequence usually implies different predictions and since similar paths also lead to different predictions depending on the current letter.

For example, the top cluster in the figure corresponds to all occurrences of the letter 'V' and is further subdivided among 'V_1' and 'V_2.' The 'V_1' cluster is itself further divided between groups where 'V_1' occurs early in the sequence (e.g., 'pV. . .') and groups where it occurs later (e.g., 'tssxxV. . .'). Note that the division according to the path does not necessarily correspond to different predictions. For example, 'V_2' always predicts 'END' and always with maximum certainty. Nevertheless, sequences up to 'V_2' are divided according to the path traversed.

Without going into the details of the organization of the remaining clusters, it can be seen that they are predominantly grouped according to the predictions associated with the corresponding portion of the sequence and then further divided according to the path traversed up to that point. For example, 'T_2,' 'X_2' and 'P_1' all predict 'T or V,' 'T_1' and 'X_1' both predict 'X or S,' and so on.

Overall, the hidden units patterns developed by the network reflect two influences: a 'top-down' pressure to produce the correct ouptut, and a 'bottom-up' pressure from the successive letters in the path which modifies the activation pattern independently of the output to be generated.

The top-down force derives directly from the back-propagation learning rule. Similar patterns on the output units tend to be associated with similar patterns on the hidden units. Thus, when two different letters yield the same prediction (e.g., 'T_1' and 'X_1'), they tend to produce similar hidden layer patterns. The bottom-up force comes from the fact that, nevertheless, each letter presented with a particular context can produce a characteristic mark or *shading* on the hidden unit pattern (see Pollack, 1989, for a further discussion of error-driven and recurrence-driven influences on the development of hidden unit patterns in recurrent networks). The hidden unit patterns are not truly an 'encoding' of the input, as is often suggested, but rather an encoding of the *association* between a particular input and the relevant prediction. It really reflects an influence from both sides.

Finally, it is worth noting that the very specific internal representations acquired by the network are nonetheless sufficiently abstract to ensure good generalization. We tested the network on the remaining untrained 22 strings of length 3 to 8 that can be generated by the grammar. Over the 165 predictions of successors in these strings, the network made an incorrect prediction (activation of an incorrect successor > 0.05) in only 10 cases, and it failed to predict one of two continuations consistent with the grammar and length constraints in 10 other cases.

3.4. Finite state automata and graded state machines

In the previous sections, we have examined how the recurrent network encodes and uses information about meaningful subsequences of events, giving it the capacity to yield different outputs according to some specific traversed path or to the length of strings. However, the network does not use a separate and explicit representation for non-local properties of the strings such as length. It only learns to associate different predictions to a subset of states; those that are associated with a more restricted choice of successors. Again there are not stacks or registers, and each different prediction is associated to a specific state on the context units. In that sense, the recurrent network that has learned to master this task still behaves like a finite-state machine, although the training set involves non-local constraints that could only be encoded in a very cumbersome way in a finite-state *grammar*.

We usually do not think of finite state automata as capable of encoding non-local information such as length of a sequence. Yet, finite state machines have in principle the same computational power as a Turing machine with a finite tape and they can be designed to respond adequately to non-local constraints. Recursive or Augmented transition networks and other Turing-equivalent automata are preferable to finite state machines because they spare memory and are modular—and therefore easier to design and modify. However, the finite state machines that the recurrent network seems to implement have properties that set them apart from their traditional counterparts:

- For tasks with an appropriate structure, recurrent networks develop their own state transition diagram, sparing this burden to the designer.
- The large amount of memory required to develop different representations for every state needed is provided by the representational power of hidden layer patterns. For example, 15 hidden units with four possible values—e.g., 0, .25, .75, 1—can support more than one billion different patterns (4^{15} = 107,374,1824).
- The network implementation remains capable of performing similarity-based processing, making it somewhat noise-tolerant (the machine does not 'jam' if it encounters an undefined state transition and it can recover as the sequence of inputs continues), and it remains able to generalize to sequences that were not part of the training set.

Because of its inherent ability to use *graded* rather than finite states, the SRN is definitely not a finite state machine of the usual kind. As we mentioned above, we have come to consider it as an exemplar of a new class of automata that we call *Graded State Machines*.

In the next section, we examine how the SRN comes to develop appropriate internal representations of the temporal context.

4. Learning

We have seen that the SRN develops and learns to use compact and effective representations of the sequences presented. These representations are sufficient to disambiguate identical cues in the presence of context, to code for length constraints and to react appropriately to atypical cases.[5] How are these representations discovered?

As we noted earlier, in an SRN, the hidden layer is presented with information about the current letter, but also—on the context layer—with an encoding of the relevant features of the previous letter. Thus, a given hidden layer pattern can come to encode information about the relevant features of two consecutive letters. When this pattern is fed back on the context layer, the new pattern of activation over the hidden units can come to encode information about three consecutive letters, and so on. In this manner, the context layer patterns can allow the network to maintain prediction-relevant features of an entire sequence.

As discussed elsewhere in more detail (Servan-Schreiber, Cleeremans & McClelland, 1988, 1989), learning progresses through three qualitatively different phases. During a first phase, the network tends to ignore the context information. This is a direct consequence of the fact that the patterns of activation on the hidden layer—and hence the context layer—are continuously changing from one epoch to the next as the weights from the input units (the letters) to the hidden layer are modified. Consequently, adjustments made to the weights from the context layer to the hidden layer are inconsistent from epoch to epoch and cancel each other. In contrast, the network is able to pick up the stable association between each *letter* and all its possible successors. For example, after only 100 epochs of training, the response pattern generated by 'S_1' and the corresponding output are almost identical to the pattern generated by 'S_2', as Figures 16a and 16b demonstrate. At the end of this phase, the network thus predicts all the successors of each letter in the grammar, independently of the *arc* to which each letter corresponds.

```
Epoch     100
String    bSsxxvps
           B  T  S  P  X  V  E
Output    00 00 36 00 33 16 17
Hidden    45 24 47 26 36 23 55 22 22 26 22 23 30 30 33
Context   44 22 56 21 36 22 64 16 13 23 20 16 25 21 40
           B  T  S  P  X  V  E
Input     00 00 100 00 00 00 00
```

a

```
Epoch     100
String    bssxxvpS
           B  T  S  P  X  V  E
Output    00 00 37 00 33 16 17
Hidden    45 24 47 25 36 23 56 22 21 25 21 22 29 30 32
Context   42 29 53 24 32 27 61 25 16 33 25 23 28 27 41
           B  T  S  P  X  V  E
Input     00 00 100 00 00 00 00
```

b

Figure 16. a) Hidden layer and output generated by the presentation of the *first* S in a sequence after 100 epochs of training. b) Hidden layer and output patterns generated by the presentation of the *second* S in a sequence after 100 epochs of training.

In a second phase, patterns copied on the context layer are now represented by a unique code designating which letter preceded the current letter, and the network can exploit this stability of the context information to start distinguishing between different occurrences of the same letter—different arcs in the grammar. Thus, to continue with the above example, the response elicited by the presentation of an 'S_1' would progressively become different from that elicited by an 'S_2.'

Finally, in a third phase, small differences in the context information that reflect the occurrence of previous elements can be used to differentiate position-dependent predictions resulting from length constraints. For example, the network learns to differentiate between 'tssxxV' which predicts either 'P' or 'V,' and 'tsssxxV' which predicts only 'V,' although both occurrences of 'V' correspond to the same arc in the grammar. In order to make this distinction, the pattern of activation on the context layer must be a representation of the entire path rather than simply an encoding of the previous letter.

Naturally, these three phases do not reflect sharp changes in the network's behavior over training. Rather, they are simply particular points in what is essentially a continuous process, during which the network progressively encodes increasing amounts of temporal context information to refine its predictions. It is possible to analyze this smooth progression towards better predictions by noting that these predictions converge towards the optimal conditional probabilities of observing a particular successor to the sequence presented up to that point. Ultimately, given sufficient training, the SRN's responses *would become* these optimal conditional probabilities (that is, the minima in the error function are located at those points in weight space where the activations equal the optimal conditional probabilities). This observation gives us a tool for anlayzing how the predictions change over time. Indeed, the conditional probability of observing a particular letter at any point in a sequence of inputs varies according to the number of preceding elements that have been encoded. For instance, since all letters occur twice in the grammar, a system basing its predictions on only the current element of the sequence will predict all the successors of the current letter, independently of the arc to which that element corresponds. If two elements of the sequence are encoded, the uncertainty about the next event is much reduced, since in many cases, subsequences of two letters are unique, and thus provide an unambiguous cue to its possible successors. In some other cases, subtle dependencies such as those resulting from length constraints require as much as 6 elements of temporal context to be optimally predictable.

Thus, by generating a large number of strings that have exactly the same statistical properties as those used during training, it is possible to estimate the conditional probabilities of observing each letter as the successor to each possible path of a given length. The *average* conditional probability (ACP) of observing a particular letter at every node of the grammar, after a given amount of temporal context (i.e., over all paths of a given length) can then be obtained easily by weighting each individual term appropriately. This analysis can be conducted for paths of any length, thus yielding a set of ACPs for each statistical order considered.[6] Each set of ACPs can then be used as the predictor variable in a regression analysis against the network's responses, averaged in a similar way. We would expect the ACPs based on short paths to be better predictors of the SRN's behavior early in training, and the ACPs based on longer paths to be better predictors of the SRN's behavior late in training, thus revealing the fact that, during training, the network learns to base its predictions on increasingly larger amounts of temporal context.

An SRN with fifteen hidden units was trained on the 43 strings of length 3 to 8 from the Reber grammar, in exactly the same conditions as described earlier. The network was trained for 1000 epochs, and its performance tested once before training, and every 50 epochs thereafter, for a total of 21 tests. Each test consisted of 1) freezing the connections, 2) presenting the network with the entire set of strings (a total of 329 patterns) once, and 3) recording its response to each individual input pattern. Next, the average activation of each response unit (i.e., each letter in the grammar) given 6 elements of temporal context was computed (i.e., after all paths of length 6 that are followed by that letter).

In a separate analysis, seven sets of ACPs (from order 0 to order 6) were computed in the manner described above. Each of these seven sets of ACPs was then used as the predictor variable in a regression analysis on each set of average activations produced by the network. These data are represented in Figure 17. Each point represents the percentage of variance explained in the network's behavior on a particular test by the ACPs of a particular statistical order. Points corresponding to the same set of ACPs are linked together, for a total of 7 curves, each corresponding to the ACPs of a particular order.

What the figure reveals is that the network's responses are approximating the conditional probabilities of increasingly higher statistical orders. Thus, before training, the performance of the network is best explained by the 0th order ACPs (i.e., the frequency of each letter in the training set). This is due to the fact that before training, the activations of the response units tend to be almost uniform, as do the 0th order ACPs. In the next two tests (i.e., at epoch 50 and epoch 100), the network's performance is best explained by the first-order ACPs. In other words, the network's predictions during these two tests were essentially based on paths of length 1. This point in training corresponds to the first phase of learning identified earlier, during which the network's responses do not distinguish between different occurrences of the same letter.

Soon, however, the network's performance comes to be better explained by ACPs of higher statistical orders. One can see the curves corresponding to the ACPs of order 2 and 3 progressively take over, thus indicating that the network is essentially basing its predictions on paths of length 2, then of length 3. At this point, the network has entered the second phase of learning, during which it now distinguishes between different occurrences of the same letter. Later in training, the network's behavior can be seen to be better captured by ACPs based on even longer paths, first of length 4, and finally, of length 5. Note that the network remains at that stage for a much longer period of time than for shorter ACPs. This reflects the fact that encoding longer paths is more difficult. At this point, the network has started to become sensitive to subtler dependencies such as length constraints, which require an encoding of the full path traversed so far. Finally, the curve corresponding to the ACPs of order 6 can be seen to raise steadily towards increasingly better fits, only to be achieved considerably later in training.

It is worth noting that there is a large amount of overlap between the percentage of variance explained by the different sets of ACPs. This is not surprising, since most of the sets of ACPs are partially correlated with each other. Even so, we see the successive correspondence to longer and longer temporal contingencies with more and more training.

In all the learning problems we examined so far, contingencies between elements of the sequence were relevant at each processing step. In the next section, we propose a detailed

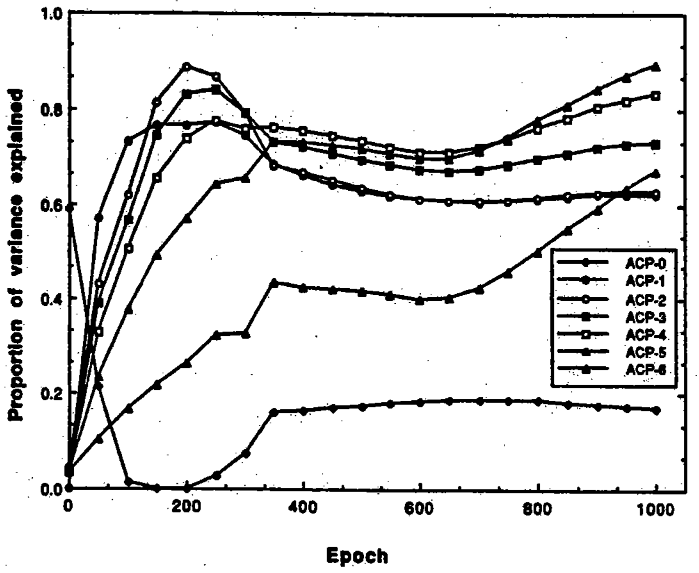

Figure 17. A graphic representation of the percentage of variance in the network's performance explained by average conditional probabilities of increasing statistical order (from 0 to 6). Each point represents the r-squared of a regression analysis using a particular set of average conditional probabilities as the predictor variable, and average activations produced by the network at a particular point in training as the dependent variable.

analysis of the constraints guiding the learning of more complex contingencies, for which information about the distant elements of the sequence to be maintained for several processing steps before they become useful.

5. Encoding non-local context

5.1. *Processing loops*

Consider the general problem of learning two arbitrary sequences of the same length and ending by two different letters. Under what conditions will the network be able to make a correct prediction about the nature of the last letter when presented with the penultimate letter? Obviously, the *necessary and sufficient condition* is that the internal representations associated with the penultimate letter are different (indeed, the hidden units patterns *have* to be different if different outputs are to be generated). Let us consider several different prototypical cases and verify if this condition holds:

PABC X and PABC V [1]

PABC X and TDEF V [2]

Clearly, problem [1] is impossible: as the two sequences are identical up to the last letter, there is simply no way for the network to make a different prediction when presented with the penultimate letter ('C' in the above example). The internal representations induced by the successive elements of the sequences will be strictly identical in both cases. Problem [2], on the other hand, is trivial, as the last letter is contingent on the penultimate letter ('X' is contingent on 'C'; 'V' on 'F'). There is no need here to maintain information available for several processing steps, and the different contexts set by the penultimate letters are sufficient to ensure that different predictions can be made for the last letter. Consider now problem [3].

PSSS P and TSSS T [3]

As can be seen, the presence of a final 'T' is contingent on the presence of an initial 'T'; a final 'P' on the presence of an initial 'P.' The shared 'S's do not supply any relevant information for disambiguating the last letter. Moreover, the predictions the network is required to make in the course of processing are identical in both sequences up to the last letter.

Obviously, the only way for the network to solve this problem is to develop *different internal representations* for *every* letter in the two sequences. Consider the fact that the network is required to make different predictions when presented with the last 'S.' As stated earlier, this will only be possible if the input presented at the penultimate time step produces different internal representations in the two sequences. However, this necessary difference cannot be due to the last 'S' itself, as it is presented in both sequences. Rather, the only way for different internal representations to arise when the last 'S' is presented is when the context pool holds different patterns of activation. As the context pool holds a copy of the internal representations of the previous step, these representations must themselves be different. Recursively, we can apply the same reasoning up to the first letter. The network must therefore develop a different representation for all the letters in the sequence. Are initial different letters a sufficient condition to ensure that each letter in the sequences will be associated with different internal representations? The answer is twofold.

First, note that developing a different internal representation for each letter (including the different instances of the letter 'S') is provided *automatically* by the recurrent nature of the architecture, even without any training. Successive presentations of identical elements to a recurrent network generate different internal representations at each step because the context pool holds different patterns of activity at each step. In the above example, the first letters will generate different internal representations. On the following step, these patterns of activity will be fed back to the network, and induce different internal representations again. This process will repeat itself up to the last 'S,' and the network will therefore find itself in a state in which it is potentially able to correctly predict the last letter of the two sequences of problem [3]. Now, there is an important caveat to this observation. Another fundamental property of recurrent networks is convergence towards an attractor state when a long sequence of idential elements are presented. Even though, initially, different patterns of activation are produced on the hidden layer for each 'S' in a sequence of 'S's, eventually the network converges towards a stable state in which every new presentation of the same input produces the same pattern of activation on the hidden layer. The number of iterations required for the network to converge depends on the number of hidden units. With more degrees of freedom, it takes more iterations for the network to settle. Thus, increasing the number of hidden units provides the network with an increased *architectural* capacity of maintaining differences in its internal representations when the input elements are identical.[7]

Second, consider the way back-propagation interacts with this natural process of maintaining information about the first letter. In problem [3], the predictions in each sequence are identical up to the last letter. As similar outputs are required on each time step, the weight adjustment procedure pushes the network into developing *identical* internal representations at each time step and for the two sequences—therefore going in the opposite direction than is required. This 'homogenizing' process can strongly hinder learning, as will be illustrated below.

From the above reasoning, we can infer that optimal learning conditions exist when both contexts and predictions are different in each sequence. If the sequences share identical sequences of predictions—as in problem [3]—the process of maintaining the differences between the internal representations generated by an (initial) letter can be disrupted by back-propagation itself. The very process of learning to predict correctly the intermediate shared elements of the sequence can even cause the total error to rise sharply in some cases after an initial decrease. Indeed, the more training the network gets on these intermediate elements, the more likely it is that their internal representations will become identical, thereby completely eliminating initial slight differences that could potentially be used to disambiguate the last element. Further training can only worsen this situation.[8] Note that in this sense back-propagation in the recurrent network is not guaranteed to implement gradient descent. Presumably, the ability of the network to resist the 'homogenization' induced by the learning algorithm will depend on its representational power—the number of hidden units available for processing. With more hidden units, there is also less pressure on each unit to take on specified activation levels. Small but crucial differences in activation levels will therefore be allowed to survive at each time step, until they finally become useful at the penultimate step.

To illustrate this point, a network with fifteen hidden units was trained on the two sequences of problem [3]. The network is able to solve this problem very accurately after approximately 10,000 epochs of training on the two patterns. Learning proceeds smoothly until a very long plateau in the error is reached. This plateau corresponds to a learning phase during which the weights are adjusted so that the network can take advantage of the small differences that remain in the representations induced by the last 'S' in the two strings in order to make accurate predictions about the identity of the last letter. These slight differences are of course due to the different context generated after presentation of the first letter of the string.

To understand further the relation between network size and problem size, four different networks (with 7, 15, 30 or 120 hidden units) were trained on each of four different versions of problem [3] (with 2, 4, 6 or 12 intermediate elements). As predicted, learning was faster when the number of hidden units was larger. There was an interaction between the size of the network and the size of the problem: adding more hidden units was of little influence when the problem was small, but had a much larger impact for larger numbers of intervening 'S's. We also observed that the relation between the size of the problem and the number of epochs to reach a learning criterion was exponential for all network sizes. These results suggest that for relatively short embedded sequences of identical letters, the difficulties encountered by the simple recurrent network can be alleviated by increasing the number of hidden units. However, beyond a certain range, maintaining different representations across the embedded sequence becomes exponentially difficult (see also Allen, 1988 and Allen, 1990 for a discussion of how recurrent networks hold information across embedded sequences).

An altogether different approach to the question can also be taken. In the next section, we argue that some sequential problems may be less difficult than problem [3]. More precisely, we will show how very slight adjustments to the predictions the network is required to make in otherwise identical sequences can greatly enhance performance.

5.2. *Spanning embedded sequences*

The previous example is a limited test of the network's ability to preserve information during processing of an embedded sequence in several respects. Relevant information for making a prediction about the nature of the last letter is at a constant distance across all patterns and elements inside the embedded sequence are all identical. To evaluate the performance of the SRN on a task that is more closely related to natural language situations, we tested its ability to maintain information about long-distance dependencies on strings generated by the grammar shown in Figure 18.

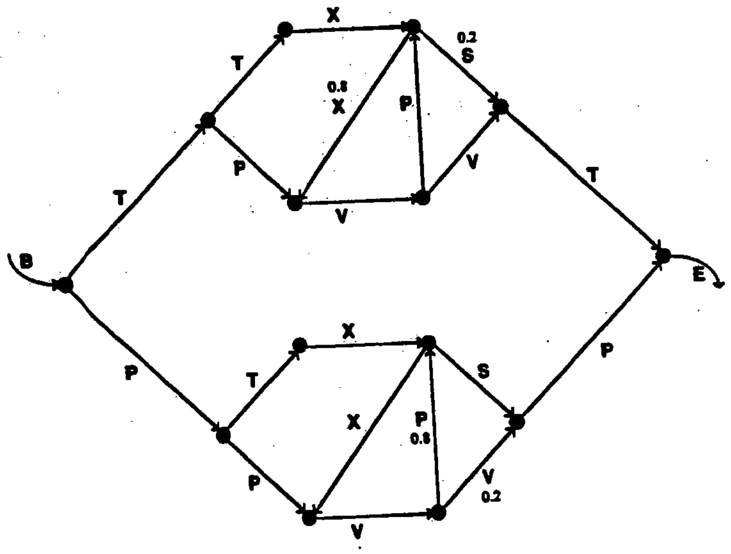

Figure 18. A complex finite-state grammar involving an embedded clause. The last letter is contingent on the first one, and the intermediate structure is shared by the two branches of the grammar. Some arcs in the asymmetrical version have different transitional probabilities in the top and bottom sub-structure as explained in the text.

If the first letter encountered in the string is a 'T,' the last letter of the string is also a 'T.' Conversely, if the first letter is a 'P,' the last letter is also a 'P.' In between these matching letters, we interposed almost the same finite state grammar that we had been using in previous experiments (Reber's) to play the role of an embedded sentence. We modified Reber's grammar by eliminating the 'S' loop and the 'T' loop in order to shorten the average length of strings.

In a first experiment we trained the network on strings generated from the finite state grammar with the same probabilities attached to corresponding arcs in the bottom and top version of Reber's grammar. This version was called the 'symmetrical grammar': contingencies inside the sub-grammar are the same independently of the first letter of the string, and all arcs had a probability of 0.5. The average length of strings was 6.5 (sd = 2.1).

After training, the performance of the network was evaluated in the following way: 20,000 strings generated from the symmetrical grammar were presented and for each string we looked at the relative activation of the predictions of 'T' and 'P' upon exit from the sub-grammar. If the Luce ratio for the prediction with the highest activation was below 0.6, the trial was treated as a 'miss' (i.e., failure to predict one or the other distinctively).[9] If the Luce ratio was greater or equal to 0.6 and the network predicted the correct alternative, a 'hit' was recorded. If the incorrect alternative was predicted, the trial was treated as an 'error.' Following training on 900,000 exemplars, performance consisted of 75% hits, 6.3% errors, and 18.7% misses. Performance was best for shorter embeddings (i.e., 3 to 4 letters) and deteriorated as the length of the embedding increased (see Figure 19).

However, the fact that contingencies inside the embedded sequences are similar for both sub-grammars greatly raises the difficulty of the task and does not necessarily reflect the nature of natural language. Consider a problem of number agreement illustrated by the following two sentences:

The dog *that chased the cat* is very playful

The dogs *that chased the cat* are very playful

We would contend that expectations about concepts and words forthcoming in the embedded sentence are different for the singular and plural forms. For example, the embedded clauses require different agreement morphemes—chases vs. chase—when the clause is in the present tense, etc. Furthermore, even after the same word has been encountered in both cases (e.g., 'chased'), expectations about possible successors for that word would remain different (e.g., a single dog and a pack of dogs are likely to be chasing different things). As we have seen, if such differences in predictions do exist the network is more likely to maintain information relevant to non-local context since that information is relevant at several intermediate steps.

To illustrate this point, in a second experiment, the same network—with 15 hidden units—was trained on a variant of the grammar shown in Figure 18. In this 'asymmetrical' version, the second X arc has a 0.8 probability of being selected during training, whereas in the bottom sub-grammar, the second P arc had a probability of 0.8 of being selected. Arcs stemming from all other nodes had the same probability attached to them in both sub-grammars. The mean length of strings generated from this asymmetrical version was 5.8 letters (sd = 1.3).

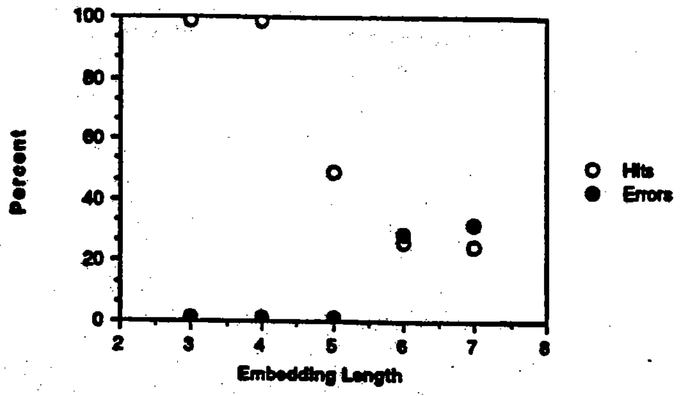

Figure 19. Percentage of hits and errors as a function of embedding length. All the cases with 7 or more letters in the embedding were grouped together.

Following training on the asymmetrical version of the grammar the network was tested with strings generated from the *symmetrical* version. Its performance level rose to 100% hits. It is important to note that performance of this network cannot be attributed to a difference in statistical properties of the test strings between the top and bottom subgrammars—such as the difference present during training—since the testing set came from the *symmetrical* grammar. Therefore, this experiment demonstrates that the network is better able to preserve information about the predecessor of the embedded sequence across identical embeddings as long as the ensemble of *potential* pathways is differentiated during training. Furthermore, differences in potential pathways may be only statistical and, even then, rather small. We would expect even greater improvements in performance if the two subgrammars included a set of non-overlapping sequences in addition to a set of sequences that are identical in both.

It is interesting to compare the behavior of the SRN on this embedding task with the corresponding FSA that could process the same strings. The FSA would have the structure of Figure 18. It would only be able to process the strings successfully by having two distinct copies of all the states between the initial letter in the string and the final letter. One copy is used after an initial P, the other is used after an initial T. This is inefficient since the embedded material is the same in both cases. To capture this similarity in a simple and elegant way, it is necessary to use a more powerful machine such as a recursive transition network. In this case, the embedding is treated as a subroutine which can be "called" from different places. A return from the call ensures that the grammar can correctly predict whether a T or a P will follow. This ability to handle long distance dependencies without duplication of the representation of intervening material lies at the heart of the arguments that have lead to the use of recursive formalisms to represent linguistic knowledge.

But the graded characteristics of the SRN allows the processing of embedded material as well as the material that comes after the embedding, without duplicating the representation of intervening material, and without actually making a subroutine call. The states of the SRN can be used *simultaneously* to indicate where the network is inside the embedding

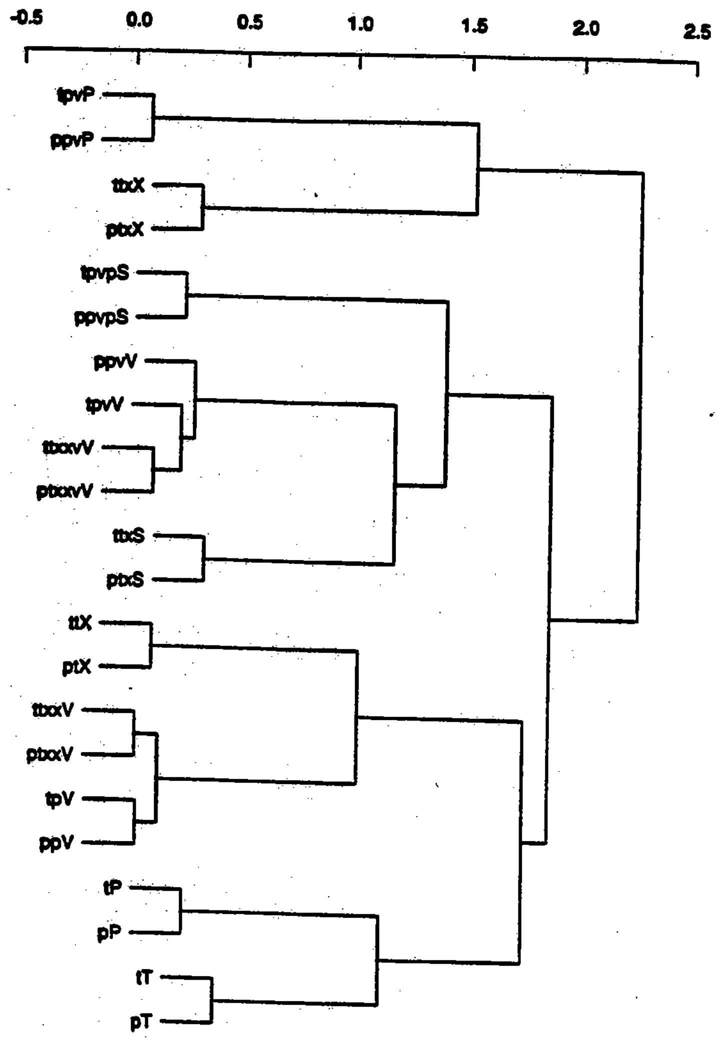

Figure 20. Cluster analysis of hidden unit activation patterns following the presentation of identical sequences in each of the two sub-grammars. Labels starting with the letter 't' come from the top sub-grammars, labels starting with the letter 'p' come from the bottom sub-grammar.

and to indicate the history of processing prior to the embedding. The identity of the initial letter simply *shades* the representation of states inside the embedding, so that corresponding nodes have similar representations, and are processed using overlapping portions of the knowledge encoded in the connection weights. Yet the shading that the initial letter provides allows the network to carry information about the early part of the string through the embedding, thereby allowing the network to exploit long-distance dependencies. This property of the internal representations used by the SRN is illustrated in Figure 20. We recorded some patterns of activation over the hidden units following the presentation of each letter inside the embeddings. The first letter of the string label in the figure (t or p) indicates whether the string corresponds to the upper or lower sub-grammar. The figure shows that the patterns of activation generated by identical embeddings in the two different sub-grammars are more similar to each other (e.g., 'tpvP' and 'ppvP') than to patterns of activation generated by different embeddings in the same sub-grammar (e.g., 'tpvP' and 'tpV'). This indicates that the network is sensitive to the similarlity of the corresponding nodes in each sub-grammar, while retaining information about what preceded entry into the sub-grammar.

6. Discussion

In this study, we attempted to understand better how the simple recurrent network could learn to represent and use contextual information when presented with structured sequences of inputs. Following the first experiment, we concluded that copying the state of activation on the hidden layer at the previous time step provided the network with the basic equipment of a finite state machine. When the set of exemplars that the network is trained on comes from a finite state grammar, the network can be used as a recognizer with respect to that grammar. When the representational resources are severely constrained, internal representations actually converge on the nodes of the grammar. Interestingly, though, this representational convergence is not a necessary condition for functional convergence: networks with more than enough structure to handle the prediction task sometimes represent the same node of the grammar using two quite different patterns, corresponding to different paths into the same node. This divergence of representations does not upset the network's ability to serve as a recognizer for well-formed sequences derived from the grammar.

We also showed that the mere presence of recurrent connections pushed the network to develop hidden layer patterns that capture information about sequences of inputs, even in the absence of training. The second experiment showed that back-propagation can be used to take advantage of this natural tendency when information about the path traversed is relevant to the task at hand. This was illustrated with predictions that were specific to particular subsequences in the training set or that took into account constraints on the length of sequences. Encoding of sequential structure depends on the fact that back-propagation causes hidden layers to encode task-relevant information. In the simple recurrent network, internal representations encode not only the prior event but also relevant aspects of the representation that was constructed in predicting the prior event from its predecessor. When fed back as input, these representations provide information that allows the network to maintain prediction-relevant features of an entire sequence. We illustrated this with cluster analyses of the hidden layer patterns.

Our description of the stages of learning suggested that the network initially learns to distinguish between events independently of the temporal context (e.g., simply distinguish between different letters). The information contained in the context layer is ignored at this point. At the end of this stage, each event is associated to a specific pattern on the hidden layer that identifies it for the following event. In the next phase, thanks to this new information, different occurrences of the same event (e.g., two occurrences of the same letter) are distinguished on the basis of immediately preceding events—the simplest form of a time tag. This stage corresponds to the recognition of the different 'arcs' in the particular finite state grammar used in the experiments. Finally, as the representation of each event progressively acquires a time tag, sub-sequences of events come to yield characteristic hidden layer patterns that can form the basis of further discriminations (e.g., between an 'early' and 'late' 'T_2' in the Reber grammar). In this manner, and under appropriate conditions, the hidden unit patterns achieve an encoding of the entire sequence of events presented.

We do not mean to suggest that simple recurrent networks can learn to recognize *any* finite state language. Indeed, we were able to predict two conditions under which performance of the simple recurrent network will deteriorate: (1) when different sequences may contain identical embedded sequences involving *exactly* the same predictions; and (2) when the number of hidden units is restricted and cannot support redundant representations of similar predictions, so that identical predictions following different events tend to be associated with very similar hidden unit patterns, thereby erasing information about the initial path. We also noted that when recurrent connections are added to a three-layer feed-forward network, back-propagation is no longer guaranteed to perform gradient descent in the error space. Additional training, by improving performance on shared components of otherwise differing sequences, can eliminate information necessary to 'span' an embedded sequence and result in a sudden rise in the total error. It follows from these limitations that the simple recurrent network could not be expected to learn sequences with a moderately complex recursive structure—such as context free grammars—if contingencies inside the embedded structures do not depend on relevant information preceding the embeddings.

What is the relevance of this work with regard to language processing? The ability to exploit long-distance dependencies is an inherent aspect of human language processing capabilities, and it lies at the heart of the general belief that a recursive computational machine is necessary for processing natural language. The experiments we have done with SRNs suggest another possibility; it may be that long-distance dependencies can be processed by machines that are simpler than fully recursive machines, as long as they make use of *graded* state information. This is particularly true if the probability structure of the grammar defining the material to be learned reflects—even very slightly—the information that needs to be maintained. As we noted previously, natural linguistic stimuli may show this property. Of course, true natural language is far more complex than the simple strings that can be generated by the machine shown in Figure 24, so we cannot claim to have shown that graded state machines will be able to process all aspects of natural language. However, our experiments indicate already that they are more powerful in interesting ways than traditional finite state automata (see also the work of Allen and Riecksen, 1989; Elman, 1990; and Pollack, in press). Certainly, the SRN should be seen as a new entry into the taxonomy of computational machines. Whether the SRN—or rather some other instance of the broader class of graded state machines of which the SRN is one of the simplest—will ultimately turn out to prove sufficient for natural language processing remains to be explored by further research.

Acknowledgments

We gratefully acknowledge the constructive comments of Jordan Pollack and an anonymous reviewer on an earlier draft of this paper. David Servan-Schreiber was supported by an

NIMH Individual Fellow Award MH-09696-01. Axel Cleeremans was supported by a grant from the National Fund for Scientific Research (Belgium). James L. McClelland was supported by an NIMH Research Scientist Career Development Award MH-00385. Support for computational resources was provided by NSF(BNS-86-09729) and ONR(N00014-86-G-0146). Portions of this paper have previously appeared in Servan-Schreiber, Cleeremans & McClelland (1989), and in Cleeremans, Servan-Schreiber & McClelland (1989).

Notes

1. Slightly modified versions of the BP program from McClelland and Rumelhart (1988) were used for this and all subsequent simulations reported in this paper. The weights in the network were initially set to random values between −0.5 and +0.5. Values of learning rate and momentum (*eta* and *alpha* in Rumelhart el al. (1986)) were chosen sufficiently small to avoid large oscillations and were generally in the range of 0.01 to 0.02 for learning rate and 0.5 to 0.9 for momentum.
2. For any single output unit, given that targets are binary, and assuming a fixed input pattern for all training exemplars, the error can be expressed as:

$$p(1 - a)^2 + (1 - p)a^2$$

where p is the probability that the unit should be on, and a is the activation of the unit. The first term applies when the target is 1, the second when the target is 0. Back-propagation tends to minimize the derivative of this expression, which is simply 2a − 2p. The minimum is attained when a = p, i.e., when the activation of the unit is equal to its probability of being on in the training set (Rumelhart, personal communication to McClelland, Spring 1989).
3. Cluster analysis is a method that finds the optimal partition of a set of vectors according to some measure of similarity (here, the euclidean distance). On the graphical representation of the obtained clusters, the contrast between two groups is indicated by the length of the horizontal links. The length of vertical links is not meaningful.
4. This fact may seem surprising at first, since the learning algorithm does not apply pressure on the weights to generate different representations for different paths to the same node. Preserving that kind of information about the path does not contribute in itself to reducing error in the prediction task. We must therefore conclude that this differentiation is a direct consequence of the recurrent nature of the architecture rather than a consequence of back-propagation. Indeed, in Servan-Schreiber, Cleeremans and McClelland (1988), we showed that some amount of information about the path is encoded in the hidden layer patterns when a succession of letters is presented, even in the absence of any training.
5. In fact, length constraints are treated exactly as atypical cases since there is no representation of the length of the string as such.
6. *For each statistical order*, the analysis consisted of three steps: First, we estimated the conditional probabilities of observing each letter after each possible path through the grammar (e.g., the probabilities of observing each of the seven letters given the sequence 'TSS'). Second, we computed the probabilities that each of the above paths leads to each node of the grammar (e.g., the probabilities that the path "TSS" finishes at node #1, node #2, etc.). Third, we obtained the average conditional probabilities (ACP) of observing each letter at each node of the grammar by summing the products of the terms obtained in steps #1 and #2 over the set of possible paths. Finally, all the ACPs that corresponded to letters that could *not* appear at a particular node (e.g., a 'V' at node #0, etc.) were eliminated from the analysis. Thus, for each statistical order, we obtained a set of 11 ACPs (one for each occurrence of the five letters, and one for 'E,' which can only appear at node #6. 'B' is never predicted).
7. For example, with three hidden units, the network converges to a stable state after an average of three iterations when presented with identical inputs (with a precision of two decimal points for each unit). A network with 15 hidden units converges after an average of 8 iterations. These results were obtained with random weights in the range [−0.5, +0.5].

8. Generally, small values for the learning rate and momentum, as well as many hidden units, help to minimize this problem.
9. The Luce ratio is the ratio of the highest activation on the output layer to the sum of all activations on that layer. This measure is commonly applied in psychology to model the strength of a response tendency among a finite set of alternatives (Luce, 1963). In this simulation, a Luce ratio of 0.5 often corresponded to a situation where 'T' and 'P' were equally activated and all other alternatives were set to zero.

References

Allen, R.B. (1988). Sequential connectionist networks for answering simple questions about a microworld. *Proceedings of the Tenth Annual Conference of the Cognitive Science Society.*

Allen, R.B., & Riecksen, M.E. (1989). Reference in connectionist language users. In R. Pfeifer, Z. Schreter, F. Fogelman-Soulié, & L. Steels (Eds.), *Connectionism in perspective.* North Holland: Amsterdam.

Allen, R.B. (1990). *Connectionist language users* (TR-AR-90-402). Morristown, NJ: Bell Communications Research.

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

Cottrell, G.W. (1985). Connectionist parsing. *Proceedings of the Seventh Annual Conference of the Cognitive Science Society.* Hillsdale, NJ: Erlbaum.

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

Elman, J.L. (1990). Representation and structure in connectionist models. In Gerry T.M. Altmann (Ed.), *Cognitive models of speech processing: Psycholinguistic and computational perspectives.* Cambridge, MA: MIT Press.

Fanty, M. (1985). *Context-free parsing in connectionist networks* (TR174). Rochester, NY: University of Rochester, Computer Science Department.

Hanson, S. & Kegl, J. (1987). PARSNIP: A connectionist network that learns natural language from exposure to natural language sentences. *Proceedings of the Ninth Annual Conference of the Cognitive Science Society.* Hillsdale, NJ: Erlbaum.

Hinton, G., McClelland, J.L., & Rumelhart, D.E. (1986). Distributed representations. In D.E. Rumelhart and J.L. McClelland (Eds.), *Parallel distributed processing, 1: Foundations.* Cambridge, MA: MIT Press.

Jordan, M.I. (1986). Attractor dynamics and parallelism in a connectionist sequential machine. *Proceedings of the Eighth Annual Conference of the Cognitive Science Society.* Hillsdale, NJ: Erlbaum.

Luce, R.D. (1963). Detection and recognition. In R.D. Luce, R.R. Bush and E. Galanter (Eds.), *Handbook of mathematical psychology (Vol. 1).* New York: Wiley.

McClelland, J.L., & Rumelhart, D.E. (1988). *Explorations in parallel distributed processing: A handbook of models, programs and exercises.* Cambridge, MA: MIT Press.

Pollack, J. (in press). Recursive distributed representations. *Artificial Intelligence.*

Reber, A.S. (1976). Implicit learning of synthetic languages: The role of the instructional set. *Journal of Experimental Psychology: Human Learning and Memory, 2,* 88–94.

Rumelhart, D.E., & McClelland, J.L. (1986). *Parallel distributed processing, 1: Foundations.* Cambridge, MA: MIT Press.

Rumelhart, D.E., Hinton, G., & Williams, R.J. (1986). Learning internal representations by error propagation. In D.E. Rumelhart and J.L. McClelland (Eds.), *Parallel distributed processing, 1: Foundations.* Cambridge, MA: MIT Press.

Sejnowski, T.J., & Rosenberg, C. (1987). Parallel networks that learn to pronounce english text. *Complex Systems, 1,* 145–168.

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

Servan-Schreiber D., Cleeremans A., & McClelland J.L. (1989). Learning sequential structure in simple recurrent networks. In D.S. Touretzky (Ed.), *Advances in neural information processing systems 1.* San Mateo, CA: Morgan Kaufmann. [Collected papers of the IEEE Conference on Neural Information Processing Systems—Natural and Synthetic, Denver, Nov. 28–Dec. 1, 1988].

St. John, M., & McClelland, J.L. (in press). Learning and applying contextual constraints in sentence comprehension. *Artificial Intelligence.*