Implicit Detection of Event Interdependencies and a PDP Model of the Process

Michael Kushner
Department of Psychology
Brooklyn College of CUNY
Brooklyn, NY 11210
hmkushnr@bklyn.bitnet

Axel Cleeremans
Department of Psychology
Carnegie Mellon University
Pittsburgh, PA 15213
cleeremans@psy.cmu.edu

Arthur Reber
Department of Psychology
Brooklyn College of CUNY
Brooklyn, NY 11210
artreber@bklyn.bitnet

Abstract

We report on an experiment in which subjects were asked to predict the location of a stimulus based on observation of a series of five events. Unbeknownst to subjects, the location of the sixth event was determined by a double contingency between the second and fourth events in the sequence. This material is therefore highly complex, since the relevant events are embedded in a large number of irrelevant contexts. The results indicated that subjects improved their prediction performance over 10 sessions encompassing over 2400 trials of training, despite the fact that they remained completely unaware of the existence of the rule, and unable to verbalize their knowledge of the contingencies in the material. We propose a model of performance in this task, in the form of a PDP model of sequence processing. The model successfully accounts for performance and illustrates how knowledge about the temporal context may develop in a way that does not necessarily yield decomposable representations. Interestingly, the model also predicts that performance would be worse if subjects were required to predict successive events rather than simply observe them.

Introduction

Implicit learning is the process whereby knowledge about complex, rule-governed stimulus environments is acquired without specific intentions to learn and largely independently of conscious knowledge about what was learned (Reber, 1989). This process has been explored in a wide variety of experimental contexts including artificial grammar learning (Reber 1967, 1989), patterned sequence learning (Lewicki, Hill, & Bizot, 1988; Nissen & Bullemer, 1987; Cleeremans & McClelland, in press), concept formation (Brooks, 1978), probability learning (Reber & Millward, 1968, 1971), and process control of simulated manufacturing plants (Berry & Broadbent, 1984). In all cases subjects learn to make decisions, classify novel stimuli, anticipate events, and solve problems that required knowledge of the regularities in the stimulus environment while showing little or no explicit, reportable knowledge about those regularities.

This research was supported by Grant BNS-89-07946 from the National Science Foundation and by a PSC-CUNY Grant from the City University of New York to Arthur S. Reber.

As has been argued elsewhere (Reber, 1989; Cleeremans & McClelland, in press), these experiments all have in common the property that the underlying knowledge base that subjects extract from their interactions with the stimulus environments can be captured by the notion of covariation. That is, subjects' behavior in all of these experiments appears to reflect a single process: the detection of covariations among events as they are instantiated in the stimulus display. The generality of this process is considerable and it has been observed across a wide range of stimulus materials (see Reber, 1989, for a review).

In this paper we explore this notion of the detection of covariation further by introducing a stimulus environment that is based on a complex array of events whose underlying structure is characterized by a remote, double- dependency rule that is far more complex than anything that has been studied to date. If implicit learning is as robust a process as some have suggested (Lewicki, 1986; Lewicki & Hill, 1989; Reber, 1989), then we ought to be able to observe this process emerging in situations where the associative links between events are complex and non-salient.

The procedure used in this study is a relatively simple prediction experiment in which subjects had to "guess" the successor of a sequence of similar events. Subjects were exposed to a series of five stimuli presented successively on a computer monitor and were asked to predict the location of the sixth stimulus. There were three possible locations, arranged as the vertices of a triangle, at which the stimuli could appear. The first five stimuli always appeared at random locations; the location of the sixth stimulus was determined on the basis of the relationship between the locations of the 2nd and 4th stimuli. The 1st, 3rd, and 5th stimuli were always irrelevant.

There are at least two reasons why this task may be quite hard. First, there are more irrelevant events than useful ones. Second, the rule that defines the location of the target stimulus is complex in that it involves a relationship between events rather than the particular events themselves. This results in each component of the rule being instantiated by different pairs of events, each of which may in turn be embedded in a large number of different irrelevant contexts. Nevertheless, we show that subjects do learn double dependencies of this kind, and that they do so independently of any explicit knowledge of the rules. We also show that

once the pattern has been picked up subjects are capable of transferring their knowledge to a "shifted" rule despite the fact that they were unaware of the rule change.

Finally, we present a simulation of the human data using a Parallel Distributed Processing (PDP) model based on the simple recurrent network ("SRN") architecture first introduced by Elman (1990). Cleeremans and McClelland (in press) showed that this model could successfully account for implicit learning of sequential material in a choice reaction situation where the sequences were generated from an artificial grammar. This model therefore appears to be a natural candidate for modeling implicit learning processes in prediction tasks such as the one we describe in this paper. We show that the model is successful in accounting for some (but not all) aspects of performance in our situation. The model illustrates how successful prediction performance may emerge from representations that are not easily decomposable, and thus possibly hard to verbalize. The model also predicts that the particular training conditions used in this experiment are critical for successful performance. We conclude that the SRN model has the potential of giving a reasonable characterization of implicit learning in a variety of stimulus environments.

A complex sequence prediction task

Method

Subjects. Six Brooklyn College undergraduates participated in the experiment. They were paid $40, and received a bonus of one cent per correct prediction beyond chance level.

Apparatus and display. The experiment was run on an IBM microcomputer. The display consisted of three numbered boxes located at the vertices of an invisible inverted triangle, and measuring 6.5 cm X 7.5. cm each. A trial consisted of five successive events. Each event consisted of the appearance of a square stimulus (2.5 cm wide) in one of the boxes. The stimulus remained on screen for 250 msec. The inter-stimulus interval was 250 msec. During a series of 5 events, the stimulus moved from one box to another. After the fifth event had occurred, subjects were asked to predict in which box the stimulus would appear next. They entered their prediction by typing a number between 1 and 3 on the keyboard. The computer then displayed the correct response by presenting the stimulus in the correct box for 2000 msec. Subjects initiated the next trial by pressing the space bar.

Design and stimulus generation. We constructed 18 different random orders of the set of 243 (3^5) possible sequences of 5 stimuli. Each of these 18 sets of 243 sequences was then blocked in 3 groups of 81 trials, for a total of 54 blocks. Subjects were exposed to a total of 4374 trials over 6 days. There was a short pause between any two blocks. All subjects were exposed to the same 18 random orders. Intentionally vague instructions described the experiment as being about "prediction behavior".

The entire experiment was broken down in three phases (the existence of which subjects were naturally kept unaware). Phase I (the "training" phase) consisted of 2430 trials (i.e., 30 blocks of 81 trials). During this phase, the location of the sixth event could be predicted perfectly based on the relationship between the locations at which the second and fourth stimuli of the current trial had appeared. If these stimuli had appeared at the same screen location, then the sixth stimulus appeared in Box 1. If they had been in a clockwise relationship, the sixth stimulus appeared in Box 2. The sixth stimulus appeared in Box 3 if the second and fourth stimuli had been in a counter-clockwise relationship.

In Phase II (the "transfer" phase), the rule was surreptitiously modified by shifting the location of the sixth stimulus by one box for each component of the rule. For instance, sixth events which had appeared in Box 1 during Phase I now appeared in Box 2. Similarly, sixth events which should have appeared in Box 2 now appeared in Box 3, and those which should have appeared in Box 3 now appeared in Box 1. Subjects were asked to make 972 predictions (i.e., 12 blocks of 81 trials) in this phase.

During Phase III (the "random" phase), the rule was again modified: the location at which the sixth stimulus may appear was now simply determined at random. Subjects were again asked to make 972 predictions (i.e., 12 blocks of 81 trials) in this phase of the experiment.

Finally, we conducted extensive interviews with each subject immediately after completion of the experiment.

Results

For each session of the experiment, we computed the average proportion of correct predictions about the location of the sixth stimulus. The results of this analysis are illustrated in Figure 1. Subjects become increasingly better at making accurate predictions over the first 10 sessions of training, and end up reaching about 45% correct responses in the 10th session. This is significantly above chance level (33%)[1], and clearly indicates that subjects have acquired knowledge about the relevant regularities embedded in the material. The second phase begins with a dramatic drop in performance (to chance level), but there is again evidence of learning over the next three sessions ($p < .01$). This suggests that subjects are able to transfer relatively easily from one stimulus-response set to another one. As expected, performance in the third, random, phase is low and fails to be significantly over chance level.

Despite this clear sensitivity to the complex regularities embedded in the material, none of the subjects exhibited explicit knowledge of the sequential structure when asked after the task. Post-experimental interviews revealed that subjects felt frustrated in their attempts to "learn the rule" that determined the location of the sixth stimulus. All subjects reported that they eventually abandoned the search for rules, and started predicting according to their "hunches" or according to "what felt right". Subjects were unable to specify which covariations were crucial in the sequence of five stimuli, not even in a general form such as "when the nth stimulus was in box X, the correct answer was usually

[1] All statistical tests reported in this section were conducted by using a normal approximation to the binomial distribution, at the .01 level.

Y". No subject reported awareness of the rule shifts between the three phases of the experiment.

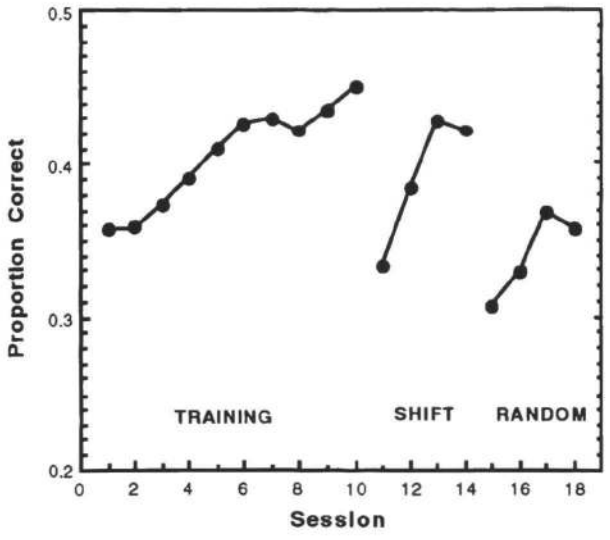


Figure 1: Mean proportion of correct predictions, over the 18 sessions of training, and for the three phases of the experiment ("Training", "Shift", & "Random").

Subjects' poor knowledge of the constraints embedded in the material was also confirmed by their performance on a ranking task, in which they were asked to rate each of the five stimuli in terms of their relevance in predicting the location of the sixth stimulus. The results failed to reveal sensitivity to the crucial events: On a scale of 1 (very important) to 5 (not important), the crucial events received average ranks of 3.5 (2nd event) and 2.67 (4th event), whereas the first, third and fifth events were ranked 3.33, 3.67 and 1.83 respectively. However, there was some evidence that particularly salient sequences which were reported by subjects also elicited very good predictions. For instance, sequences in which the first five stimuli had appeared at the same location always predicted Box 1 as the location of the sixth trial. (i.e., "11111 → 1", "22222 → 1", "33333 → 1"). Similarly, alternating sequences such as "12121" always predicted Box 1 as well. Subjects could correctly predict the successor of repeating sequences in about 61% of the cases of Phase I (49% for the alternating sequences) — considerably better than average. This result clearly indicates that, at least in some specific cases like this, subjects have become aware of some of the regularities embedded in the material. Subjects' successful prediction performance is far from being based only on these salient patterns, however: The average prediction score during Phase I only dropped by .0046 percentage points when the 3 possible "repeating" and 6 possible "single alternating" sequences were eliminated from the analysis. Clearly, subjects have become sensitive to contingencies about which they are unable to report.

A simulation of human prediction performance

This study is a natural candidate for exploring how well a model of sequence processing first introduced by Elman (1990) may be used to simulate human prediction performance. Cleeremans & McClelland (in press) used the "Simple Recurrent Network", or "SRN", to model implicit learning processes in a choice reaction situation. The SRN (Figure 2) is a standard fully connected three-layers back-propagation network (see Rumelhart, Hinton & Williams, 1986), with the added property that the hidden unit layer is allowed to feed back on itself with a delay of one time step, so that the intermediate results of processing at time *t-1* can influence the intermediate results of processing at time *t*. In practice, the SRN is implemented by copying the pattern of activation on the hidden units onto a set of "context units", which feed back into the hidden layer along with the next input. All the forward-going connections in this architecture are modified by back-propagation. The recurrent connections from the hidden layer to the context layer implement a simple copy operation and are not subject to training.

As reported elsewhere (Cleeremans, Servan-Schreiber & McClelland, 1989, in press), we have explored the computational aspects of this architecture in considerable detail. Following Elman (1990), we have shown that an SRN trained to *predict* the successor of each element of a sequence presented one element at a time can learn to perform this "prediction task" perfectly on moderately complex material. For instance, the SRN can learn to predict optimally each element of a continuous sequence generated from small finite-state grammars such as those used by Reber (1989). After training, the network produces responses that closely approximate the optimal conditional probabilities of presentation of all possible successors of the sequence at each step. Note that the network is never presented with more than one element of the sequence at a time. Thus, it has to elaborate its own internal representations of as much temporal context as needed to achieve optimal predictions. Through training, the network progressively comes to discover which features of the previous sequence are relevant to the prediction task.

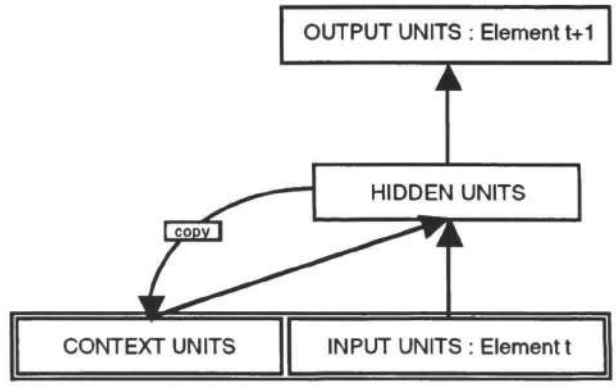


Figure 2: The simple recurrent network (SRN). Adapted from Cleeremans & McClelland (in press).

This architecture — as well as other connectionist architectures with which the SRN shares several basic features — appears to be a good candidate for modeling implicit learning phenomena. For instance, because all the knowledge of the system is stored in its connections, this knowledge may only be expressed through performance. Further, the back-propagation learning procedure implements the kind of elementary associative learning that seems characteristic of many implicit learning processes. However, there is also substantial evidence that knowledge acquired implicitly is very complex and structured (Reber, 1989) — not the kind of knowledge one thinks would emerge from associative learning processes. The work of Elman (1990), in which the SRN architecture was applied to language processing, has demonstrated that the representations developed by the network are highly structured and accurately reflect subtle contingencies, such as those entailed by pronominal reference in complex sentences. Thus, it appears that the SRN embodies two important aspects of implicit learning performance: elementary learning mechanisms that yield complex and structured knowledge. The SRN model shares these characteristics with many other connectionist models, but its specific architecture makes it particularly suitable for processing sequential material.

To model our experimental situation, we made the following assumptions: First, each of the three possible locations at which the stimulus may appear was represented by activating a single unit in either the input pool or the output pool. Second, we assumed that the activations of the output units represent the network's predictions about the location of the next stimulus. Third, since no predictions were required from subjects during presentation of the first five trials of a series, learning was turned off for those trials. The network was therefore merely storing successive events during presentation of the first four trials. When the fifth trial was presented as input to the network, learning was turned on, and the network was trained to activate the output unit corresponding to the location at which the sixth stimulus would appear. Finally, since trials (i.e. blocks of five events) were totally independent from each other, the context units were reset to zero at the beginning of each trial. Thus, the temporal context could influence processing within a block of five events, but it was prevented from carrying over to the next block.

Procedure. Three SRNs with 15 hidden units were trained in exactly the same conditions as human subjects. Each network used a different set of initial random weights, and the learning rate was set to 0.5. (the momentum parameter was not used in these simulations). On each of the 18 sessions of training (10 during Phase I, and 4 each during Phases II and III), each network was exposed to 243 sequences of five events. On each trial, the network was exposed to one of these sequences by activating the input unit corresponding to each event in turn. Note that the network was not trained at this point; it merely processed the sequence of events. When the fifth event was presented, however, learning was turned on, and the network was trained to activate the unit corresponding to the sixth event on its output layer. The error between its prediction and the actual sixth event (as specified by the rules appropriate for the current phase of training) was then computed and back-propagated to modify the weights. To evaluate the model's performance, we recorded the activation of the output units, and determined which unit was most active. A prediction response was considered correct if the activation of the unit corresponding to the actual sixth event was higher than the activation of the two other units. On the next trial, the context units were reset to zero, learning was turned off, and the network was presented with another sequence of five events.

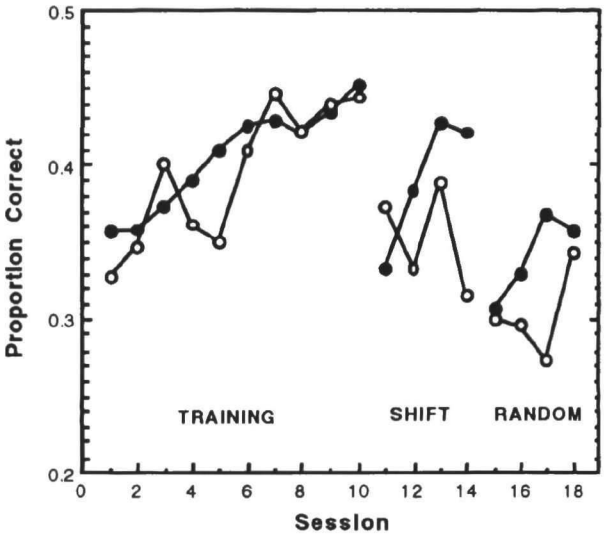
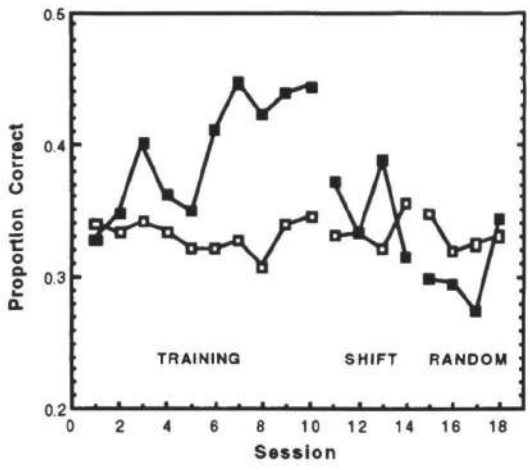


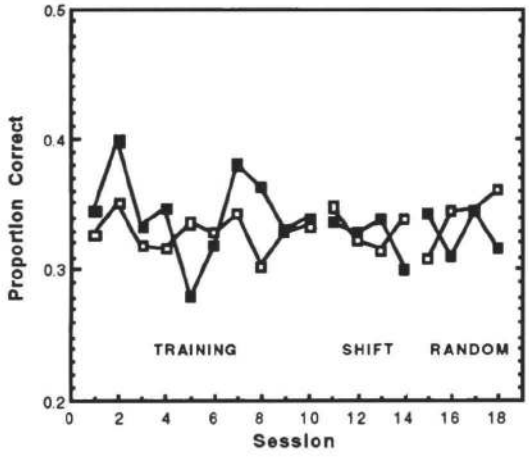
Figure 3: Mean proportion of correct predictions over the 18 sessions of training, and for the three phases of the experiment. Filled symbols represent human data; open symbols represent simulated data (epsilon = 0.5).

Figure 3 compares the human data with the average proportion of correct prediction responses produced by the networks, for each of the 18 sessions of training. It is clear that the model is learning the regularity in Phase I. There are other aspects of the data for which the correspondence with the simulations was far from perfect. In particular, the main discrepancy is located in Phase II: the model isn't nearly as good as the human subjects to readjust its performance to the new rule. This is interesting because it points to a shortcoming shared by most current simulation models of implicit phenomena: their relative inability to transfer to a different set of stimuli and responses which is structurally identical to the original, but instantiated by different tokens. We will return to this point in the discussion.

Figure 4a contrasts the model's performance in predicting the sixth event with its average performance in predicting events 2-5. Recall that the network was not trained to predict successive events during presentation of the first four events. However, it still produced responses, which may be used to compute a baseline "prediction" score against which to compare actual prediction responses of the sixth event. Figure 4a shows that this baseline curve remains flat and at about chance level throughout training, thereby confirming that the network is indeed acquiring information which is specific to predicting the sixth event.



[a]



[b]

Figure 4: Proportion of correct prediction responses plotted separately for Events 2-5 (open symbols) and for Event 6 (filled symbols). [a]: Data from an SRN trained only on presentation of the 5th event. [b]: Data from an SRN trained on all events.

This analysis is most useful when comparing its results with those of the same analysis conducted on the responses of an SRN trained to predict the successor of *each* event. Figure 4b represents these data. Interestingly, the model's performance is much worse under these conditions than when it is trained to predict only the sixth event. Indeed, there is almost no difference between the curves corresponding to predictions of events 2-5 and to event 6: both curves remain flat and at about chance level throughout training.This is because the network is trained to predict random events (events 2 to 5) four times as much as it is trained to predict structured events (event 6). Because of the error minimization resulting from back-propagation, the responses of the network tend to represent the average probability of each output unit to be active *in the entire training set.* Since the probability of an output unit to be on in the entire training set is about 0.33, the responses of the SRN are essentially random when it is trained under these conditions. Thus, the SRN model makes the somewhat counter-intuitive prediction that subjects would fail to learn the contingencies embedded in the material if the task was such that they were required to guess the nature of *each* successive event.

Discussion

Subjects were run on a complex task in which they were required to predict where the sixth event of a series would appear. Unbeknownst to them, the location of the sixth stimulus was determined based on the relationship between the second and fourth events of a sequence. Subjects' prediction performance improved with training, despite the fact that they remained unable to specify which events were relevant, or even that there was any kind of structure present in the material. Further, subjects could successfully transfer the knowledge acquired during training to a different, "shifted", rule.

The simulation work reported in this paper has demonstrated how the SRN model may be applied to this experimental situation. There are several interesting issues here worth commenting on:

First, the results show that the model is capable of learning the complex rule instantiated in the training material, and to do so at about the same rate as human subjects[2]. What are the mechanisms underlying this sensitivity? During the first five events of each trial, the model is merely storing information about each event, in the form of a time-varying pattern of activation over the hidden units. When the fifth trial is presented to the network, this pattern of activation now represents the entire sequence of five events. Different sequences will result in different such representations, even in the absence of any training . This is simply the result of the recurrent nature of the architecture (see Cleeremans, Servan-Schreiber & McClelland, in press, for a discussion of this point). It is on the basis of the differences and similarities between these internal representations that the network becomes capable of predicting the sixth event. To do so, however, it has to learn how to map clusters of internal representations corresponding to sequences resulting in the same prediction onto the output units corresponding to these predictions.

Another, more interesting point, is the observation that the representations developed by the network are completely opaque, in the sense that the information encoded by the network allows successful performance, but is not readily decomposable into critical features. Indeed, it would take sophisticated analysis methods (such as hierarchical clustering, see Cleeremans, Servan-Schreiber & McClelland, 1990) to uncover the regularities embedded in the internal representations developed by the network. This characteristic of the representational system of the SRN, and of PDP networks in general, provides a natural explanation for the fact that human subjects are unable to describe what

[2] Obviously, the learning rate is a free parameter in this model. We conducted a large number of simulations, each with different parameters. In some cases, the network failed to learn. In some other cases, performance was better than in the human data. The simulations discussed in this paper yielded the best fits with the human data.

elements of their knowledge are responsible for successful performance. Whether models are to be taken literally is arguable, particularly when it comes to their representational system, but this provocative interpretation of the SRN's behavior is quite compelling.

A further point is related to training. The results revealed that the network is unable to learn the task when trained on each event of a series. The corresponding experiment with human subjects remains to be done, but it is interesting to speculate on the reasons for this interference. A typical result in implicit learning experiments is that asking subjects to look for structure where there is none results in worse performance (see for instance Reber, 1976). In our situation, asking the network to predict successive events rather than simply observe them appears to have the same effects, and for the same reasons: the model tends to elaborate representations that fail to capture the relevant covariations, because the structure is deeply embedded in many irrelevant contexts.

One aspect of the SRN's performance in this task seems to suggest that it is not yet a complete model of the human data, however. Indeed, the lack of transfer to the new, shifted, rule system used in Phase II of the experiment may be incompatible with the idea that performance is based only on the kind of mechanisms instantiated by the SRN. Why is it hard for the SRN model to adjust its responses to the new rule? The answer lies in an examination of how the SRN model stores and processes knowledge about sequences. Basically, in the case of this particular task, the connections between the input units and the hidden units implement a mapping between sets of sequences of events and distinct internal representations. Optimally, each set of sequences which results in the same prediction about the sixth event should be associated to a unique code on the hidden units. More likely, there is a number of distinct clusters of internal representations, with each cluster grouping those internal representations which result in the same response. The connections between the hidden units and the output units, in turn, map these clusters of internal representations onto the responses. The rule shift introduced in the experiment is in effect a change in this latter mapping. Indeed, it does not result in any change in which events are important, or in the number of possible responses, etc. As a result, all the network really needs to do in order to produce the shifted responses is to adjust the connections from the hidden units to the output units. But, even though weight adjustments are known to be much faster in the last layer of connections than in the others, this process of remapping the internal representations appears to be quite a slow one: even though it would eventually learn the correct new mapping, it appears unable to do so within the limited amount of training available during Phase II.

This difficulty in dealing with "shifted" material exhibited by the SRN model is a very interesting shortcoming. Subjects do indeed seem to be able to readjust their responses to the new shifted rule rather quickly. Another, similarly intriguing transfer result has been repeatedly described in grammar learning experiments. For instance, Reber (1969) reported that transfer performance was significantly over chance with material generated from a grammar which had the same structural properties as the grammar used during training, but used a different set of letters. By contrast (but not surprisingly), transfer performance was much worse in a control condition in which the same set of letters but a different grammar were used. There appears to be no simple way of accounting for this kind of result. The basic problem is that successful models of implicit learning, including the SRN, base their performance on the processing of exemplars. If the representations that these models develop encode the relevant structural properties of the material, they are nevertheless expressed in terms of the exemplars themselves, and not as more abstract characterizations of these structural properties. The fact that subjects do transfer successfully in this situation suggests that some additional mechanisms may play an important role in implicit learning performance. There is no doubt that the current generation of simulation models of implicit learning phenomena will have to address this issue in the future.

References

Berry, D.C. & Broadbent, D.E. (1984). On the relationship between task performance and associated verbalizable knowledge. *QJEP, 36*, 209-231.

Brooks, L.R. (1978). Nonanalytic concept formation and memory for instances. In R. Rosch and B.B. Lloyd (Eds.), *Cognition and Categorization*, New York: Wiley.

Cleeremans, A. & McClelland, J.L. (in press). Learning the structure of event sequences. *JEP: General.*

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

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

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

Lewicki, P. (1986). *Nonconscious Social Information Processing.* New York: Academic Press.

Lewicki, P., Hill, T., & E. Bizot, E. (1988). Acquisition of procedural knowledge about a pattern of stimuli that cannot be articulated. *Cognitive Psychology, 20*, 24-37.

Lewicki, P. & Hill, T. (1989). On the status of nonconscious processes in human cognition: Comment on Reber. *JEP: General, 118*, 239-241.

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

Reber, A.S. (1967). Implicit learning of artificial grammars. *JVLVB, 6*, 855-863.

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

Reber, A.S. (1989). Implicit learning and tacit knowledge. *JEP: General, 118*, 219-235.

Reber, A.S. & Millward, R.B. (1968). Event observation in probability learning. *Journal of Experimental Psychology, 77*, 317-327.

Reber, A.S. & Millward, R.B. (1971). Event tracking in probability learning. *American Journal of Psychology, 84*, 85-99.

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