# Incremental Sequence Learning

**Axel Cleeremans** (AXCLEER@ULB.AC.BE)
**Arnaud Destrebecqz** (ADESTRE@ULB.AC.BE)
Séminaire de Recherche en Sciences Cognitives
Université Libre de Bruxelles
Avenue F.-D. Roosevelt, 50 — CP 122
1050 Bruxelles — Belgium

## Abstract

As linguistic competence so clearly illustrates, processing sequences of events is a fundamental aspect of human cognition. For this reason perhaps, sequence learning behavior currently attracts considerable attention in both cognitive psychology and computational theory. In typical sequence learning situations, participants are asked to react to each element of sequentially structured visual sequences of events. An important issue in this context is to determine whether essentially associative processes are sufficient to understand human performance, or whether more powerful learning mechanisms are necessary. To address this issue, we explore how well human participants and connectionist models are capable of learning sequential material that involves complex, disjoint, long-distance contingencies. We show that the popular Simple Recurrent Network model (Elman, 1990), which has otherwise been shown to account for a variety of empirical findings (Cleeremans, 1993), fails to account for human performance in several experimental situations meant to test the model's specific predictions. In previous research (Cleeremans, 1993) briefly described in this paper, the structure of center-embedded sequential structures was manipulated to be strictly identical or probabilistically different as a function of the elements surrounding the embedding. While the SRN could only learn in the second case, human subjects were found to be insensitive to the manipulation. In the new experiment described in this paper, we tested the idea that performance benefits from "starting small effects" (Elman, 1993) by contrasting two conditions in which the training regimen was either incremental or not. Again, while the SRN is only capable of learning in the first case, human subjects were able to learn in both. We suggest an alternative model based on Maskara & Noetzel's (1991) Auto-Associative Recurrent Network as a way to overcome the SRN model's failure to account for the empirical findings.

## Introduction

Over the past few years, sequence learning has become one of the major paradigms through which to study elementary learning processes, particularly in the context of implicit learning research (see Berry & Dienes, 1993; Cleeremans, 1993 for reviews). In typical sequence learning situations, participants are asked to react to each element of sequentially structured visual sequences of events (e.g., Nissen & Bullemer, 1987). There is now a large literature showing that human subjects can exhibit very detailed sensitivity to the sequential constraints present in the material through the differences in their reaction time to stimuli that are or are not predictable based on the temporal context set by previous elements of the sequence.

An important issue in this context is to determine whether essentially associative processes are sufficient to understand human performance, or whether more powerful learning mechanisms are necessary. This issue has been typically approached by exposing participants to complex material that involves disjoint temporal contingencies, that is, contingencies between elements of a sequence that are separated by a number of other irrelevant elements. For instance, Reed and Johnson (1994) trained their participants on so-called second-order conditional sequences in which each element *t* of the sequence can only be predicted based on the identity of both elements *t - 2* and *t 1*. Other research has focused specifically on the question of determining whether human participants can maintain information about long-distance contingencies over embedded material, as illustrated by the following two natural language expressions:

The **dog** - *that chased the cat* - **is** playful
The **dogs** - *that chased the cat* - **are** playful

Both expressions share an embedding that is completely irrelevant in determining the number of the verb. When processing such expressions, information about the number of the head (dog vs. dogs) therefore has to be maintained in memory until processing of the embedding information has been completed.

Such expressions present interesting challenges for popular sequential connectionist architectures such as the Simple Recurrent Network (henceforth, SRN). The SRN, first proposed by Elman (1990), and subsequently adapted by Cleeremans & McClelland (1991) to simulate sequential effects in reaction time tasks, is shown in Figure 1. The network uses back-propagation to learn to predict the next element of a sequence based only on the current element and on a representation of the temporal context that the network has elaborated itself. To do so, it uses information provided by so-called context units which, on every step, contain a copy of the network's hidden unit activation vector at the previous time step. Over training, the relative activation of

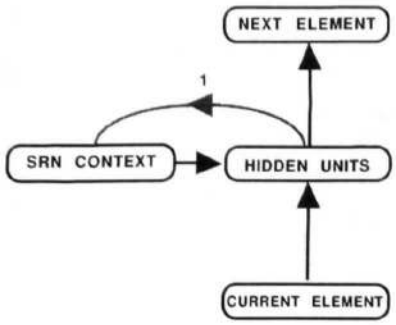


**Figure 4:** The simple recurrent network (SRN). See text for details.

the output units representing each possible successor come to approximate the optimal conditional probabilities associated with their appearance in the current context, and can thus be interpreted as representing implicit preparation for the next event when the network is used as a model of human sequence learning performance. Previous work (see Cleeremans & McClelland, 1991; Cleeremans, 1993) has shown that the SRN is able to account for about 80% of the variance in sequential choice reaction time data.

The SRN, however, also suffers from an important limitation in its ability to learn sequential material. Indeed, one key aspect of learning in the SRN is that the material need to be "prediction-relevant" at every step for its representation to be maintained in the context layer (see Servan-Schreiber, Cleeremans & McClelland, 1991). In other words, each element of the sequence has to be useful in predicting the next one, even if only probabilistically so. This specific limitation is not shared by other architectures of similar complexity, such as the Jordan network (Jordan, 1986) or buffer networks (see Cleeremans, 1993). How well human participants perform with such material therefore has interesting diagnostic value in determining which model is best fit to account for human sequence learning performance.

This is the issue we focus on in the rest of this paper. We start by reviewing existing data in light of previous research on the SRN. Next, we report on an experiment meant to compare human and simulated performance in conditions where the SRN is known to fail.

## Overcoming SRN limitations

Previous research on this issue has revealed a number of important facts about the SRN's limitations. In particular, several authors have attempted to show that the SRN's sensitivity to prediction-relevance can be overcome by changing features of the stimulus environment to which the network is exposed during training. Two arguments have been developed. First, one may argue that natural language situations seldom correspond to the artificially hard situation described above. The following example illustrates this *naturalistic argument* by showing that embedded structures are in fact often dependent on the information conveyed by the head, if only in subtle ways:

The **dog** - *that chased its tail* - **is** playful
The **dogs** - *that chased each other* - **are** playful

In these sentences, the embeddings can no longer be switched between the two expressions because the number of the head constraints to some extent which embeddings can follow. Hence the gist of this argument is that completely independent embeddings are the exception rather than the rule: Typical embeddings do contain (syntaxic or semantic) information that is relevant for the processing of subsequent information.

Based on this idea, Servan-Schreiber, Cleeremans and McClelland (1991) trained an SRN on sequential material generated from a finite-state grammar producing center-embedded structures of the following form:

T E* - T
P - E* - P

In these expressions, the last element (T or P) is contingent on the first one (the head), from which it is separated by a number of embedded elements E. In the symmetrical condition of Servan-Schreiber et al.'s simulations, the embedding was always identical regardless of which head had occurred. In the other, asymmetrical condition, the probability of occurrence of some embedded elements varied as a function of which head had been presented. In both conditions, the SRN was assessed on how well it could predict the tail after neutral embeddings. Servan-Schreiber et al. found that the SRN could only master the material in the asymmetrical condition. The SRN can therefore maintain information across irrelevant embedded elements, but only when the embedding as a whole is probabilistically dependent on the head.

Cleeremans (1993) tested human subjects in a choice-reaction time task using the same design as described above but with the grammar illustrated in Figure 2, which can generate center-embedded structures similar to the ones described in the previous paragraphs. He found that subjects were able to successfully anticipate which tail was most likely to occur after a given head in both conditions. In contrast, the SRN was only able to encode the long-distance contingencies between the outer elements in the asymmetrical condition, that is, when the embeddings were probabilistically dependent on the head.

Overall, these results therefore suggest that while the SRN's limitations can be overcome by changing the probability structure of the stimulus environment it is exposed to, human participants do not appear to suffer from these limitations at all.

A second way to overcome the SRN's limitations was proposed by Elman (1993). Elman found that the SRN was able to learn the kind of complex and hierarchically organized information that typically occurs in natural language when training is incremental, that is, when the network is only progressively exposed to the more complex sequential contingencies contained in the stimulus material. To quote Elman (1993): "The network fails to learn the task when the entire data set is presented all at once [but] when the training data were selected such that simple sentences were presented first, the network succeeded not only in mastering these, but then going on to master the complex

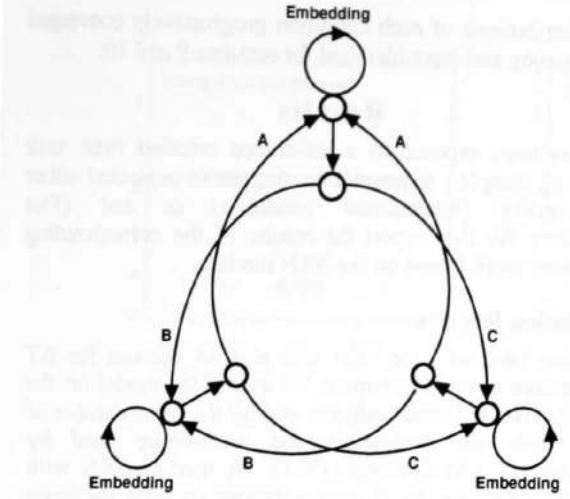


**Figure 2:** Finite-state grammar used to generate the stimulus material. See text for details.

sentences as well" (p. 74). Elman (1993) also found that directly manipulating the network's memory by forcing the activations of the context units to be reset at progressively larger intervals over training produced equivalent beneficial effects — a result that prompted Elman to frame his argument in terms of development, based on the observation that the capacity of children's short-term memory also increases with age.

In this paper, we propose to explore this idea in the context of sequence learning situations. We report on an experiment designed to test the effects of the training regimen in a six-choice sequential reaction time task. We contrasted two conditions that involved the same stimulus material but that used two different training regimens: an incremental one, and a non-incremental one. We detail this experiment in the next section.

## Experimental Design

### Method

The experiment consisted of 10 training sessions during which subjects were exposed to a serial six-choice RT task. Each session consisted of 20 blocks of 150 trials each, for a total of 30,000 trials. On each trial, a stimulus appeared at one of six positions arranged horizontally on a computer screen, and subjects were to press as fast and as accurately as possible on the corresponding key.

The sequential structure of the material was manipulated by generating the sequence based on the finite-state grammar illustrated in Figure 2, as described below. The sequences contained three different long-distance contingencies, the elements of which (i.e., the head and the tail) were separated by a varying number of *embedded* elements. To determine whether the training regimen has an impact on performance, we contrasted two conditions. In the *Flat* training condition, the distribution of embedding lengths was the same throughout training. Subjects were therefore exposed to the most complex material right from the start. By contrast, in the *Incremental* training condition, training started with mostly short embeddings, and the proportion of long embeddings was increased only progressively during training.

### Subjects

Twelve subjects were randomly assigned to either condition. Subjects were paid about $50 for their participation in the experiment and could earn an additional bonus of $10 to $20 based on performance.

### Apparatus and Display

The experiment was run on PowerPC Macintosh computers. The display consisted of six dots arranged in a horizontal line on the computer's screen and separated by intervals of 3 cm. Each screen position corresponded to a key on the computer's keyboard. The spatial configuration of the keys was fully compatible with the screen positions. The stimulus was a small black circle 0.35 cm high that appeared on a white screen background, centered 1 cm below one of the six dots. The RSI was 120 msec.

### Procedure

The procedure used in this experiment followed very closely the design described in Jiménez, Méndez and Cleeremans (1996). Subjects were exposed to two sessions each day for 5 consecutive days. All subjects were kept unaware of the fact that the material contained sequential contingencies, and were merely told that the experiment was about the effects of prolonged practice on motor performance. The instructions stressed both accuracy and speed. Short user-controlled rest breaks occurred between any two experimental blocks. During these breaks, subjects were given feedback about their performance and informed about how much bonus money they had earned so far. This amount was computed for each block based both on accuracy and speed.

### Stimulus material

Stimuli were generated based on the finite-state grammar illustrated in Figure 2. On each of the 30,000 trials, stimulus generation proceeded in two phases. First, an arc coming out of the current node was randomly selected, and its label recorded. The current node was initialized randomly at the start of each block, and was updated on each trial to be the node pointed to by the selected arc. Next, the recorded label was used to determine the screen position at which the stimulus would appear by following a 6 x 6 Latin square design, so that each label corresponded to each screen position for exactly one of the six subjects in each condition.

The grammar generates sequences that all share the following form:

$$H - E^* - T$$

where H designates a head element, E designates a embedded element, and T designates a tail element. The tail element of any such sequence also served as head for the next sequence. In our grammar, both heads and tails were instantiated by the

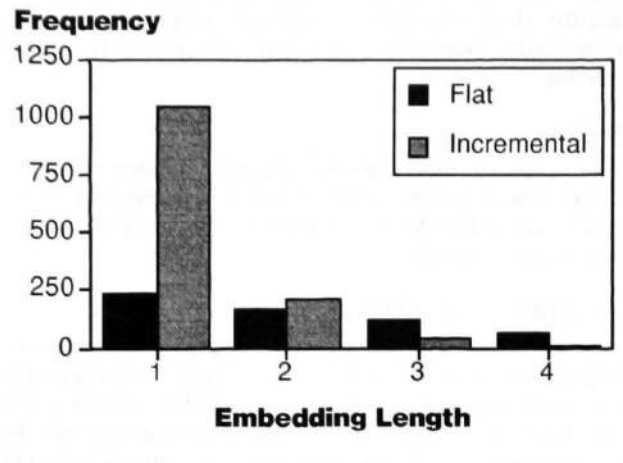


**Figure 3:** Initial frequency distribution of embeddings containing up to four elements in the Flat and Incremental conditions.

labels 'A', 'B', and 'C', which we subsequently refer to as the *outer elements*. The grammar was designed so that any particular head could be followed (after at least one element of the embedding) by the other two outer elements, and so that each head tended to be more strongly associated with one of the two legal tails than with the other. Thus for instance, if 'A' appears as the head of a given sequence, 'A' itself can not appear as the tail of this sequence, 'B' will tend to appear as the tail element in 80% of the cases, and 'C' in the remaining 20% of the cases. This difference in tail likelihood provides us with a simple way of assessing whether subjects are sensitive to the regularities contained in the material. Indeed, any difference in the reactions times elicited by likely vs. unlikely tails would clearly indicate that participants have encoded information about the head element, because only the head provides information about the distribution of tails.

As Figure 2 illustrates, heads and tails were always separated by an embedding. The embedding was instantiated by three different tokens (not representd in Figure 2): the labels 'X','Y', and 'Z'. The grammar was designed so that one element was mandatory. Subsequent elements of the embedding were chosen at random with the constraint that direct repetitions of any element were forbidden. Embedded elements are therefore totally irrelevant with respect to the task of predicting the tail. A random number of additional embedded elements could appear with probability $l$ on each of the three loops of the grammar. In the Flat condition, $l$ was set to 0.666 during the entire experiment. This means that each embedded element had a 0.666 chance of being followed by another embedded element. In the Incremental condition, by contrast, the probability to stay in the loop increased from 0.22 to 0.66 in steps of 0.11 every four sessions. Figure 3 shows the initial frequency distribution of embeddings up to length 4. One can see that short embeddings are much more frequent in the Incremental condition than in the Flat condition, and that this distribution reverses for embeddings of length 3 and higher. The distributions of each condition progressively converged over training and were identical for sessions 9 and 10.

## Results

Subjects were exposed to a six-choice reaction time task involving complex sequential contingencies presented either incrementally (Incremental condition) or not (Flat condition). We first report the results of the corresponding simulation work based on the SRN model.

### Simulation Results

To assess how well the SRN was able to account for RT performance in this experiment, we trained the model on the same material as human subjects and for the same number of trials, with the parameters and architecture used by Cleeremans and McClelland (1991). We used an SRN with 15 hidden units and local representations on both the input and output pools (i.e., each unit corresponded to one of the 6 stimuli). To account for short-term priming effects, the network used dual connection weights, as described in Cleeremans and McClelland (1991). The network was trained to predict each element of a continuous sequence of stimuli generated in exactly the same conditions as for human subjects. On each step, a label was generated from the grammar and presented to the network by setting the activation of the corresponding input unit to 1.0. Activation was then allowed to spread to the other units of the network, and the error between its response and the actual successor of the current stimulus was then used to modify the weights. During training, the running average activation of each output unit was recorded on every trial and transformed into Luce ratios to normalize the responses. For the purpose of comparing simulated and observed responses, we assumed (1) that the normalized activations of the output units represent response tendencies, and (2) that there is a linear reduction in RT proportional to the relative strength of the unit corresponding to the correct response. The network's responses were subtracted from 1.0 to make increases in response strength compatible with reduction in RT, and were finally transformed into zscores for easy comparison with human data.

The results are illustrated in the left panels of Figure 4. The data represent differences in the response strengths associated to either likely or unlikely tails of sequences containing up to three embedded elements, for the Flat condition (top panel) or the Incremental condition (bottom panel). The figure makes it clear that the SRN is incapable of learning even the shortest contingencies in the Flat condition: There are no differences between its responses to likely and unlikely tails regardless of the length of the embedding. The network simply fails to learn.

In contrast, the model appears capable of successfully predicting the tail element of sequences containing a single embedded element in the Incremental condition, producing responses that are about 5% stronger when the tail is likely to occur (given the head) vs. when it is not, at the end of training.

As predicted, the model is therefore quite sensitive to the difference between the training regimens used to present the

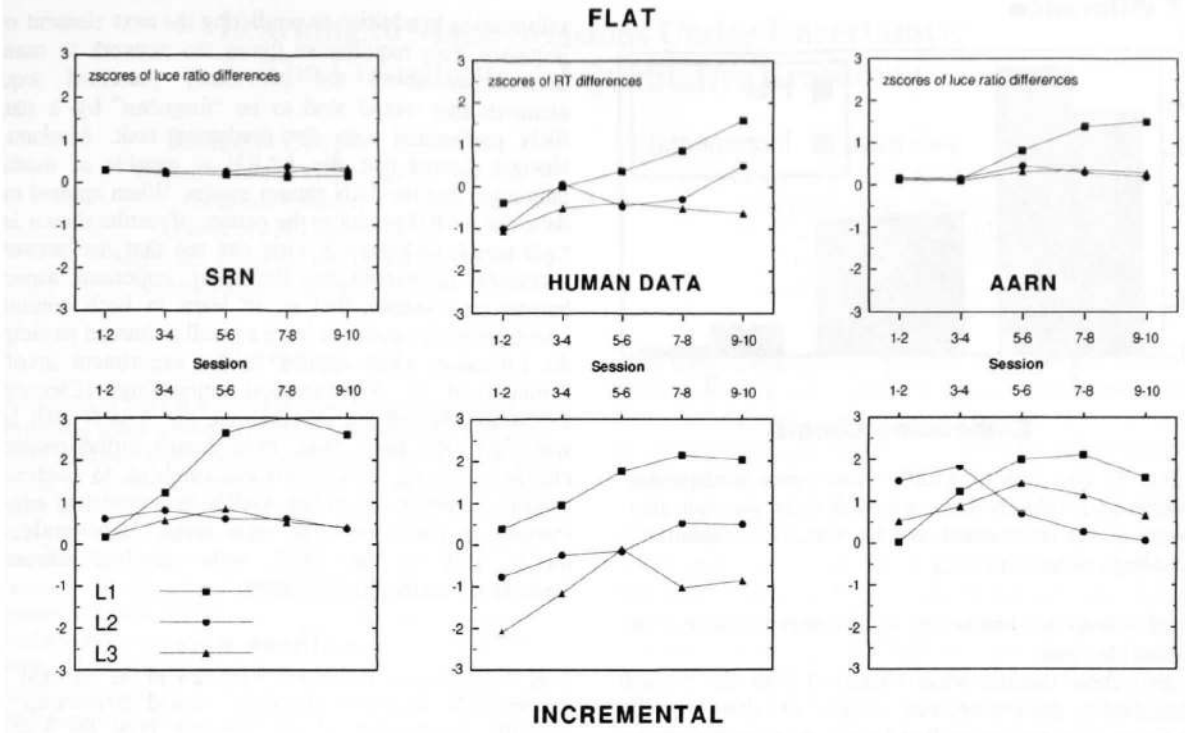


**Figure 4:** Human and simulated data (i.e., average difference scores between responses to likely and unlikely tail elements) for the SRN model (left panels), human subjects (middle panels) and for the AARN (right panels), plotted separately for embeddings of length 1 (L1, squares), 2 (L2, circles), and 3 (L3, triangles). Top panels: Flat condition. Bottom panels: Incremental condition. All data have been separately transformed into zscores based on the data obtained for each source over the two conditions. See the text for additional details.

stimulus material. Are human participants similarly sensitive to this difference? This is the focus of the next section.

### Human performance

The human data are illustrated in the middle panels of Figure 4. One can see that subjects appear to learn in both the Flat condition and the Incremental condition. To determine how well participants were able to discriminate between likely and unlikely tails after embeddings of different lengths, we conducted an ANOVA on the data obtained over the last two sessions of the experiment. These data are represented in Figure 5. The figure indicates (1) that participants appear to be sensitive to the likelihood of tails occurring after embeddings up to length 2, and (2) that performance appears to be quite similar in the two conditions. Averaging over both conditions, likely tails had a 52 msec advantage over unlikely tails when the embedding had a length of 1, and a 28 msec advantage when the embedding was of length 2.

These impressions were confirmed by the results of a mixed-measures ANOVA with condition as a between-subjects factor [Flat vs. Incremental condition) and session [2 levels], embedding length [4 levels], and tail probability [likely vs. unlikely], as repeated measures factors. Condition and session both failed to reach significance. The analysis revealed a significant effect of tail probability, $F(1, 10) = 10.37$, $Mse = 2560.62$, $p < 0.01$, thereby indicating that participants indeed tended to produce faster responses when reacting to a likely tail as compared with their responses to unlikely tails. As suggested by the data shown in Figure 4, however, this sensitivity to tail likelihood interacted significantly with embedding length, $F(3, 30) = 8.81$, $Mse = 624.03$, $p < .001$. Contrasts conducted separately for the different levels of embedding length showed that participants exhibited significant differences between their responses to likely and unlikely tails after embeddings up to length 2, but not for embeddings involving 3 or more elements.

Overall then, in contrast with the SRN data, participants appear to learn the material equally well in both conditions.

## Discussion

How do we learn about disjoint temporal contingencies? Is such learning influenced by the training regimen? In this paper, we addressed these issues by exploring human and simulated performance in an experiment meant to test whether a specific prediction of the SRN model of sequence processing was borne out empirically. The experiment involved assessing reaction time performance on complex sequential material containing center-embedded elements and presented either incrementally or not. Following up on Elman's (1993) work, we first showed that the SRN could only learn the material when its more complex instances were introduced progressively during training. In contrast, the human data showed that participants could learn the

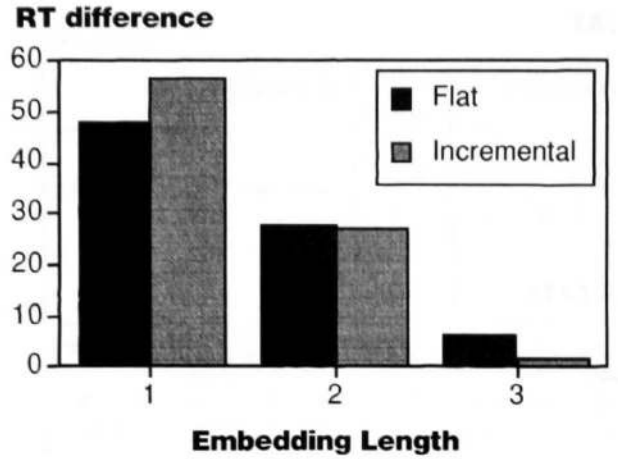


**Figure 5:** Reaction time differences between responses to likely and unlikely tails averaged over the last two sessions of the experiment, and for sequences containing embeddings of length 1 to 3.

material in both conditions, up to instances containing two embedded elements.

Overall, these results, when combined with the research we described in the introduction, suggest (1) that the SRN exhibits specific limitations that human participants are not sensitive to, and (2) that human participants are generally capable of mastering sequential material that the SRN is unable to learn. This does not necessarily make the SRN an inadequate model of human performance in sequential choice reaction time tasks, in that the model clearly captures many central aspects of performance in such situations (see Cleeremans, 1993, for a review), but it should prompt us to look for alternative architectures that build on the SRN's strengths while also overcoming its limitations. One such architecture has recently been proposed by Maskara and Noetzel (1992). Their Auto-Associative Recurrent Network (henceforth, AARN) is illustrated in Figure 6. As its name suggests, this network is essentially an SRN that is also required to act as an encoder on both the current element and the context information. On each time step, the network is thus required to produce the current element and the context

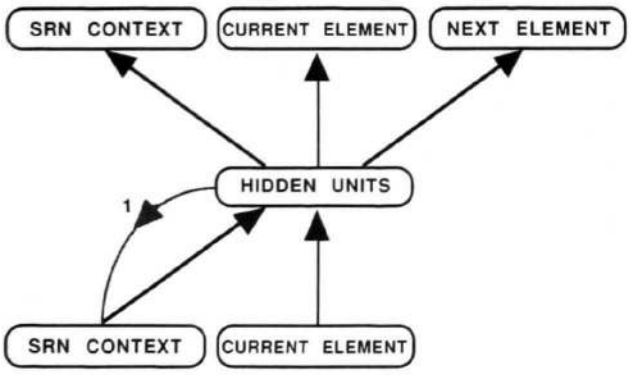


**Figure 6:** The auto-associative recurrent network (AARN). See text for details.

information in addition to predicting the next element of the sequence. This requirement forces the network to maintain information about the previously presented sequence elements that would tend to be "forgotten" by a standard SRN performing only the prediction task. Maskara and Noetzel showed that the AARN is capable of mastering languages that the SRN cannot master. When applied to our data, the AARN produces the pattern of results shown in the right panels of Figure 5. One can see that the network is successful in reproducing the most important aspect of human performance, that is, to learn in both conditions, albeit the model does not learn as well as human participants do. Likewise, when applied to the experiment involving symmetrical vs. asymmetrical embeddings (Cleeremans, 1993) described in the introduction, the AARN also learns where the SRN fails. Thus, even though further research is clearly necessary, such results encourage us to continue to explore the properties of the AARN as a model of sequence learning in choice reaction time tasks. Importantly, this model, just as the SRN, only involves elementary associative learning mechanisms.

## Authors note

Axel Cleeremans is a Research Associate of the National Fund for Scientific Research (Belgium). Arnaud Destrebecqz is a scientific collaborator of the National Fund for Scientific Research (Belgium). This work was supported by FRFC grant #2.4605.95 F to Axel Cleeremans.

## References

Berry, D.C., and Dienes, Z. (1993). Implicit Learning: Theoretical and empirical issues, Erlbaum.

Cleeremans, A. (1993). Mechanisms of Implicit learning: Connectionist models of sequence processing. Cambridge, MA: MIT Press.

Cleeremans, A. & McClelland, J.L. (1991). Learning the structure of event sequences. *JEP:G*, 120, 235-253.

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

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

Jiménez, L, Méndez, C., & Cleeremans, A. (1996). Comparing direct and indirect measures of sequence learning. *JEP:LMC*, 22:948-969.

Jordan, M.I. (1986). Attractor dynamics and parallelism in a connectionist sequential machine. In *Proceedings of the 8th. Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Erlbaum.

Maskara, A., & Noetzel, A. (1992). Forced simple recurrent neural network and grammatical inference. In *Proceedings of the Fourteenth Annual Conference of the Cognitive Science Society*, 420-425, NJ: LEA.

Nissen, M.J. & Bullemer, P. (1987). Attentional requirements of learning: Evidence from performance measures. *Cognitive Psychology*,19,1-32.

Reed, J. & Johnson, P. (1994). Assessing implicit learning with indirect tests: Determining what is learned about sequence structure. *JEP:LMC*, 20:585–594.

Servan-Schreiber, D., Cleeremans, A. & McClelland, J.L. (1991). Graded State Machines: The representation of temporal contingencies in simple recurrent networks. *Machine Learning* , 7:161–193.