Integrating Learning into Models of Human Memory: The Hebbian Recurrent Network[1]

Simon Dennis† and Janet Wiles‡
Department of Computer Science†
The University of Queensland, 4072, Australia
mav@cs.uq.oz.au

Departments of Psychology and Computer Science‡
The University of Queensland, 4072, Australia
janetw@cs.uq.oz.au

Abstract

We develop an *interactive* model of human memory called the Hebbian Recurrent Network (HRN) which integrates work in the mathematical modeling of memory with that in error correcting connectionist networks. It incorporates the Matrix Model (Pike, 1984) into the Simple Recurrent Network (SRN, Elman, 1989). The result is an architecture which has the desirable memory characteristics of the matrix model such as low interference and massive generalization, but which is able to *learn* appropriate encodings for items, decision criteria and the control functions of memory which have traditionally been chosen a priori in the mathematical memory literature. Simulations demonstrate that the HRN is well suited to a recognition task inspired by typical memory paradigms. In comparison to the SRN, the HRN is able to learn longer lists, and is not degraded significantly by increasing the vocabulary size.

Introduction

Within the recent literature there has been considerable debate about the ability of connectionist models to capture human memory phenomena. While there is a feeling that connectionist models have much to offer, the results thus far have been mixed (Ratcliff, 1990; Lewandowsky, 1991). Connectionist models provide mechanisms by which encodings, decision criteria and control functions can be learned, yet, on some very basic variables like the degree of interference, they have not performed as well as conventional memory models.

We start by discussing the contribution of connectionist models to the modeling of memory. Next we examine some of the major aspects of memory phenomena which remain obstacles for connectionist models of memory. Finally we outline the Hebbian Recurrent Network (HRN) which integrates learning and memory models to address these issues.

Learning Issues

In an interactive model, it is the interplay of the environment and the architecture which leads to performance (Dennis, Wiles, & Humphreys, 1992). For example in optimizing connectionist models, performance is determined both by the architecture (i.e., structure of the interconnections, the transfer function, and the values of the parameters) and the statistical contingencies embodied by the training set. There are in essence three major advantages of such a system for the modeling of human memory (Dennis et al., 1992).

Firstly, while the performance of current memory models depends solely on architectural assumptions, the performance of an interactive model depends both on architectural and environmental assumptions. Since the environment is observable, environmental assumptions tend to be more amenable to empirical verification (Anderson & Schooler, 1991; Dennis et al., 1992) and hence interactive models can provide more parsimonious accounts of the phenomena of memory.

Secondly, interactive architectures can address the developmental course of memory phenomena (Bates & Elman, 1992). The developmental literature demonstrates that memory strategies such as imag-

[1] This paper has been supported by an Australian Postgraduate Research Award to the first author, and an Australian Research Council grant to Humphreys, Burt, Wiles and Tehan.

ing, elaboration, naming and rehearsal are learned (Hasher & Zacks, 1979), yet there has been little progress in modeling such processes. Connectionist models may be able to redress the situation.

Thirdly, there are a number of well documented learning to learn phenomena which are not addressed by current architectures (Postman, 1969). While there has been some work done on specific transfer effects (i.e., those which depend on the manipulation of relations of successive tasks) there has been no attempt to address nonspecific transfer effects (i.e., higher order transfer of skills, e.g., response learning) despite a well developed empirical database. Connectionist models offer an opportunity to develop such a theory.

So, which aspects of the memory system are learned? Current models can be summarized in terms of their representations, their decision criteria and their control paradigms. The following sections address each of these in turn.

Learning Representation. The formation of appropriate encodings for items which will be entered into memory has been a difficult and largely unexplored area in the memory literature. While it is known that items with similar meanings and perceptual forms interact with each other to a greater extent than unrelated items there has been little progress in determining how the formation of such an encoding landscape may come about.

Connectionist models however, construct their own internal representations, and hence offer a way of avoiding many representational assumptions. Furthermore, connectionist networks learn these representations as a consequence of the environment in which they are situated. They introduce a principled way in which "abstract features" might be formed. Hence the first of our design criteria for a model of memory is that internal representation be learned.

Learning Decision Criteria. Another aspect of the memory task which is usually held constant is the decision mechanism. In most current models a signal detection framework is applied and there is no sense in which the matching function could be said to have been acquired.

In the majority of memory paradigms, however, subjects become more accurate as they gain experience. While some of the improvement may be due to the refinement of the representation, a component of this improvement is attributable to an improvement in the ability to decide upon a response (Postman, 1969). Hence the second design criterion of the HRN is that it be able to acquire decision criteria.

Learning Control. Within the memory literature the nature of the control processes is often taken for granted. For instance, how is it that the subject decides when to give a response? Traditional mathematical models of memory assume that the answer to this question is embedded in the program.

To learn control regimes within a connectionist architecture one must turn to recurrent networks. Feedforward models learn representations and decisions but cannot embody the temporal relations which characterize control problems. There have been attempts to apply recurrent networks in this area (Nolfi, Parisi, Vallar, & Burani, 1990; Wiles & Phillips, 1991), and we will review these in more detail when we attempt to fulfill the third design criterion, that is, the acquisition of control processes.

Memory Issues

In the previous section we outlined the aspects of the memory system which might be acquired by a connectionist system. To be serious contenders as memory models, however, there are a number of obstacles which must be overcome. It is to these that we now turn.

Capacity and Interference. The problem of catastrophic interference has received a great deal of attention in the recent literature (Ratcliff, 1990; McCloskey & Cohen, 1989; Lewandowsky, 1991; French, 1991; Brousse & Smolensky, 1989; Hetherington & Seidenberg, 1989; Wiles & Phillips, 1991). The difficulty arises when what has been learned is disrupted dramatically by subsequent learning, i.e. there is significant retroactive interference. The problem is of particular importance in the modeling of recognition memory where the capacity is very large and the degree of interference is small.

Within the literature two major strategies have emerged in order to deal with the problem of catastrophic interference. The first involves increasing the orthogonality of items and hence decreasing the extent to which they interact (Lewandowsky, 1991; French, 1991).

The alternative approach has been to use recurrent architectures to encode lists of items rather than single items on their hidden units (Nolfi et al., 1990; Wiles & Phillips, 1991). The network then has the task of learning a single higher order encoding function rather than a sequence of items. Unfortunately, this encoding function becomes much more difficult to acquire as the number of items to be encoded increases, and hence current recurrent architectures have strict capacity restrictions. The first of the memory criteria, then, is that the model avoid catastrophic interference while maintaining capacity.

Generalization. Another issue closely related to interference is the degree of generalization. What proportion of the entire space of possible inputs must the network be presented with during its training phase in order to perform well on unseen cases? In the context of memory, the important variable is the size of the vocabulary. Subjects have extensive vocabularies, yet are able to perform memory tasks involving any of the items within that vocabulary. The second memory criterion is to be able to generalize well as the size of the vocabulary increases.

Rapid Binding. The last of the memory criteria revolves around an important issue raised by Wiles and Phillips (1991) concerning the distinction between memorization and learning. Memory involves the rapid binding of already established representations rather than the acquisition of new representations. Enduring memories can be laid with presentation durations of just a few hundred milliseconds. The time course of learning to learn effects, in contrast, tends to be in the order of hours or days. A model of memory should be capable of explaining the difference between memorization and learning, and be able to account for the difference in the time scales.

The Hebbian Recurrent Network

The aim is to address the above criteria by incorporating the Matrix Model into the SRN in a fashion which takes advantage of the strengths of each. There are two important points about the structure of the HRN (see figure 1). Firstly, the representations should be determined by the dynamics of the network as a consequence of learning and not chosen a priori by the experimenter. Hence the inputs to the memory system should come from the hidden activations of the backpropagation network. Secondly, two layers of backpropagation weights are available to map the outputs of memory onto an appropriate response and result of probing memory can be used to construct the next cue to memory, allowing the possibility of chains of recollection.

There are two ways in which one may view the HRN. The first is to consider it a matrix memory which has some optimized weights around it to handle the control, decision and representation aspects. The other way of thinking about it is as an SRN with long term memory.

Simulations and Results

The SRN and HRN were applied to an episodic recognition task. The network was presented with a sequence of study items during which it was to respond with the "Blank" symbol. These were followed by a test item and on the next time step the network had to respond "Yes" if the test item was in the study list and "No" otherwise.

List Length. Figure 2(A) shows the effect of increasing the list length on the performance of the SRN, and HRN respectively. While the performance of the SRN has decreased to chance after only 5 items, the HRN sustained its performance until it reached 10 item lists. In these simulations the vocabulary size was set to be twice the list length so as to maintain a 0.5 probability of a positive test item. Hence when list length reaches 10 the vocabulary has reached 20 items. The inability of the network to memorize larger lists may be a consequence of the fact that the vocabulary size has reached the number of hidden units rather than an indication of the memory capacity.

Vocabulary Size. Figure 2(B) shows performance as the vocabulary size is manipulated when the HRN is applied to the 4-item recognition task. As was noted earlier, performance is virtually unaffected by vocabulary size until it reaches the number of hidden units. At this point there is a sharp drop.

Hidden Unit Analysis. While the SRN must form a representation of the entire list in its hidden unit activation patterns, the HRN can rely on the hebbian memory to store items. Figure 3 shows Hierarchical Cluster Diagrams (HCA) of the hidden unit patterns after the final study item has been input which demonstrate this point.

Discussion and Conclusions

The introduction outlines six criteria for a model of human memory and at this stage we evaluate the HRN's performance on these. The architecture of the HRN is designed so that it will learn representation, decision criteria and control and hence the first three criteria are met. In such a simple task the degree of control required was limited. The only major distinction to be made was between the study phase, and the decision phase. This distinction though was typically learned very quickly, usually well within one hundred epochs.

The HRN avoids catastrophic interference, inheriting the performance characteristics from the matrix memory. Its ability to generalize well even as the number of vocabulary items increased is of particular importance, and is one of the major distinguishing factors between it and the SRN. The last of the criteria was the rapid binding in memory of already estab-

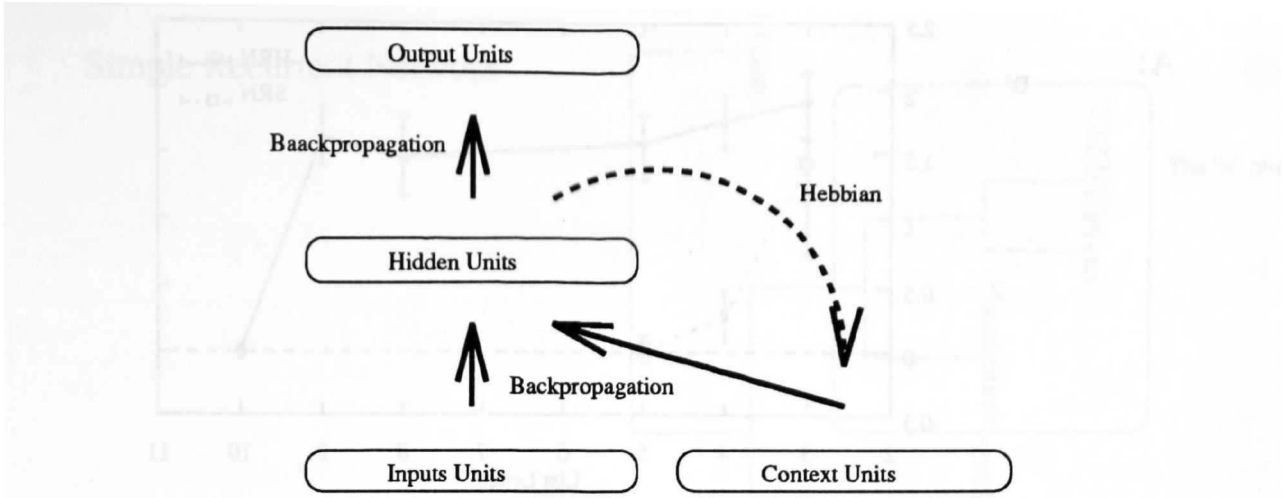

Figure 1: The Hebbian Recurrent Network Architecture. The solid arrows are sets of weights which are modified using the backpropagation algorithm. The dashed line represents the feeding of hidden unit activations through a set of Hebbian weights to the context units in preparation for the next timestep. The Hebbian weights are updated after the activations are fed through. In addition the input pattern was required on the output to ensure that the states corresponding to different inputs were separated.

lished representations. While the possible representations are developed by the backpropagation mechanism, specific memories are stored in the hebbian weights. It is the dual memory architecture which avoids catastrophic interference, allows for the significant improvement in generalization, and accounts for the dramatically different timespans of memory and learning.

References

Anderson, J. R., & Schooler, L. J. 1991. Reflections of the environment in memory. *Psychological Science*, *2*(6):396–408.

Bates, E. A., & Elman, J. L. 1992. Connectionism and the study of change. Technical Report, 9202, Center for Research in Language. University of California, San Diego, CA.

Brousse, O., & Smolensky, P. 1989. Virtual memories and massive generalization in connectionist combinatorial learning. In *Program of the Eleventh Annual Conference of the Cognitive Science Society*, 380–387 Hillsdale, NJ. Lawrence Erlbaum Associates.

Dennis, S., Wiles, J., & Humphreys, M. S. 1992. What does the environment look like? Setting the scene for interactive models of human memory. Technical Report, 249, Department of Computer Science. University of Queensland, Australia.

Elman, J. L. 1989. Representation and structure in connectionist models. Technical Report, 8903, Center for Research in Language. University of California, San Diego, CA.

French, R. 1991. Using semi-distributed representations to overcome catastrophic forgetting in connectionist networks. Technical Report, 51, Center for Research on Concepts and Cognition. Indiana University, IN.

Hasher, L., & Zacks, R. T. 1979. Automatic and effortful processes in memory. *Journal of Experimental Psychology: General*, *108*:356–388.

Hetherington, P., & Seidenberg, M. S. 1989. Is there catastrophic interference in connectionist networks?. In *Program of the Eleventh Annual Conference of the Cognitive Science Society*, 26–33 Hillsdale, NJ. Lawrence Erlbaum Associates.

Lewandowsky, S. 1991. Gradual unlearning and catastrophic interference: A comparison of distributed architectures. In Hockley, W. E., & Lewandowsky, S. (Eds.), *Relating Theory and Data: Essays on Human Memory in Honor of Bennet B. Murdock*, 445–476. Lawrence Erlbaum Associates, Hillsdale, NJ.

McCloskey, M., & Cohen, N. J. 1989. Catastrophic interference in connectionist networks: The se-

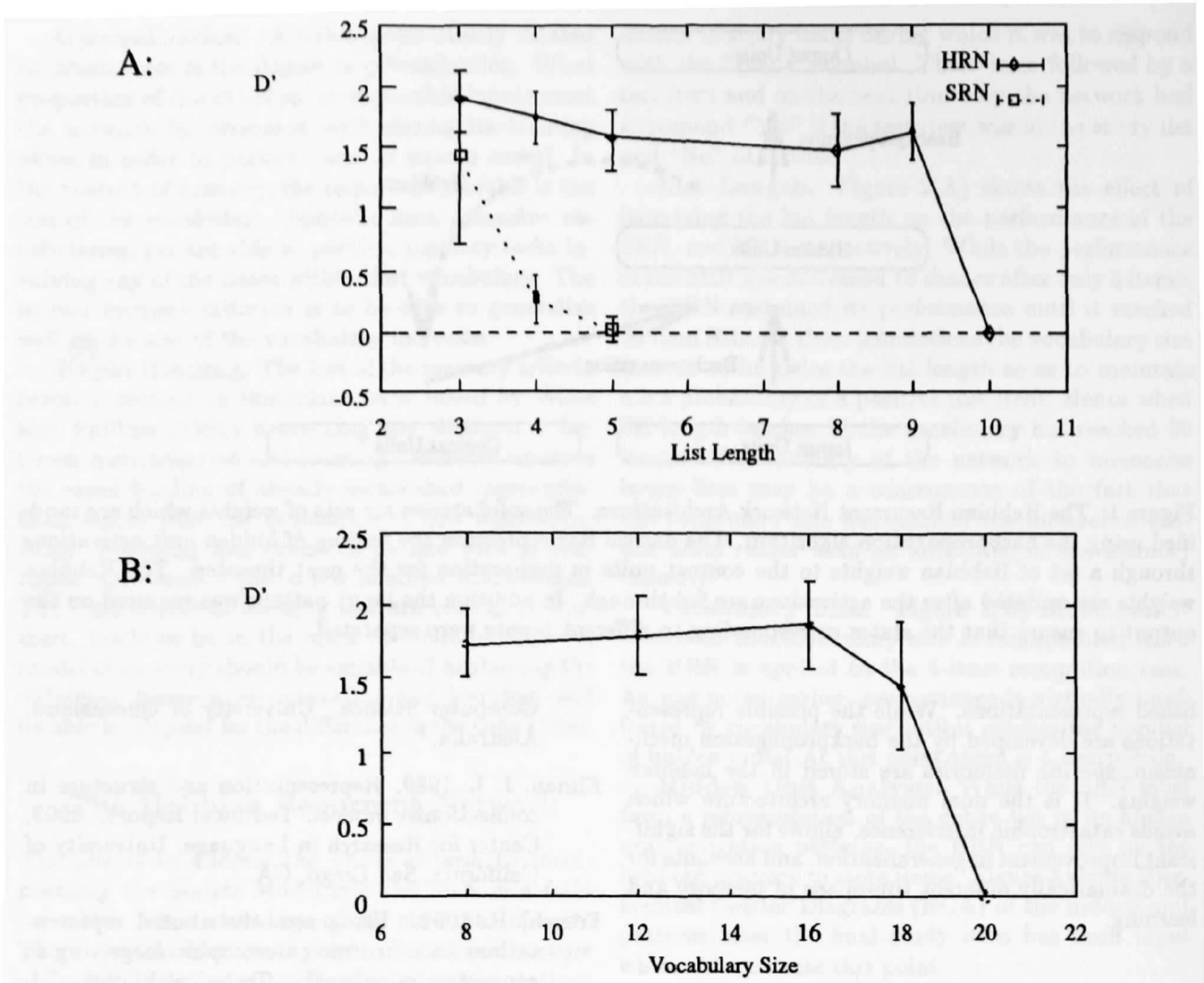

Figure 2: (A) Performance (d') as a function of List Length on the HRN and SRN. The HRN performs much better than the SRN on lists of length 9 or less. On lists of length 10 the HRN's performance falls to chance. (B) Performance (d') as a function of the Number of Vocabulary Items using the HRN on the 4 item recognition task. The size of the vocabulary seems to make little difference until it reaches 20 when it drops to below chance. The bars indicate the 95% confidence intervals.

quential learning problem.. In Bower, G. H. (Ed.), *The psychology of learning and motivation*, 109–165. Academic Press, NY.

Nolfi, S., Parisi, D., Vallar, G., & Burani, C. 1990. Recall of sequences of items by a neural network. In Touretsky, D. S., Elman, J. L., Sejnowski, T. J., & Hinton, G. E. (Eds.), *Proceedings of the 1990 Connectionist Models Summer School*. Morgan Kaufmann, San Mateo, CA.

Pike, R. 1984. Comparison of convolution and matrix distributed memory systems for associative recall and recognition. *Psychological Review*, *91*(3):281–293.

Postman, L. 1969. Experimental analysis of learning to learn. In Bower, G. H., & Spence, J. T. (Eds.), *Psychology of Learning and Motivation*, Vol. 3, 241–297. Academic Press.

Ratcliff, R. 1990. Connectionist models of recognition memory: Constraints imposed by learning and forgetting functions. *Psychological Review*, *97*(2):285–308.

Wiles, J., & Phillips, S. 1991. Serial recall of binary sequences. Unpublished manuscript.

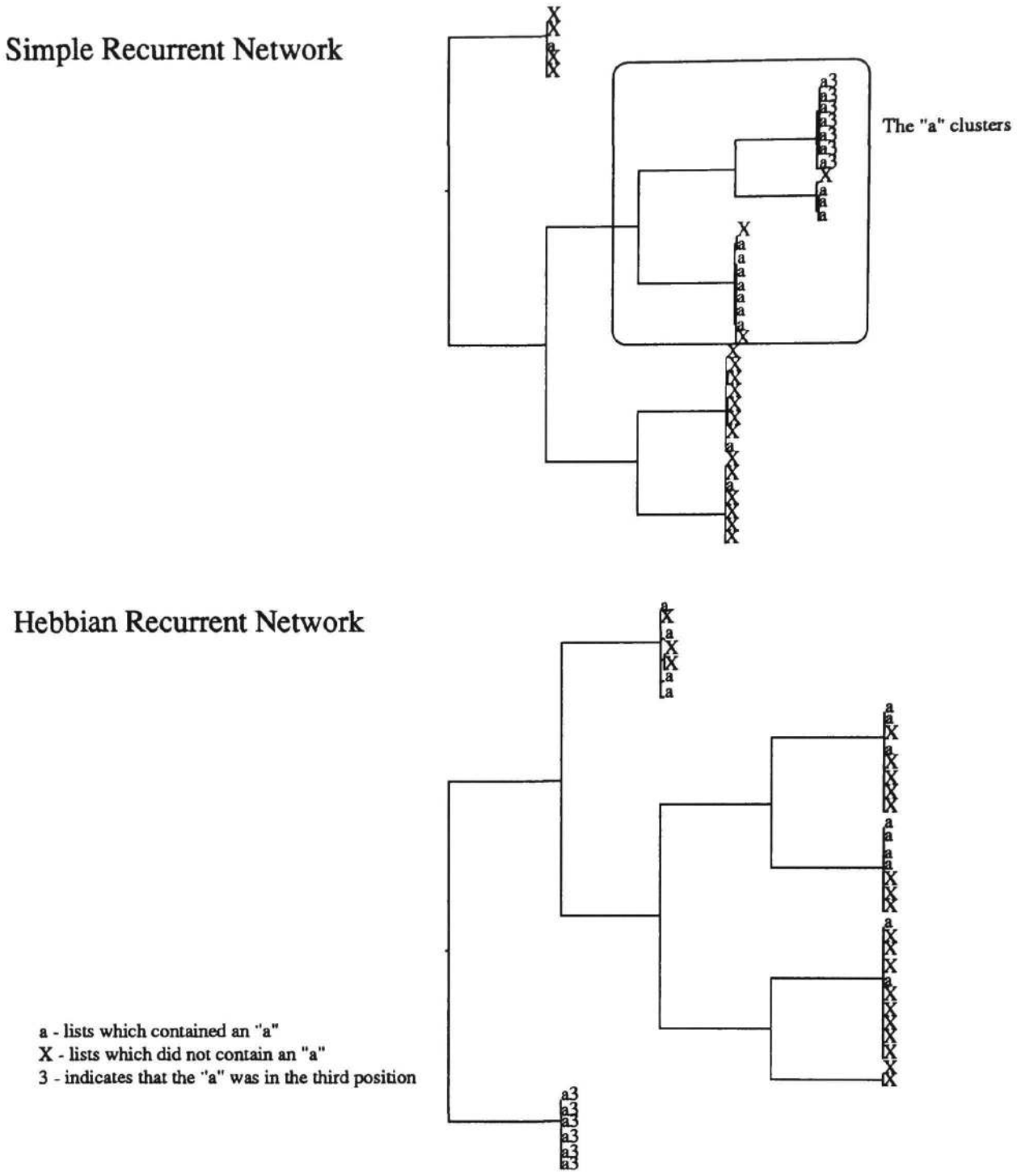

Figure 3: Hierarchical Cluster Analysis (HCA) of the hidden unit patterns after the study list has been input. For the SRN the sequences which contain an "a" regardless of position are clustered together. In contrast the HRN clusters only the third position "a"s well. The first and second position "a"s are mixed in with the non "a" patterns (i.e. Xs). The HRN does not need to separate these items in the hidden unit vectors since they are retained in the hebbian memory.