Journal of Experimental Psychology: General
1991, Vol. 120, No. 3, 235-253
Copyright 1991 by the American Psychological Association, Inc.
0096-3445/91/$3.00

Learning the Structure of Event Sequences

Axel Cleeremans and James L. McClelland
Carnegie Mellon University

How is complex sequential material acquired, processed, and represented when there is no intention to learn? Two experiments exploring a choice reaction time task are reported. Unknown to Ss, successive stimuli followed a sequence derived from a "noisy" finite-state grammar. After considerable practice (60,000 exposures) with Experiment 1, Ss acquired a complex body of procedural knowledge about the sequential structure of the material. Experiment 2 was an attempt to identify limits on Ss ability to encode the temporal context by using more distant contingencies that spanned irrelevant material. Taken together, the results indicate that Ss become increasingly sensitive to the temporal context set by previous elements of the sequence, up to 3 elements. Responses are also affected by priming effects from recent trials. A connectionist model that incorporates sensitivity to the sequential structure and to priming effects is shown to capture key aspects of both acquisition and processing and to account for the interaction between attention and sequence structure reported by Cohen, Ivry, and Keele (1990).

In many situations, learning does not proceed in the explicit and goal-directed way characteristic of traditional models of cognition (Newell & Simon, 1972). Rather, it appears that a good deal of our knowledge and skills are acquired in an incidental and unintentional manner. The evidence supporting this claim is overwhelming: In his recent review article, Reber (1989) analyzes about 40 empirical studies that document the existence of learning processes that do not necessarily entail awareness of the resulting knowledge or of the learning experience itself. At least three different "implicit learning" paradigms have yielded robust and consistent results: artificial grammar learning (Dulany, Carlson, & Dewey, 1984; Mathews et al., 1989; Reber, 1967, 1989; Servan-Schreiber & Anderson, 1990), system control (Berry & Broadbent, 1984; Hayes & Broadbent, 1988), and sequential pattern acquisition (Cohen, Ivry, & Keele, 1990; Lewicki, Czyzewska, & Hoffman, 1987; Lewicki, Hill, & Bizot, 1988; Nissen & Bullemer, 1987; Willingham, Nissen, & Bullemer, 1989). The classic result in these experimental situations is that "subjects are able to acquire specific procedural knowledge (i.e., processing rules) not only without being able to articulate what they have learned, but even without being aware that they had learned anything" (Lewicki et al., 1987, p. 523). Related research with neurologically impaired patients (see Schacter, 1987, for a review) also provides strong evidence for the existence of a functional dissociation between *explicit memory* (conscious recollection) and *implicit memory* (a facilitation of performance without conscious recollection).

Despite this wealth of evidence documenting implicit learning, few models of the mechanisms involved have been proposed. Reber's (1989) analysis of the field, for instance, leaves one with the impression that little has been done beyond mere demonstrations of existence. This lack of formalization can doubtless be attributed to the difficulty of assessing subjects' knowledge when it does not lend itself easily to verbalization. Indeed, although concept formation or traditional induction studies can benefit from experimental procedures that reveal the organization of subjects' knowledge and the strategies they use, such procedures often appear to disrupt or alter the very processes they are supposed to investigate in implicit learning situations (see Dulany et al., 1984; Dulany, Carlson, and Dewey, 1985; Reber, Allen, & Regan, 1985, for a discussion of this point). Thus, research on implicit learning has typically focused more on documenting the conditions under which one might expect the phenomenon to manifest itself than on obtaining the fine-grained data needed to elaborate information-processing models.

Nevertheless, a detailed understanding of such learning processes seems to be an essential preliminary step toward developing insights into the central questions raised by recent research, such as the relationship between task performance and "verbalizable" knowledge, the role that attention plays in unintentional learning, or the complex interactions between conscious thought and the many other functions of the cognitive system. Such efforts at building simulation models of implicit learning mechanisms in specific experimental situations are already underway. For instance, Servan-Schreiber and Anderson (1990) and Mathews et al. (1989) have both developed models of the Reber task that successfully account for key aspects of learning and classification performance.

In this article, we explore performance in a different experimental situation, which has recently attracted increased attention as a paradigm for studying unintentional learning: sequential pattern acquisition. We report on two experiments, which investigate sequence learning in a novel way that allows detailed data on subjects' sequential expectations to be ob-

This research was supported by a grant from the National Fund for Scientific Research (Belgium) to Axel Cleeremans and by a National Institute of Mental Health Research Scientist Development Award to James L. McClelland.

We thank Steven Keele for providing us with details about the experimental data reported in Cohen, Ivry, and Keele (1990) and Emile and David Servan-Schreiber for several insightful discussions. Arthur Reber and an anonymous reviewer contributed many helpful comments.

Correspondence concerning this article should be addressed to Axel Cleeremans, Department of Psychology, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213. Electronic mail may be sent to cleeremans@psy.cmu.edu.

tained, and explore an information-processing model of the task.

Sequence Learning

An increasingly large number of empirical studies have begun to explore the conditions under which one might expect subjects to display sensitivity to sequential structure despite limited ability to verbalize their knowledge. Most of these studies have used a choice reaction time paradigm. Thus, Lewicki et al. (1988) used a four-choice reaction time (RT) task during which the stimulus could appear in one of four quadrants of a computer screen on any trial. Unknown to subjects, the sequential structure of the material was manipulated by generating sequences of 5 elements according to a set of simple rules. Each rule defined where the next stimulus could appear as a function of the locations at which the two previous stimuli had appeared. As the set of sequences was randomized, the first 2 elements of each sequence were unpredictable. By contrast, the last 3 elements of each sequence were determined by their predecessors. Lewicki et al. (1988) hypothesized that this difference would be reflected in response latencies to the extent that subjects are using the sequential structure to respond to successive stimuli. The results confirmed the hypothesis: A progressively widening difference between the number of fast and accurate responses elicited by predictable and unpredictable trials emerged with practice. Furthermore, subjects were exposed to a different set of sequences in a later part of the experiment. These sequences were constructed using the same transition rules, but applied in a different order. Any knowledge about the sequential structure of the material acquired in the first part of the experiment thus became suddenly useless, and a sharp increase in response latency was expected. The results were consistent with this prediction. Yet, when asked after the task, subjects failed to report having noticed any pattern in the sequence of exposures, and none of them even suspected that the sequential structure of the material had been manipulated.

Obviously, repeated exposure to structured material elicits performance improvements that depend specifically on the fact that the material is structured (as opposed to general practice effects). Similar results have been described in different tasks. For instance, Miller (1958) reported higher levels of free recall performance for structured strings over random strings. Hebb (1961) reported an advantage for repeated strings over nonrepeated strings in a recall task, even though subjects were not aware of the repetitive nature of the material. Pew (1974) found that tracking performance was better for a target that followed a consistent trajectory than for a random target. Again, subjects were unaware of the manipulation and failed to report noticing any pattern. More recently, Lewicki et al. (1987) reported improved performance in a search task when combinations of trials as remote as six steps contained information about the location of the target. Other subjects given as much time as they wished to identify the crucial information failed in doing so, thereby suggesting that the relevant patterns were almost impossible to detect explicitly.

However, lack of awareness, or inability to recall the material, does not necessarily entail that these tasks require no attentional capacity. Nissen and Bullemer (1987) demonstrated that a task similar to that used by Lewicki et al. (1988) failed to elicit performance improvements with practice when a memory-intensive secondary task was performed concurrently. More recently, A. Cohen et al. (1990) refined this result by showing that the ability to learn sequential material under attentional distraction interacts with sequence complexity. Only sequences composed entirely of ambiguous elements (i.e., elements that cannot be predicted solely on the basis of their immediate predecessor) are difficult to learn when a secondary task is present.

To sum up, there is clear evidence that subjects acquire specific procedural knowledge when exposed to structured material. When the material is sequential, this knowledge is about the temporal contingencies between sequence elements. Furthermore, it appears that the learning processes underlying performance in sequential choice reaction experiments do not entail or require awareness of the relevant contingencies, although attention is needed to learn even moderately complex material. Several important questions, however, remain unanswered.

First, it is not clear how sensitivity to the temporal context develops over time. How do responses to specific sequence elements vary with practice? Does sensitivity to more or less distant contingencies develop in parallel, or in stages, with the shortest contingencies being encoded earlier than the longer ones? Is there an upper limit to the amount of sequential information that can be encoded, even after considerable practice?

Second, most recent research on sequence processing has used very simple material (but see Lewicki et al., 1987), sometimes even accompanied by explicit cues to sequence structure (Lewicki et al., 1988). Are the effects reported in these relatively simple situations also observed when subjects are exposed to much more complex material involving, for instance, some degree of randomness, or sequence elements that differ widely in their predictability?

Third, and perhaps most important, no detailed information-processing model of the mechanisms involved has been developed to account for the empirical findings reviewed above. In other words, what kind of mechanisms may underlie sequence learning in choice RT situations?

In the rest of this article, we explore the first 2 questions by proposing an answer to the third. We first describe a parallel distributed processing (PDP) model in which processing of events is allowed to be modulated by contextual information. The model learns to develop its own internal representations of the temporal context despite very limited processing resources and produces responses that reflect the likelihood of observing specific events in the context of an increasingly large temporal "window." We then report on two experiments using a choice RT task. Unknown to subjects, successive stimuli followed a sequence derived from a "noisy" finite-state grammar, in which random stimuli were interspersed with structured stimuli in a small proportion of the trials throughout training. This procedure allowed us to obtain detailed data about subjects' expectations after specific stimuli at any point in training. After considerable practice (60,000 exposures) with Experiment 1, subjects acquired a complex body of procedural knowledge about the sequential structure of the material. We analyze this data in detail. Experiment 2 attempts to identify limits on subjects' ability to encode the temporal context by using more distant contingencies that spanned irrelevant material. Next, we argue that the mechanisms implemented in our model may constitute a viable model of implicit learning in sequence learning situations and support this claim by a detailed analysis of the correspondence between the model and our experimental data. Finally, we examine how well the model captures the interaction between attention and sequence structure reported by A. Cohen et al. (1990).

A Model of Sequence Learning

Early research on sequence processing has addressed two related but distinct issues: probability learning situations, in which subjects are asked to predict the next event in a sequence, and choice reaction situations, in which subjects simply respond to the current stimulus, but nevertheless display sensitivity to the sequential structure of the material. Most of the work in this latter area has concentrated on relatively simple experimental situations, such as two-choice reaction time paradigms, and relatively simple effects, such as repetition and stimulus frequency effects. In both cases, most early models of sequence processing (e.g., Estes, 1976; Falmagne, 1965; Laming, 1969; Restle, 1970) have typically assumed that subjects somehow base their performance on an estimation of the conditional probabilities characterizing the transitions between sequence elements, but failed to show how subjects might come to represent or compute them. Laming (1969), for instance, assumed that subjects continuously update running average estimates of the probability of occurrence of each stimulus, on the basis of an arbitrarily limited memory of the sequence. Restle (1970) emphasized the role that explicit recoding strategies play in probability learning, but presumably this work is less relevant in situations for which no explicit prediction responses are expected from the subjects.

Two points seem to be problematic with these early models. First, it seems dubious to assume that subjects actually base their performance on some kind of explicit computation of the optimal conditional probabilities, except possibly in situations in which such computations are required by the instructions (such as in probability learning experiments). In other words, these early models are not process models. They may be successful in providing good descriptions of the data, but fail to give any insights into how processing is actually conducted.

Second, it is not clear how the temporal context gets integrated in these early models. Often, an assumption is made that subjects estimate the conditional probabilities of the stimuli given the relevant temporal context information, but no functional account is provided of how the context information—and how much of it—is allowed to influence processing of the current event.

In the following paragraphs, we present a model that learns to encode the temporal context as a function of whether it is relevant in optimizing performance at the task. The model consists of a simple recurrent network (SRN; see Cleeremans, Servan-Schreiber, & McClelland, 1989; Elman, 1990). The SRN (Figure 1) is a standard, fully connected, three-layer, back-propagation network, 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" that feed into the hidden layer, along with the input units. 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.

At first sight, this architecture appears to be a good candidate for modeling implicit learning phenomena. Indeed, as other connectionist architectures, it has a number of basic features that seem to make it highly appropriate 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, a central characteristic of implicit learning. Furthermore, the back-propagation learning procedure implements the kind of elementary associative learning that also seems characteristic of many implicit learning processes. However, there is also substantial evidence that knowledge acquired implicitly is very complex and structured (Reber, 1989), that is, not the kind of knowledge one thinks would emerge from associative learning processes. The work of Elman (in press), in which the SRN architecture was applied to language processing, 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. In the following paragraphs, we examine how the SRN model is able to encode temporal contingencies.

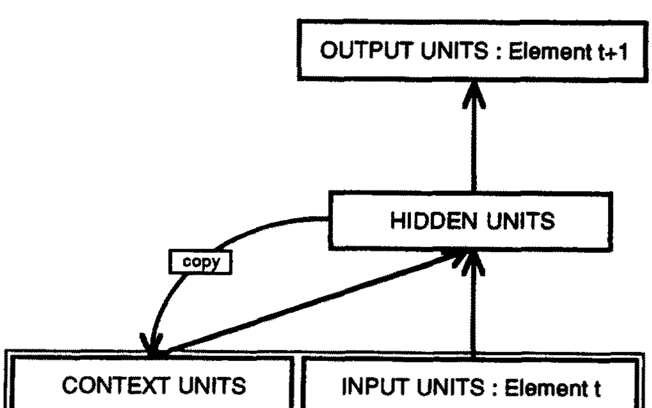

Figure 1. The simple recurrent network (SRN).

As reported elsewhere (Cleeremans et al., 1989), 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 the one represented in Figure 2.[1] 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. Because all letters of the grammar were inherently ambiguous (i.e., optimal predictions required more than the immediate predecessor to be encoded), the network must have developed representations of entire subsequences of events. 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.

A complete analysis of the learning process is beyond the scope of this article (a full account is given in Servan-Schreiber, Cleeremans, & McClelland, 1988), but the key points are as follows: As the initial articles about back-propagation (e.g., Rumelhart, Hinton, & Williams, 1986) pointed out, the hidden unit patterns of activation represent an "encoding" of the features of the input patterns that are relevant to the task. In the 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 learn to maintain prediction-relevant features of an entire sequence of events. Naturally, the actual process through which temporal context is integrated into the representations that the network develops is much more continuous than the above description implies. That is, the "phases of learning" outlined above are but particular points on a continuum.

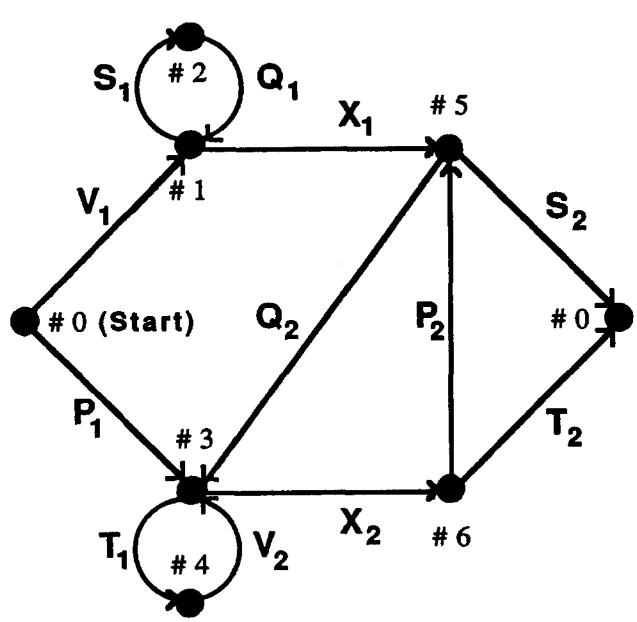

Figure 2. The finite state grammar used to generate the stimulus sequence in Experiment 1. (Note that the first and last nodes are one and the same.)

To summarize, learning and processing in the SRN model have several properties that make it attractive as an architecture for sequence learning. First, the model only develops sensitivity to the temporal context if it is relevant in optimizing performance on the current element of the sequence. As a result, there is no need to make specific assumptions regarding the size of the temporal window that the model is allowed to receive input from. Rather, the size of this self-developed window appears to be essentially limited by the complexity of the sequences to be learned by the network. Representational resources (i.e., the number of hidden units available for processing) are also limiting factors, but only marginal ones. Second, the model makes minimal assumptions regarding processing resources: Its architecture is elementary, and all computations are local to the current element (i.e., there is no explicit representation of the previous elements). Processing is therefore strongly driven by the constraints imposed by the prediction task. As a consequence, the model tends to become sensitive to the temporal context in a very gradual way and will tend to fail to discriminate between the successors of identical subsequences preceded by disambiguating predecessors when the embedded material is not itself dependent on the preceding information. We return to this last point in the General Discussion section.

To evaluate the model as a theory of human learning in sequential choice reaction time situations, we assumed (a) that the activations of the output units represent response tendencies and (b) that the RT to a particular response is proportional to some function of the activation of the corresponding output unit. The specific instantiations of these assumptions that are adopted in this research are detailed later. With these assumptions in place, the model produces responses that can be directly compared with experimental data. In the following sections, we report on two experiments that were designed to allow for such detailed comparisons to be conducted.

[1] In a finite-state grammar, sequences can be generated by randomly choosing an arc among the possible arc emanating from a particular node and by repeating this process with the node pointed to by the selected arc. A continuous sequence can be generated by assuming that the grammar loops onto itself, that is, that its first and last nodes are one and the same.

Experiment 1

Subjects were exposed to a six-choice RT task. The entire experiment was divided into 20 sessions. Each session consisted of 20 blocks of 155 trials. On any of the 60,000 recorded trials, a stimulus could appear at one of six positions arranged in a horizontal line on a computer screen. The task consisted of pressing as fast and as accurately as possible on one of six corresponding keys. Unknown to subjects, the sequential structure of the stimulus material was manipulated. Stimuli were generated using a small finite-state grammar that defined legal transitions between successive trials. Some of the stimuli, however, were not "grammatical." On each trial, there was a 15% chance of substituting a random stimulus to the one prescribed by the grammar. This "noise" served two purposes. First, it ensured that subjects could not simply memorize the sequence of stimuli and hindered their ability to detect regularities in an explicit way. Second, because each stimulus was possible on every trial (if only in a small porportion of the trials), we could obtain detailed information about what stimuli subjects did or did not expect at each step.

If subjects become increasingly sensitive to the sequential structure of the material over training, one would thus predict an increasingly large difference in the RTs elicited by predictable and unpredictable stimuli. Furthermore, detailed analyses of the RTs to particular stimuli in different temporal contexts should reveal differences that reflect subjects' progressive encoding of the sequential structure of the material.

Method

Subjects. Six subjects (Carnegie Mellon University [CMU] staff and students), aged 17–42, participated in the experiment. Subjects were each paid $100 for their participation in the 20 sessions of the experiment and received a bonus of up to $50 on the basis of speed and accuracy.

Apparatus and display. The experiment was run on a Macintosh II computer. The display consisted of six dots arranged in a horizontal line on the computer's screen and separated by intervals of 3 cm. At a viewing distance of 57 cm, the distance between any two dots subtended a visual angle of 3°. Each screen position corresponded to a key on the computer's keyboard. The spatial configuration of the keys was entirely compatible with the screen positions (i.e., the leftmost key corresponded to the leftmost screen position, and so on). The stimulus was a small black circle 0.40 cm in diameter that appeared centered 1 cm below one of the six dots. The timer was started at the onset of the stimulus and stopped by the subjects' response. The response–stimulus interval was 120 ms.

Procedure. Subjects received detailed instructions during the first meeting. They were told that the purpose of the experiment was to "learn more about the effect of practice on motor performance." Both speed and accuracy were stressed as being important. After receiving the instructions, subjects were given three practice blocks of 15 random trials each at the task. A schedule for the 20 experimental sessions was then set up. Most subjects followed a regular schedule of 2 sessions a day.

The experiment itself consisted of 20 sessions of 20 blocks of 155 trials each. Each block was initiated by a *get ready* message and a warning beep. After a short delay, 155 trials were presented to the subjects. The first 5 trials of each block were entirely random so as to eliminate initial variability in the responses. These data points were not recorded. The next 150 trials were generated according to the procedure described below (in the *Stimulus material* section). Errors were signaled to the subjects by a short beep. After each block, the computer paused for approximately 30 s. The message *rest break* was displayed on the screen, along with information about subjects' performance. This feedback consisted of the mean RT and accuracy values for the last block and of information about how these values compared with those for the next-to-last block. If the mean RT for the last block was within a 20-ms interval of the mean RT for the next-to-last block, the words *as before* were displayed; otherwise, either *better* or *worse* appeared. A 2% interval was used for accuracy. Finally, subjects were also told about how much they had earned during the last block and during the entire session up to the last block. Bonus money was allocated as follows: Each reaction time under 600 ms was rewarded by .078¢, and each error entailed a penalty of 1.11¢. These values were calculated so as to yield a maximum of $2.50 per session.

Stimulus material. Stimuli were generated on the basis of the small finite-state grammar shown in Figure 2. Finite-state grammars consist of nodes connected by labeled arcs. Expressions of the language are generated by starting at Node #0, choosing an arc, recording its label, and repeating this process with the next node. Note that the grammar loops onto itself: The first and last nodes, both denoted by the digit 0, are actually the same. The vocabulary associated with the grammar consists of six letters (T, S, X, V, P, and Q), each represented twice on different arcs (as denoted by the subscript on each letter). This results in highly context-dependent transitions, as identical letters can be followed by different sets of successors as a function of their position in the grammar (For instance, S_1 can only be followed by Q, but S_2 can be followed by either V or P). Finally, the grammar was constructed so as to avoid direct repetitions of a particular letter, because it is known (Bertelson, 1961; Hyman, 1953) that repeated stimuli elicit shorter RTs independently of their probability of presentation. (Direct repetitions can still occur because a small proportion of the trials were generated randomly, as described below.)

Stimulus generation proceeded as follows. On each trial, three steps were executed in sequence. First, an arc was selected at random among the possible arcs coming out of the current node, and its corresponding letter recorded. The current node was set to be Node #0 on the sixth trial of any block and was updated on each trial to be the node pointed to by the selected arc. Second, in 15% of the cases, another letter was substituted to the letter recorded at Step 1 by choosing it at random among the five remaining letters in the grammar. Third, the selected letter was used to determine the screen position at which the stimulus would appear. A 6 × 6 Latin square design was used, so that each letter corresponded to each screen position for exactly 1 of the 6 subjects. (Note that subjects were never presented with the actual letters of the grammar.)

Postexperimental interviews. All subjects were interviewed after completion of the experiment. The experimenter asked a series of increasingly specific questions in an attempt to gain as much information about subjects' explicit knowledge of the manipulation and the task.

Results and Discussion

Task performance. Figure 3 shows the average RTs on correct responses for each of the 20 experimental sessions, plotted separately for predictable and unpredictable trials. We discarded responses to repeated stimuli (which are necessarily ungrammatical) because they elicit fast RTs independently of their probability of presentation, as discussed above. Figure 3

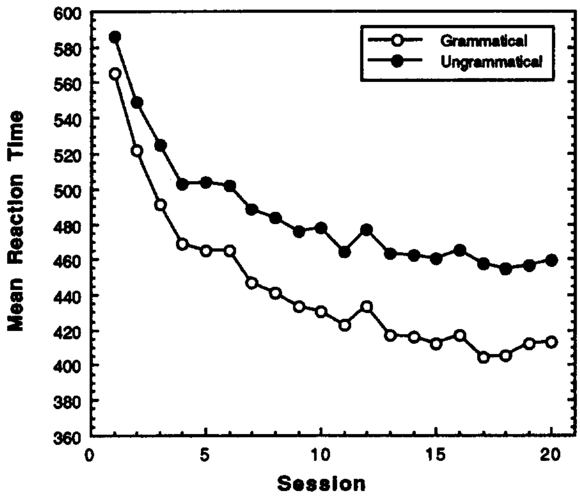

Figure 3. Mean reaction times for grammatical and ungrammatical trials for each of the 20 sessions of Experiment 1.

shows that a general practice effect is readily apparent, as is an increasingly large difference between predictable and unpredictable trials. A two-way analysis of variance (ANOVA) with repeated measures on both factors (practice [20 levels] by trial type [grammatical vs. ungrammatical]) revealed significant main effects of practice, $F(19, 95) = 9.491$, $p < .001$, $MS_e = 17710.45$, and of trial type, $F(1, 5) = 105.293$, $p < .001$, $MS_e = 104000.07$, as well as a significant interaction, $F(19, 95) = 3.022$, $p < .001$, $MS_e = 183.172$. It appears that subjects become increasingly sensitive to the sequential structure of the material. To assess whether the initial difference between grammatical and ungrammatical trials was significant, a similar analysis was conducted on the data from the first session only, using the 20 blocks of this session as the levels of the practice factor. This analysis revealed that there were significant main effects of practice, $F(19, 95) = 4.006$, $p < .001$, $MS_e = 2634.295$, and of trial type, $F(1, 5) = 8.066$, $p < .05$, $MS_e = 3282.914$, but no interaction, $F(19, 95) = 1.518$, $p > .05$, $MS_e = 714.558$. We provide an interpretation for this initial difference when examining the model's performance.

Accuracy averaged 98.12% over all trials. Subjects were slightly more accurate on grammatical trials (98.40%) than on ungrammatical trials (96.10%) throughout the experiment. A two-way ANOVA with repeated measures on both factors (practice [20 levels] by trial type [grammatical vs. ungrammatical]) confirmed this difference, $F(1, 5) = 7.888$, $p < .05$, $MS_e = .004$. The effect of practice did not reach significance, $F(19, 95) = .380$, $p > .05$, $MS_e = .0003$; neither did the interaction, $F(19, 95) = .727$, $p > .05$, $MS_e = .00017$.

Postexperimental interviews. Each subject was interviewed after completion of the experiment. We loosely followed the scheme used by Lewicki et al. (1988). Subjects were first asked about "whether they had anything to report regarding the task." All subjects reported that they felt their performance had improved a lot during the 20 sessions, but much less so in the end. Two subjects reported that they felt frustrated because of the lack of improvement in the last sessions.

Next, subjects were asked whether they "had noticed anything special about the task or the material." This question failed to elicit more detailed reports. All subjects tended to repeat the comments they had given in answering the first question.

Finally, subjects were asked directly whether they "had noticed any regularity in the way the stimulus was moving on the screen." All subjects reported noticing that short sequences of alternating stimuli did occur frequently. When probed further, 5 subjects were able to specify that they had noticed two pairs of positions between which the alternating pattern was taking place. On examination of the data, it appeared that these reported alternations corresponded to the two small loops on Nodes #2 and #4 of the grammar. One subject also reported noticing another more complex pattern between three positions, but was unable to specify the exact locations when asked. All subjects felt that the sequence was random when not involving these salient patterns. When asked whether they "had attempted to take advantage of the patterns they had noticed in order to anticipate subsequent events," all subjects reported that they had attempted to do so at times (for the shorter patterns), but that they felt that it was detrimental to their performance as it resulted in more errors and slower responses. Thus, it appears that subjects only had limited reportable knowledge of the sequential structure of the material and that they tried not to use what little knowledge they had.

Gradual encoding of the temporal context. As discussed earlier, one mechanism that would account for the progressive differentiation between predictable and unpredictable trials consists of assuming that subjects, in attempting to optimize their responses, progressively come to prepare for successive events on the basis of an increasingly large temporal context set by previous elements of the sequence. In the grammar we used, the uncertainty associated with the next element of the sequence can, in most cases, be optimally reduced by encoding two elements of temporal context. However, some sequence elements require three or even four elements of temporal context to be optimally disambiguated. For instance, the path *SQ* (leading to Node #1) occurs only once in the grammar and can only be legally followed by S or by X. In contrast, the path *TVX* can lead to either Node #5 or Node #6 and is therefore not sufficient to perfectly distinguish between stimuli that occur only (in accordance with the grammar) at Node #5 (S or Q) and stimuli that occur only at Node #6 (T or P). One would assume that subjects initially respond to the contingencies entailed by the shortest paths and progressively become sensitive to the higher order contingencies as they encode more and more temporal context.

A simple analysis that would reveal whether subjects are indeed basing their performance on an encoding of an increasingly large temporal context was conducted. The analysis' general principle consists of comparing the data with the probability of occurrence of the stimuli, given different amounts of temporal context.

First, we estimated the overall probability of observing each letter as well as the conditional probabilities (CPs) of observing each letter as the successor of every grammatical path of length 1, 2, 3, and 4, respectively. This was achieved by generating 60,000 trials in exactly the same way as during the experiment and by recording the probability of observing every letter after every observed sequence of every length up to four elements. Only grammatical paths (i.e., sequences of letters that conform to the grammar) were then retained for further analysis. There are 70 such paths of length 4, each possibly followed by each of the six letters, thus yielding a total of 420 data points. There are fewer types of shorter paths, but each occurs more often.

Next, the set of average correct RTs for each successor to every grammatical path of length 4 was computed, separately for groups of four successive experimental sessions.

Finally, 25 separate regression analyses were conducted, using each of the five sets of CPs (0–4) as predictors, and each of the five sets of mean RTs as dependent variables. Because the human data are far from being perfectly reliable at this level of detail, the obtained correlation coefficients were then corrected for attenuation. Reliability was estimated by the split-halves method (Carmines & Zeller, 1987), using data from even and odd experimental blocks.

Figure 4 illustrates the results of these analyses. Each point on the figure represents the corrected r^2 of a specific regression analysis. Points corresponding to analyses conducted with the same amount of temporal context (0–4 elements) are linked together.

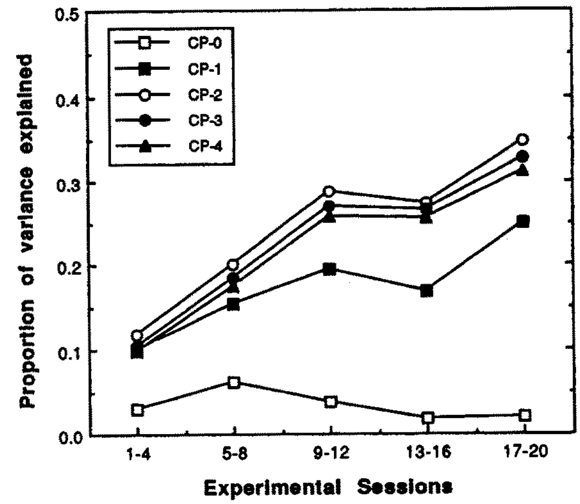

Figure 4. Correspondence between the human responses and conditional probabilities (CP) after paths of length 0–4 during successive blocks of four simulated sessions.

If subjects are encoding increasingly large amounts of temporal context, we would expect the variance in the distribution of their responses at successive points in training to be better explained by CPs of increasingly higher statistical orders. Although the overall fit is rather low (note that the vertical axis only extends to 0.5), Figure 4 nevertheless reveals the expected pattern: First, the correspondence between human responses and the overall probability of appearance of each letter (CP–0) is very close to zero. This clearly indicates that subjects are responding on the basis of an encoding of the constraints imposed by previous elements of the sequence. Second, one can see that the correspondence with the first-order CPs tends to level off below the fits for the second, third, and fourth orders early in training. By contrast, the correspondence between the data and the higher order CPs keeps increasing throughout the entire experiment. The fits to the second-, third-, and fourth-order paths are highly similar in part because their associated CPs are themselves highly similar. This in turn is due to the fact that only a small proportion of sequence elements are ambiguous up to the third or fourth position. Furthermore, even though the data may appear to be most closely consistent with the second order CPs throughout the task, a separate analysis restricted to the first 4 sessions of training indicated that the first-order CPs were the best predictor of the data in the first 2 sessions. Finally, it is still possible that deviations from the second-order CPs are influenced by the constraints reflected in the third- or even fourth-order CPs. The next section addresses this issue.

Sensitivity to long-distance temporal contingencies. To assess more directly whether subjects are able to encode three or four letters of temporal context, several analyses on specific successors of specific paths were conducted. One such analysis involved several paths of length 3. These paths were the same in their last two elements, but differed in their first element as well as in their legal successors. For example, we compared XTV with PTV and QTV and examined RTs for the letters *S* (legal only after XTV) and *T* (legal only after PTV or QTV). If subjects are sensitive to three letters of context, their response to an S should be relatively faster after XTV than in the other cases, and their response to a T should be relatively faster after PTV or QTV than after XTV. Similar contrasting contexts were selected in the following manner: First, as described above, we only considered grammatical paths of length 3 that were identical but for their first element. Specific ungrammatical paths are too infrequent to be represented often enough in individual subject's data. Second, some paths were eliminated to control for priming effects to be discussed later. For instance, the path VTV was eliminated from the analysis because the alternation between V and T favors a subsequent T. This effect is absent in contrasting cases, such as XTV, and may thus introduce biases in the comparison. Third, specific successors to the remaining paths were eliminated for similar reasons. For instance, we eliminated S from comparisons on the successors of SQX and PQX because both Q and S prime S in the case of SQX but not in the case of PQX. As a result of this residual priming, the response to S after SQX tends to be somewhat faster than what would be predicted on the basis of the grammatical constraints only, and the comparison is therefore contaminated. These successive eliminations left the following contrasts available for further analysis: SQX–Q and PQX–T (grammatical) versus SQX–T and PQX–Q (ungrammatical); SVX–Q and TVX–P versus SVX–P and TVX–Q; and XTV–S, PTV–T, and QTV–T versus XTV–T, PTV–S, and QTV–S.

Figure 5 shows the RTs elicited by grammatical and ungrammatical successors of these remaining paths, averaged over blocks of 4 successive experimental sessions. The figure reveals that there is a progressively widening difference between the two curves, thereby suggesting that subjects become increasingly sensitive to the contingencies entailed by elements of the temporal context as removed as three elements

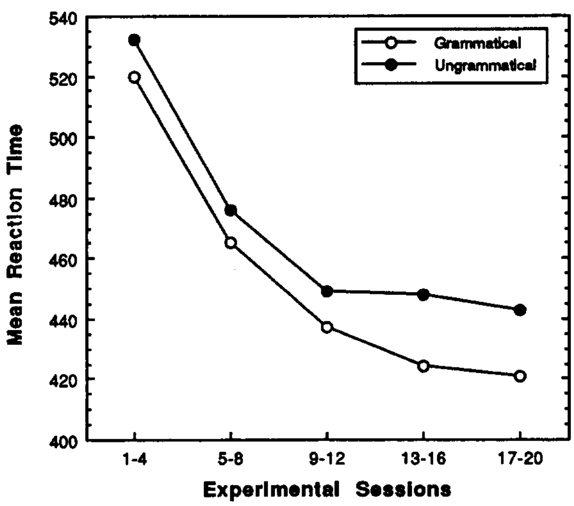

Figure 5. Mean reaction times for predictable and unpredictable successors of selected paths of length 3 and for successive blocks of four experimental sessions.

from the current trial. A two-way ANOVA with repeated measures on both factors (practice [four levels] by successor type [grammatical vs. ungrammatical]) was conducted on these data and revealed significant main effects of successor type, $F(1, 5) = 7.265, p < .05, MS_e = 530.786$, and of practice, $F(4, 20) = 11.333, p < .001, MS_e = 1602.862$. The interaction just missed significance, $F(4, 20) = 2.530$, $p < .07$, $MS_e = 46.368$, but it is obvious that most of the effect is located in the later sessions of the experiment. This was confirmed by the results of a one-tailed paired t test conducted on the difference between grammatical and ungrammatical successors, pooled over the first 8 and the last 8 sessions of training. The difference score averaged −11.3 ms early in training and −22.8 ms late in training. It was significantly bigger late in training, $t(5) = -5.05$, $p < .005$. Thus, there appears to be evidence of a gradually increasing sensitivity to at least three elements of temporal context.

A similar analysis was conducted on selected paths of length 4. After selecting candidate contexts as described above, the following paths remained available for further analysis: XTVX-S, XTVX-Q, QTVX-T, QTVX-P, PTVX-T, and PTVX-P (grammatical) versus XTVX-T, XTVX-P, QTVX-S, QTVX-Q, PTVX-S, and PTVX-Q (ungrammatical). No sensitivity to the first element of these otherwise identical paths of length 4 was found, even during Sessions 17–20: A paired, one-tailed t test on the difference between grammatical and ungrammatical successors failed to reach significance $t(5) = .076$, $p > .1$. Although one cannot reject the idea that subjects would eventually become sensitive to the constraints set by temporal contingencies as distant as four elements, there is no indication that they do so in this situation.

Experiment 2

Experiment 1 demonstrated that subjects progressively become sensitive to the sequential structure of the material and seem to be able to maintain information about the temporal context for up to three steps. The temporal contingencies characterizing this grammar were relatively simple, however, because in most cases, only two elements of temporal context are needed to disambiguate the next event perfectly.

Furthermore, contrasting, long-distance dependencies were not controlled for their overall frequency. In Experiment 2, a more complex grammar (Figure 6) was used in an attempt to identify limits on subjects' ability to maintain information about more distant elements of the sequence. In this grammar, the last element (A or X) is contingent on the first one (also A or X). Information about the first element, however, has to be maintained across either of the two identical embeddings in the grammar and is totally irrelevant for predicting the elements of the embeddings. Thus, to accurately prepare for the last element at Nodes #11 or #12, one needs to maintain information for a minimum of four steps. Accurate expectations about the nature of the last element would be revealed by a difference in the RT elicited by the letters *A* and *X* at Nodes #11 and #12 (A should be faster than X at Node #11 and vice versa). Naturally, there was again a 15% chance of substituting another letter for the one prescribed by the grammar. Furthermore, a small loop was inserted at Node #13 so as to avoid direct repetitions between the letters that precede and follow Node #13. One random letter was always presented at this point; after which there was a 40% chance of staying in the loop on subsequent steps.

Finally, to obtain more direct information about subjects' explicit knowledge of the training material, we asked them to try to generate the sequence after the experiment was completed. This "generation" task involved exactly the same stimulus sequence generation procedure as during training. On every trial, subjects had to press on the key corresponding to the location of the next event.

Method

The design of Experiment 2 was almost identical to that of Experiment 1. The changes are detailed below.

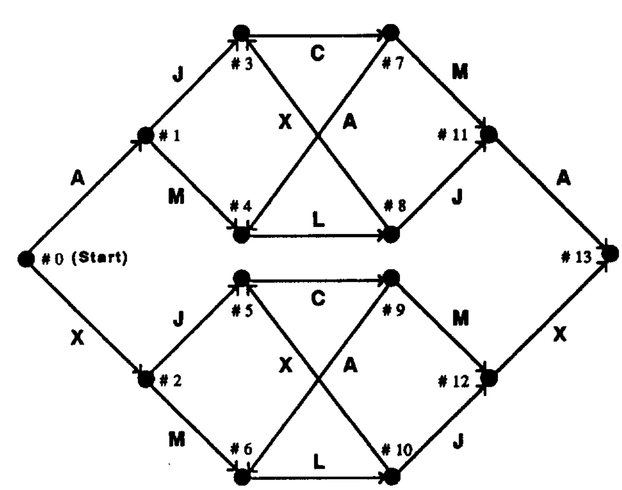

Figure 6. The finite state grammar used to generate the stimulus sequence in Experiment 2.

Subjects. Six new subjects (CMU undergraduates and graduates), aged 19–35, participated in Experiment 2.

Generation task. Experiment 1 did not include any strong test of subjects' verbalizable knowledge about the stimulus material. In the present experiment, we attempted to remedy this situation by using a generation task inspired by Nissen and Bullemer (1987). After completing the 20 experimental sessions, subjects were informed of the nature of the manipulation and asked to try to predict the successor of each stimulus. The task consisted of three blocks of 155 trials of events generated in exactly the same way as during training. (As during the experiment itself, the 5 initial random trials of each block were not recorded.) On each trial, the stimulus appeared below one of the six screen positions, and subjects had to press on the key corresponding to the position at which they expected the next stimulus to appear. Once a response had been typed, a cross 0.40 cm in width appeared centered 1 cm above the screen position corresponding to the subjects' prediction, and the stimulus was moved to its next location. A short beep was emitted by the computer on each error. Subjects were encouraged to be as accurate as possible.

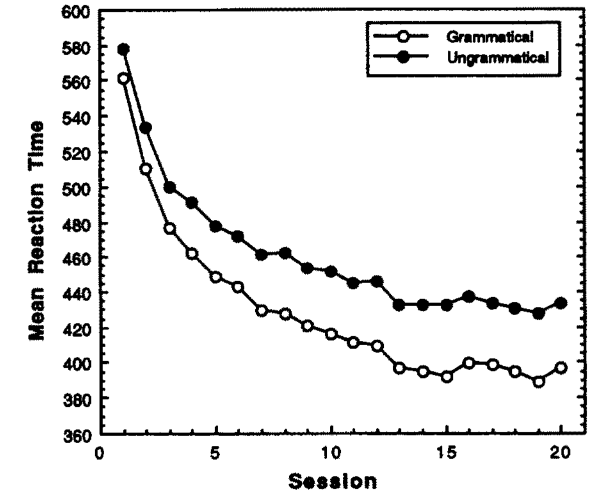

Figure 7. Mean reaction times for grammatical and ungrammatical trials for each of the 20 sessions of Experiment 2.

Results and Discussion

Task performance. Figure 7 shows the main results of Experiment 2. They closely replicate the general results of Experiment 1, although subjects were a little bit faster overall in Experiment 2. A two-way ANOVA with repeated measures on both factors (practice [20 levels] by trial type [grammatical vs. ungrammatical]) again revealed significant main effects of practice, $F(19, 95) = 32.011$, $p < .001$, $MS_e = 21182.79$, and of trial type, $F(1, 5) = 253.813$, $p < .001$, $MS_e = 63277.53$, as well as a significant interaction, $F(19, 95) = 4.670$, $p < .001$, $MS_e = 110.862$. A similar analysis conducted on the data from only the first session again revealed significant main effects of practice, $F(19, 95) = 4.631$, $p < .001$, $MS_e = 1933.331$, and of trial type, $F(1, 5) = 19.582$, $p < .01$, $MS_e = 861.357$, but no interaction, $F(19, 95) = 1.383$, $p > .1$, $MS_e = 343.062$.

Accuracy averaged 97% over all trials. Subjects were again slightly more accurate on grammatical (97.60%) than on ungrammatical (95.40%) trials. However, a two-way ANOVA with repeated measures on both factors (practice [20 levels] by trial type [grammatical vs. ungrammatical]) failed to confirm this difference, $F(1, 5) = 5.351$, $p > .05$, $MS_e = .005$. The effect of practice did reach significance, $F(19, 95) = 4.112$, $p < .001$, $MS_e = .00018$, but not the interaction, $F(19, 95) = 1.060$, $p > .05$, $MS_e = .00008$. Subjects became more accurate on both grammatical and ungrammatical trials as the experiment progressed.

Sensitivity to long-distance temporal contingencies. Of greater interest are the results of analyses conducted on the responses elicited by the successors of the four shortest paths starting at Node #0 and leading to either Node #11 or Node #12 (AJCM, AMLJ, XJCM, and XMLJ). Among those paths, those beginning with A predict A as their only possible successor and vice versa for paths starting with X. Because the subpaths *JCM* and *MLJ* undifferentially predict A or X as their possible successors, subjects need to maintain information about the initial letter to accurately prepare for the successors. The RTs on legal successors of each of these four paths (i.e., A for AJCM and AMLJ and X for XJCM and XMLJ) were averaged together and compared with the average RT on the illegal successors (i.e., X for AJCM and AMLJ and A for XJCM and XMLJ), thus yielding two scores. Any significant difference between these two scores would mean that subjects are disciminating between legal and illegal successors of these four paths, thereby suggesting that they have been able to maintain information about the first letter of each path over three irrelevant steps. The mean RT on legal successors over the last four sessions of the experiment was 385, and the corresponding score for illegal successors was 388. A one-tailed paired t test on this difference failed to reach significance, $t(5) = 0.571$, $p > .05$. Thus, there is no indication that subjects were able to encode even the shortest long-distance contingency of this type.

Generation task. To determine whether subjects were better able to predict grammatical elements than ungrammatical elements after training, a two-way ANOVA with repeated measures on both factors (practice [three levels] by trial type [grammatical vs. ungrammatical]) was conducted on the accuracy data of 5 subjects (one subject had to be eliminated because of a technical failure).

For grammatical trials, subjects averaged 23.00%, 24.40%, and 26.20% correct predictions for the three blocks of practice, respectively. The corresponding data for the ungrammatical trials were 18.4%, 13.8%, and 20.10%. Chance level was 16.66%. It appears that subjects are indeed better able to predict grammatical events than ungrammatical events. The ANOVA confirmed this effect: There was a significant main effect of trial type, $F(1, 4) = 10.131$, $p < .05$, $MS_e = .004$, but no effect of practice, $F(2, 8) = 1.030$, $p > .05$, $MS_e = .004$, and no interaction, $F(2, 8) = .1654$, $p > .05$, $MS_e = .001$. Although overall accuracy scores are very low, these results nevertheless clearly indicate that subjects have acquired some explicit knowledge about the sequential structure of the material in the course of training. This is consistent with previous studies (A. Cohen et al., 1990; Willingham et al., 1989) and not surprising given the extensive training to which subjects have been exposed. At the same time, it is clear that whatever knowledge was acquired during training is of limited use in predicting grammatical elements, because subjects were only able to do so in about 25% of the trials of the generation task.

Simulation of the Experimental Data

Taken together, the results of both experiments suggest that subjects do not appear to be able to encode long-distance dependencies when they involve four elements of temporal context (i.e., three items of embedded independent material); at least, they cannot do so under the conditions used here. However, there is clear evidence of sensitivity to the last three elements of the sequence (Experiment 1). Furthermore, there is evidence for a progressive encoding of the temporal context information: Subjects rapidly learn to respond on the basis of more than the overall probability of each stimulus and become only gradually sensitive to the constraints entailed by higher order contingencies.

Application of the SRN Model

To model our experimental situation, 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 six stimuli). The network was trained to predict each element of a continuous sequence of stimuli generated in exactly the same conditions as for human subjects in Experiment 1. On each step, a letter was generated from the grammar as described in the Method section of Experiment 1 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 activation of each output unit was recorded on every trial and transformed into Luce ratios (Luce, 1963) to normalize the responses.[2] For the purpose of comparing the model's and the subjects' responses, we assumed (a) that the normalized activations of the output units represent response tendencies and (b) that there is a linear reduction in RT proportional to the relative strength of the unit corresponding to the correct response.

This data was first analyzed in the same way as for Experiment 1 subjects and compared with the CPs of increasingly higher statistical orders in 20 separate regression analyses. The results are illustrated in Figure 8.

In stark contrast with the human data (Figure 4; note the scale difference), the variability in the model's responses appears to be very strongly determined by the probabilities of particular successor letters given the temporal context. Figure 8 also reveals that the model's behavior is dominated by the first-order CPs for most of the training, but that it becomes progressively more sensitive to the second- and higher order CPs. Beyond 60,000 exposures, the model's responses come to correspond most closely to the second-, then third-, and then finally fourth-order CPs.

Figure 9 illustrates a more direct comparison between the model's responses at successive points in training with the corresponding human data. We compared human and simulated responses after paths of length 4 in 25 separate analyses, each using one of the five sets of simulated responses as predictor variable and one of the five sets of experimental responses as dependent variable. The obtained correlation coefficients were again corrected for attenuation. The results are illustrated in Figure 9. Each point in the figure represents the corrected r^2 of a specific analysis. One would expect the model's early performance to be a better predictor of the subjects' early behavior and vice versa for later points in training.

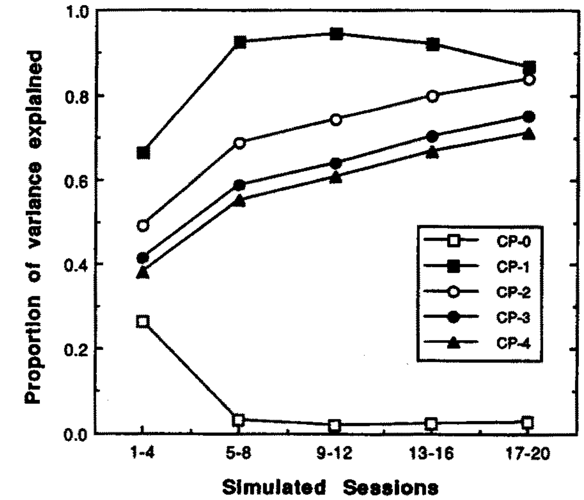

Figure 8. Correspondence between the simple recurrent network's responses and conditional probabilities (CP) after paths of length 0–4 during successive blocks of four simulated sessions.

It is obvious that the model is not very good at capturing subjects' behavior: The overall fit is relatively low (note that the vertical axis only goes up to .5) and reflects only weakly the expected progressions. It appears that too much of the variance in the model's performance is accounted for by sensitivity to the temporal context.

However, exploratory examination of the data revealed that factors other than the conditional probability of appearance of a stimulus exert an influence on performance in our task. We identified three such factors and incorporated them in a new version of the simulation model.

The Augmented SRN model

First of all, it appears that a response that is actually executed remains primed for a number of subsequent trials (Bertelson, 1961; Hyman, 1953; Remington, 1969). In the last sessions of our data, we found that if a response follows

[2] This transformation amounts to dividing the activation of the unit corresponding to the response by the sum of the activations of all units in the output pool. Because the strength of a particular response is determined by its relative, rather than absolute, activation, the transformation implements a simple form of response competition.

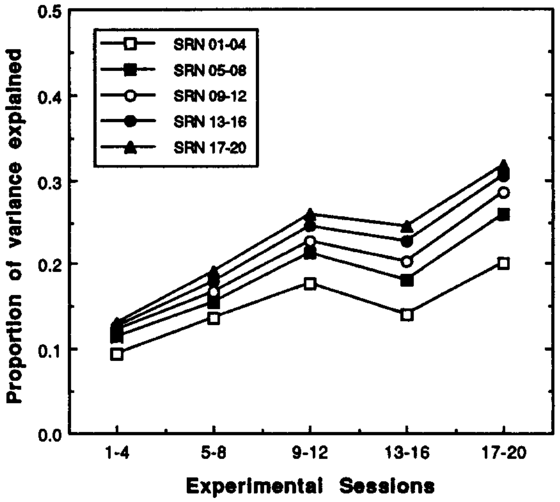

Figure 9. Correspondence between the simple recurrent network's (SRN's) responses and the human data during successive blocks of four sessions of training (Experiment 1).

itself immediately, there is about 60 to 90 ms of facilitation, depending on other factors. If it follows after a single intervening response (as in VT–V in Experiment 1, for example), there is about 25 ms of facilitation if the letter is grammatical at the second occurrence and 45 ms if it is ungrammatical.

The second factor may be related: Responses that are grammatical at Trial t but do not actually occur remain primed at Trial $t + 1$. The effect is somewhat weaker, averaging about 30 ms.

These two factors may be summarized by assuming (a) that activations at Time t decay gradually over subsequent trials and (b) that responses that are actually executed become fully activated, whereas those that are not executed are only partially activated.

The third factor is a priming, not of a particular response, but of a particular sequential pairing of responses. This can best be illustrated by a contrasting example, in which the response to the second X is compared in QXQ–X and VXQ–X. Both transitions are grammatical; yet the response to the second X tends to be about 10 ms faster in cases similar to QXQ–X, in which the X follows the same predecessor twice in a row, than it is in cases similar to VXQ–X, in which the first X follows one letter and the second follows a different letter.

This third factor can perhaps be accounted for in several ways. We have explored the possibility that it results from a rapidly decaying component to the increment to the connection weights mediating the associative activation of a letter by its predecessor. Such "fast" weights have been proposed by a number of investigators (Hinton & Plaut, 1987; McClelland & Rumelhart, 1985). The idea is that when X follows Q, the connection weights underlying the prediction that X will follow Q receive an increment that has a short-term component in addition to the standard long-term component. This short-term increment decays rapidly, but is still present in sufficient force to influence the response to a subsequent X that follows an immediately subsequent Q.

In the light of these analyses, one possibility for the relative failure of the original model to account for the data is that the SRN model is partially correct, but that human responses are also affected by rapidly decaying activations and adjustments to connection weights from preceding trials. To test this idea, we incorporated both kinds of mechanisms into a second version of the model. This new simulation model was exactly the same as before, except for two changes.

First, it was assumed that preactivation of a particular response was based not only on activation coming from the network, but also on a decaying trace of the previous activation:

$$\text{ravact[i]}(t) = \text{act[i]}(t) + (1 - \text{act[i]}(t))*k*\text{ravact[i]}(t - 1)$$

where act(t) is the activation of the unit based on the network at Time t, and *ravact*(t), that is, *r*unning *av*erage *act*ivation at time t, is a nonlinear running average that remains bounded between 0 and 1. After a particular response had been executed, the corresponding *ravact* was set to 1.0. The other *ravacts* were left at their current values. The constant k was set to 0.5, so that the half-life of a response activation is one time step.

The second change consisted of assuming that changes imposed on the connection weights by the back-propagation learning procedure have two components. The first component is a small (*slow* ϵ = 0.15) but effectively permanent change (i.e., a decay rate slow enough to ignore for present purposes), and the other component is a slightly larger (*fast* ϵ = 0.2) change, but which has a half-life of only a single time step. (The particular values of ϵ were chosen by trial and error, but without exhaustive search.)

With these changes in place, we observed that, of course, the proportion of the variance in the model accounted for by predictions based on the temporal context is dramatically reduced, as illustrated in Figure 10 (compare with Figure 8). More interesting, the pattern of change in these measures as well as the overall fit is now quite similar to that observed in the human data (Figure 4).

Indeed, there is a similar progressive increase in the correspondence with the higher order CPs, with the curve for the first-order CPs leveling off relatively early with respect to those corresponding to CPs based on paths of length 2, 3, and 4.

A more direct indication of the good fit provided by the current version of the model is given by the fact that it now correlates very well with the performance of the subjects (Figure 11; compare with the same analysis illustrated in Figure 9, but note the scale difference). Late in training, the model explains about 81% of the variance of the corresponding human data. Close inspection of the figure also reveals that, as expected, the SRN's early distribution of responses is a slightly better predictor of the corresponding early human data. This correspondence gets inverted later on, thereby suggesting that the model now captures key aspects of acquisition as well. Indeed, at almost every point, the best prediction of the human data is the simulation of the corresponding point in training.

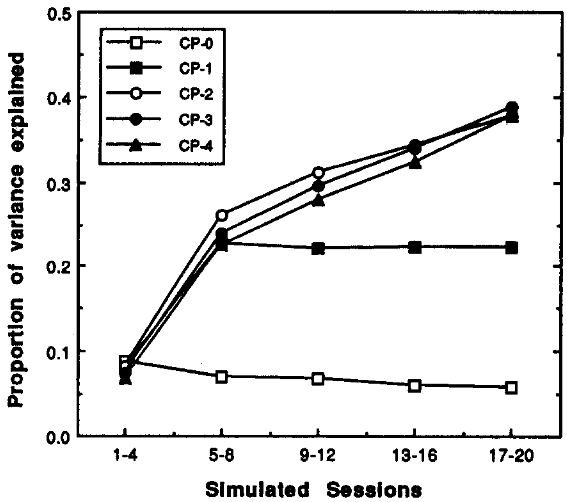

Figure 10. Correspondence between the augmented simple recurrent network's responses and conditional probabilities (CPs) after paths of length 0–4 during successive blocks of four simulated sessions.

Two aspects of these data need some discussion. First, the curves corresponding to each set of CPs are close to each other because the majority of the model's responses retain their relative distribution as training progresses. This is again a consequence of the fact that only a few elements of the sequence require more than two elements of temporal context to be perfectly disambiguated.

Second, the model's responses correlate very well with the data, but not perfectly. This raises the question as to whether there are aspects of the data that cannot be accounted for by the postulated mechanisms. There are three reasons why this need not be the case. First, the correction for attenuation assumes homogeneity, but because of different numbers of trials in different cells there is more variability in some cells than in others (typically, the cells corresponding to grammatical successors of paths of length 4 are much more stable than those corresponding to ungrammatical successors). Second, the set of parameters we used is probably not optimal. Although we examined several combinations of parameter values, the possibility of better fits with better parameters cannot be excluded. Finally, in fitting the model to the data, we have assumed that the relation between the models' responses and reaction times was linear, whereas in fact it might be somewhat curvilinear. These three facts would all tend to reduce the r^2 well below 1.0 even if the model is in fact a complete characterization of the underlying processing mechanisms.

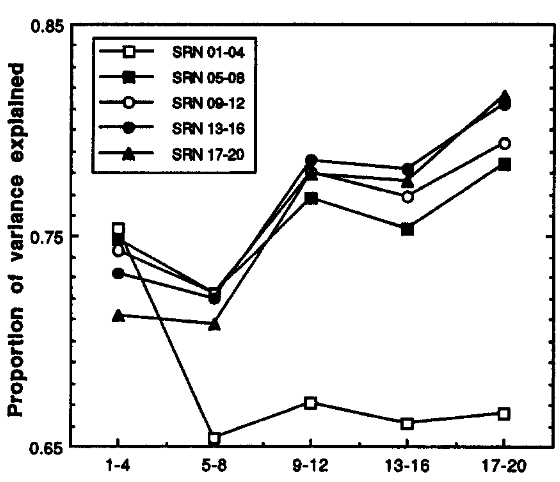

Figure 11. Correspondence between the augmented simulated recurrent network's (SRN's) responses and the human data during successive blocks of four sessions of training (Experiment 1).

The close correspondence between the model and the subjects' behavior during learning is also supported by an analysis of the model's responses to paths of length 3 and 4 (Experiment 1). Using exactly the same selection of paths as for the subjects in each case, we found that a small but systematic difference between the model's responses to predictable and unpredictable successors to paths of length 3 emerged in Sessions 9–12 and kept increasing over Sessions 13–16 and 17–20. The difference was .056 (i.e., a 5.6% difference in the mean response strength) when averaged over the last four sessions of training. By contrast, this difference score for paths of length 4 was only .003 at the same point in training, thereby clearly indicating that the model was not sensitive to the fourth-order temporal context.

Finally, to further illustrate the correspondence between the model and the experimental data, we wanted to compare human and simulated responses on an ensemble of specific successors of specific paths, but the sheer number of data points renders an exhaustive analysis virtually intractable. There are 420 data points involved in each of the analyses discussed above. However, one analysis that is more parsimonious, but that preserves much of the variability of the data, consists of comparing human and simulated responses for each letter at each node of the grammar. Because the grammar used in Experiment 1 counts seven nodes (0–6), and because each letter can occur at each node because of the noise, this analysis yields 42 data points, a comparatively small number. Naturally, some letters are more likely to occur at some nodes than at others, and therefore, one expects the distribution of average RTs over the six possible letters to be different for different nodes. For instance, the letters *V* and *P* should elicit relatively faster responses at Node #0, where both letters are grammatical, than at Node #2, where neither of them is. Figure 12 represents the results of this analysis. Each individual graph shows the response to each of the six letters at a particular node, averaged over the last four sessions of training, for both human and simulated data. Because there is an inverse relationship between activations and RTs, the model's responses have been subtracted from 1. All responses were then transformed into standard scores to allow for direct comparisons between the model and the experimental data, and the figures therefore represent deviations from the general mean.

Visual examination reveals that the correspondence between the model and the data is very good. This was confirmed by the high degree of association between the two data sets: The corrected r^2 was .88. Commenting in detail on each

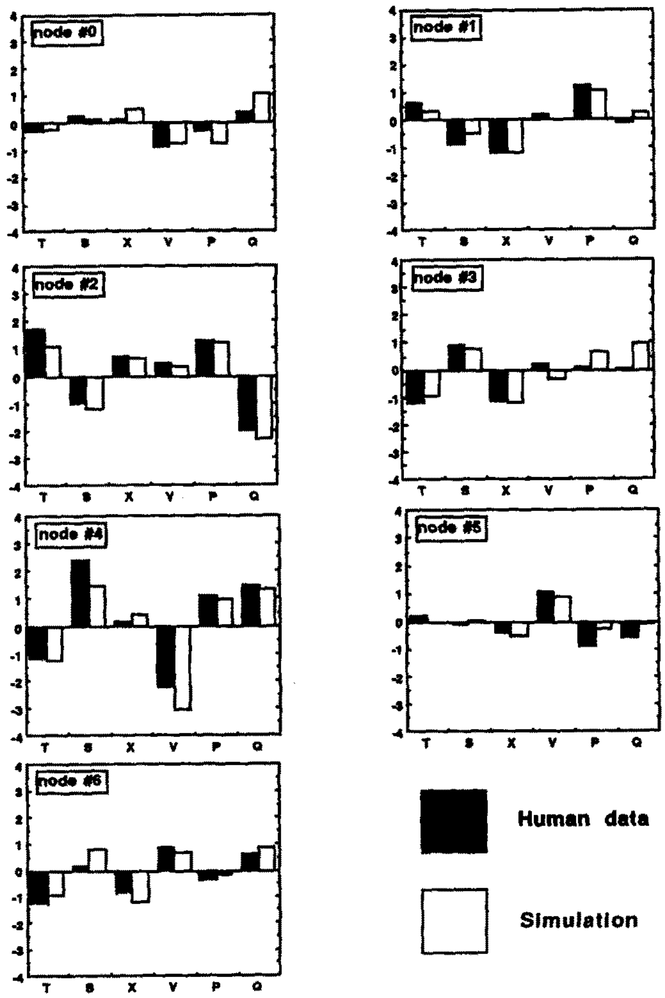

Figure 12. Human and simulated responses to each of the six letters, plotted separately for each node (#0 to #6) of the grammar (Experiment 1). (All responses have been transformed into standard scores with respect to the mean of the entire distribution.)

of the figures seems unnecessary, but some aspects of the data are worth remarking on. For instance, one can see that the fastest response overall is elicited by a V at Node #4. This is not surprising, because the T–V association is both frequent (note that it also occurs at Node #0) and consistent (i.e., the letter *T* is a relatively reliable cue to the occurrence of a subsequent *V*). Furthermore, V also benefits from its involvement in a TVT–V alternation in a number of cases. On Figure 12, one can also see that T elicits a relatively fast response, even though it is ungrammatical at Node #4. This is a direct consequence of the fact that a T at Node #4 follows itself immediately. It is therefore primed despite its ungrammaticality. The augmented SRN model captures both of these effects quite adequately, if not perfectly.

The impact of the short-term priming effects is also apparent in the model's overall responses. For instance, the initial difference between grammatical and ungrammatical trials observed in the first session of both experiments is also present in the simulation data. In both cases, this difference results from the fact that responses to first-order repetitions (which are necessarily ungrammatical) were eliminated from the ungrammatical trials, whereas second-order repetitions and trials involved in alternations were not eliminated from the grammatical trials. Each of these two factors contribute to widen the difference between responses to grammatical and ungrammatical trials, even though learning of the sequential structure is only minimal at that point. The fact that the SRN model also exhibits this initial difference is a further indication of its aptness at accounting for the data.

Attention and Sequence Structure

Can the SRN model also yield insights into other aspects of sequence learning? A. Cohen et al. (1990) reported that sequence structure interacts with attentional requirements. Subjects placed in a choice reaction situation were able to learn sequential material under attentional distraction, but only when it involved simple sequences in which each element has a unique successor (such as in 12345 . . .). More complex sequences involving ambiguous elements (i.e., elements that could be followed by several different successors, as in 123132 . . .) could only be learned when no secondary task was performed concurrently. A third type of sequence—hybrid sequences—in which some elements were uniquely associated to their successor and some other elements were ambiguous (such as in 143132 . . .) elicited intermediate results. A. Cohen et al. (1990) hypothesized that the differential effects of the secondary task on the different types of sequences might be due to the existence of two different learning mechanisms: one that establishes direct pairwise associations between an element of the sequence and its successor, and another that creates hierarchical representations of entire subsequences of events. The first mechanism would require less attentional resources than the second and would thus not suffer as much from the presence of a secondary task. A. Cohen et al. further point out that there is no empirical basis for distinguishing between this hypothesis and a second one, namely, that all types of sequences are processed hierarchically, but that ambiguous sequences require a more complex "parsing" than unique sequences. Distraction would then have differential effects on these two kinds of hierarchical coding.

We propose a third possibility: that sequence learning may be based solely on associative learning processes of the kind found in the SRN.[3] Through this learning mechanism, associations are established between prediction-relevant features of previous elements of the sequence and the next element. If two subsequences have the same successors, the model will tend to develop identical internal representations in each case.

[3] In work done independently of our simulations, J. K. Kruschke (personal communication, June 5, 1990) explored the possibility of simulating the effects of attention on sequence learning in SRNs. In one of his simulations, the learning rate of the connections from the context units to the hidden units was set to a lower value than for the other connections of the network.

If two otherwise identical subsequences are followed by different successors as a function of their predecessors, however, the network will tend to develop slightly different internal representations for each subsequence. This ability of the network to simultaneously represent similarities and differences led us to refer to the SRN model as an instantiation of a graded state machine (McClelland, Cleeremans, & Servan-Schreiber, 1990). This notion emphasizes the fact that, although there is no explicit representation of the hierarchical nature of the material, the model nevertheless develops internal representations that are shaded by previous elements of the sequence.

The key point in the context of this discussion is that the representations of sequence elements that are uniquely associated with their successors are not different in kind from those of elements that can be followed by different successors as a function of their own predecessors. How, then, might the model account for the interaction between attention and sequence structure reported by A. Cohen et al. (1990)? One possibility is that the effect of the presence of a secondary task is to hamper processing of the sequence elements. A simple way to implement this notion in our model consists of adding normally distributed random noise to the input of specific units of the network (Cohen & Servan-Schreiber, 1989, explored a similar idea by manipulating gain to model processing deficits in schizophrenia). The random variability in the net input of units in the network tends to disrupt processing, but in a graceful way (i.e., performance does not break down entirely). The intensity of the noise is controlled by a scale parameter, σ. We explored how well changes in this parameter as well as changes in the localization of the noise captured the results of Experiment 4 of A. Cohen et al. (1990).

A Simulation of Attentional Effects in Sequence Learning

In this experiment, subjects were exposed to 14 blocks of either 100 trials for the unique sequence (12345 . . .) condition or 120 trials for the ambiguous sequence (123132 . . .) and hybrid sequence (143132 . . .) conditions. Half of the subjects receiving each sequence performed the task under attentional distraction (in the form of a tone-counting task); the other half only performed the sequence learning task. In each of these six conditions, subjects first received two blocks of random material (Blocks 1–2), followed by eight blocks of structured material (Blocks 3–10), then another two blocks of random material (Blocks 11–12), and a final set of two blocks of structured material (Blocks 13–14). The interesting comparisons are between performance on the last two random blocks (Blocks 11–12), on the one hand, and on the four last structured blocks (Blocks 9–10 and 13–14), on the other hand. Any positive difference between the average RTs on these two groups of blocks would indicate interference when the switch to random material occurred, thereby suggesting that subjects have become sensitive to the sequential structure of the material.

We have represented the standard scores of the six relevant RT differences in the left panel of Figure 13. When the sequence learning task is performed alone ("single" condition), unique and hybrid sequences are better learned than ambiguous sequences, as indicated by the larger difference between random and structured material elicited by unique and hybrid sequences. The same pattern is observed when the sequence learning task is performed concurrently with the tone-counting task ("dual" condition), but overall performance is much lower. In the actual data, the difference between random and structured material for the ambiguous sequence is very close to zero. In other words, the ambiguous sequence is not learned at all under dual-task conditions. The crucial point that this analysis reveals, however, is that learning of the unique and hybrid sequences is also hampered by the presence of the secondary task.

To capture this pattern of results, an SRN with 15 hidden units was trained in exactly the same conditions as subjects in the study by A. Cohen et al. (1990). We recorded the response of the network to each stimulus and separately averaged these responses over the last random and structured blocks, as described above. These mean responses were then substrated from one and transformed into standard scores to allow for direct comparisons with the data.

We explored three different ways of modeling the secondary task by means of noise. One consists of adding noise to the connections from the context units to the hidden units only. We found that this resulted in specific interference with acquisition of the ambiguous sequence. Basically, the network learns to ignore the noisy information coming from the context units and minimizes the error using the main processing pathway only. However, this is not what is observed in the data: The presence of the secondary task also hampers learning of the unique and hybrid sequences. Therefore, we focused on two other ways of allowing noise to interfere with processing: adding noise to the net input of each unit of the network or adding noise to the net input of each hidden unit only. In both cases, activation propagating from the context units and from the input units to the rest of the network was affected equally.

In a first simulation, the secondary task was modeled by adding normally distributed random noise ($\sigma = 0.7$) to the net input of each unit in the network. The learning rates were set to 0.35 (*slow* ϵ) and to 0.45 (*fast* ϵ). The values of the other parameters were identical to those used in our previous simulations. The results are illustrated in the middle panel of Figure 13. The response pattern produced by the network is quite similar to the human data. In particular, the noise (a) affected learning of all three types of sequences and (b) virtually eliminated learning of the ambiguous sequence. Indeed, the difference score for the ambiguous sequence was 0.019 in the dual condition, only 1.9%. Thus, at this level of noise, learning of the ambiguous sequence is almost entirely blocked, as for subjects in the A. Cohen et al. (1990) study. By contrast, learning of the unique and hybrid sequences is relatively preserved, although the hybrid sequence was not learned as well by the model as by the subjects.

The right panel of Figure 13 illustrates the results of a similar analysis conducted on a simulation using higher learning rates (*slow* $\epsilon = 0.7$, *fast* $\epsilon = 0.8$) and in which noise ($\sigma = 1.9$) was only allowed to affect the net input to each hidden

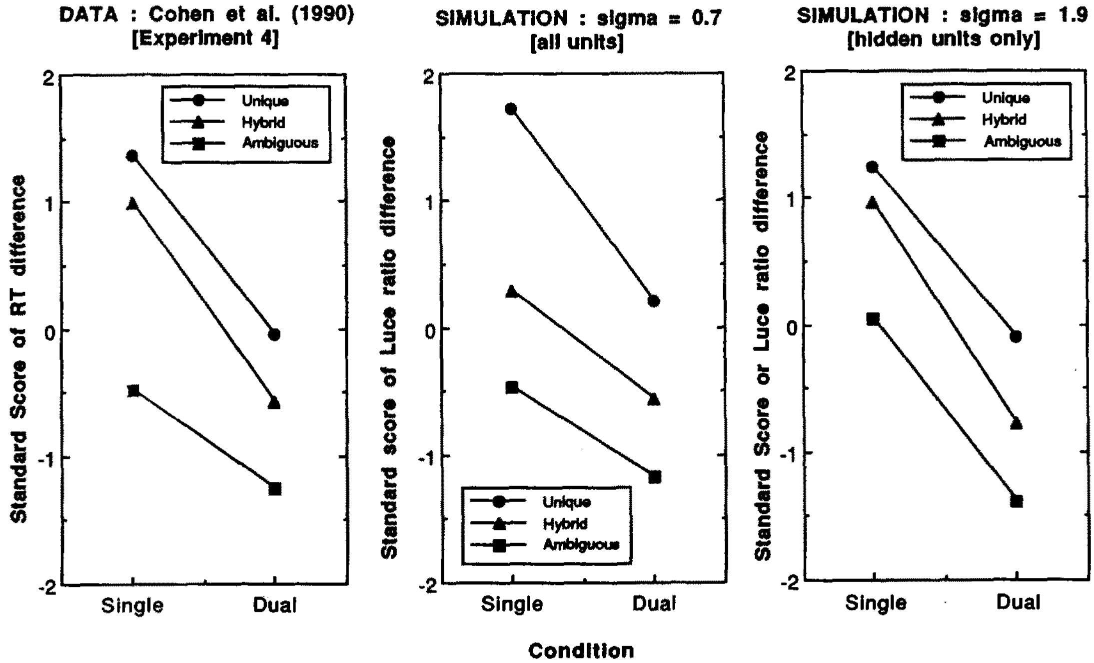

Figure 13. Standard scores of human and simulated mean difference scores between responses on random and structured material, for unique, hybrid, and ambiguous sequences, and under single- or dual-task conditions.

unit of the network. The figure shows that with these very different parameters, the model still captures the basic pattern of results observed in the data. The difference score for the ambiguous sequence in the dual condition was 0.023, again very close to zero. In contrast with the previous simulation, however, the hybrid sequence now appears to be learned as well as by human subjects. The ambiguous sequence, on the other hand, seems to be learned somewhat too well with this particular set of parameters.

The important result is that both simulations produced an interference pattern qualitatively similar to the empirical data. We found that quite a wide range of parameter values would produce this effect. For instance, the basic pattern is preserved if the learning rates and the noise parameter are varied proportionally or, as our two simulations illustrate, if the noise is allowed to interfere with all the units in the network or with only the hidden units. This just shows that fitting simulated responses to empirical data ought to be done at a fairly detailed level of analysis. A precise, quantitative match with the data seems inappropriate at this relatively coarse level of detail. Indeed, there is no indication that exactly the same pattern of results would be obtained in a replication, and overfitting is always a danger in simulation work. The central point is that we were able to reproduce this pattern of results by manipulating a single parameter in a system that makes no processing or representational distinction between unique, hybrid, and ambiguous sequences.

To summarize, these results have two important implications. First, it appears that the secondary task exerts similar detrimental effects on both types of sequences. Learning of ambiguous sequences is almost entirely blocked when performed concurrently with the tone-counting task. Unique and hybrid sequences can be learned under attentional distraction, but to a lesser extent than under single-task conditions. Both of these effects can be simulated by varying the level of noise in the SRN model.

Second, our simulations suggest that unique and ambiguous sequences are represented and processed in the same way. Therefore, a distinction between associative and hierarchical sequence representations does not appear to be necessary to explain the interaction between sequence structure and attention observed by A. Cohen et al. (1990).

General Discussion

In Experiment 1, subjects were exposed to a six-choice serial reaction time task for 60,000 trials. The sequential structure of the material was manipulated by generating successive stimuli on the basis of a small finite-state grammar. On some of the trials, random stimuli were substituted to those prescribed by the grammar. The results clearly support the idea that subjects become increasingly sensitive to the sequential structure of the material. Indeed, the smooth differentiation between grammatical and ungrammatical trials can only be explained by assuming that the temporal context set by previous elements of the sequence facilitates or interferes with the processing of the current event. Subjects progressively come to encode more and more temporal context by attempting to optimize their performance on the next trial. Experiment 2 showed that subjects were relatively unable to maintain information about long-distance contingencies that span irrelevant material. Taken together, these results suggest that, in this type of task, subjects gradually acquire a complex body of procedural knowledge about the sequential structure of the material. Several issues may be raised regarding the form of this knowledge and the mechanisms that underlie its acquisition.

Sensitivity to the Temporal Context and Sequence Representation

Subjects are clearly sensitive to more than just the immediate predecessor of the current stimulus; indeed, there is evidence of sensitivity to differential predictions based on two and even three elements of context. However, sensitivity to the temporal context is also clearly limited: Even after 60,000 trials of practice, there is no evidence that subjects discriminate between the different possible successors entailed by elements of the sequence four steps away from the current trial. The question of how much temporal context subjects may be able to encode has not been thoroughly explored in the literature, and it is therefore difficult to compare our results with the existing evidence. Remington (1969) demonstrated that subjects' responses in a simple two-choice reaction task were affected by elements as removed as five steps, but the effects were very small and did not depend on the sequential structure of the material. Rather, they were essentially the result of repetition priming. Early studies by Millward and Reber (1968, 1972), however, documented sensitivity to as much as seven elements of temporal context in a two-choice probability learning paradigm that used structural material. In the Millward and Reber (1972) study, the sequences were constructed so that the event occurring on Trial t was contingent on an earlier event occurring at Trial $t - L$. The lag L was progressively increased from 1 to 7 over successive experimental sessions. The results indicated that subjects were slightly more likely to produce the contingent response on the trial corresponding to the lag than on any other trial, thereby suggesting that they encoded the contingency. A number of factors, however, make this result hard to generalize to our situation. First, subjects were asked to predict the next element of a sequence, rather than simply react to it. It is obvious that this requirement will promote explicit encoding of the sequential structure of the material much more than in our situation. Second, the task only involved two choices, which is much fewer than the six choices used here. There is little doubt that detecting contingencies is facilitated when the number of stimuli is reduced. Third, the training schedule (in which the lag between contingent events was progressively increased over successive practice sessions) used in this study is also likely to have facilitated encoding of the long-distance contingencies. Finally, the differences in response probabilities observed by Millward and Reber (1972) were relatively small for the longer lags (for instance, they reported a .52 probability of predicting the contingent event at Lag 7 vs. .47 for the noncontingent event).

More recently, Lewicki et al. (1987), and also Stadler (1989), reported that subjects seemed to be sensitive to six elements of temporal context in a search task in which the location of the target on the seventh trial was determined by the locations of the target on the six previous trials. This result may appear to contrast with ours, but close inspection of the structure of the sequences used by Lewicki et al. (1987) revealed that 50% of the uncertainty associated with the location of the target on the seventh trial may be removed by encoding just three elements of temporal context. This could undoubtedly account for the facilitation observed by Lewicki et al. and is totally consistent with the results obtained here.

In summary, none of the above studies provided firm evidence that subjects become sensitive to more than three or four elements of temporal context in situations that do not involve explicit prediction of successive events. It is interesting to speculate on the causes of these limitations. Long-distance contingencies are necessarily less frequent than shorter ones. However, this should not prevent them per se from becoming eventually encoded should the regularity-detection mechanism be given enough time and resources. A more sensible interpretation is that memory for sequential material is limited and that the traces of individual sequence elements decay with time. More recent traces would replace older ones as they are processed. This notion is at the core of many early models of sequence processing (e.g., Laming, 1969). In the SRN model, however, sequence elements are not represented individually, and memory for context does not spontaneously decay with time. The model nevertheless has clear limitations in its ability to encode long-distance contingencies. The reason for these limitations is that the model develops representations that are strongly determined by the constraints imposed by the prediction task. That is, the current element is represented together with a representation of the prediction-relevant features of previous sequence elements. As learning progresses, representations of subsequences followed by identical successors tend to become more and more similar. For instance, we have shown that an SRN with three hidden units develops internal representations that correspond exactly to the nodes of the finite-state grammar from which the stimulus sequence was generated (Cleeremans et al., 1989). This is a direct consequence of the fact that all the subsequences that entail the same successors (i.e., that lead to the same node) tend to be represented together. As a result, it also becomes increasingly difficult for the network to produce different responses to otherwise identical subsequences preceded by disambiguating elements. In a sense, more distant elements are subject to a loss of resolution, the magnitude of which depends exponentially on the number of hidden units available for processing (Servan-Schreiber et al., 1988). Encoding long-distance contingencies is greatly facilitated if each element of the sequence is relevant—even only in a probabilistic sense—for predicting the next one. Whether subjects also exhibit this pattern of behavior is a matter for further research.

Awareness of the Sequential Structure

It is often claimed that learning can proceed without explicit awareness (e.g., Reber, 1989; Willingham et al., 1989). However, in the case of sequence learning, as in most other implicit learning situations, it appears that subjects become aware of at least some aspects of the structure inherent in the stimulus material. Our data suggest that subjects do become aware of the alternations that occur in the grammar (e.g., SQSQ and VTVT in Experiment 1), but have little reportable knowledge of any other contingencies. The loops also produced marked effects on performance. Indeed, as Figure 12 illustrates, the greatest amount of facilitation occurs at Nodes #2 and #4, and for the letters involved in the loops (Q at Node #2 and V at Node #4). However, this does not necessarily entail that explicit knoweldge about these alternations played a significant role in learning the sequential structure of the material. Indeed, a great part of the facilitation observed for these letters results from the fact that they are subject to associative priming effects because of their involvement in alternations. Furthermore, our data contain many instances of cases in which performance facilitation resulting from sensitivity to the sequential structure was not accompanied by corresponding explicit knowledge. For instance, the results of the analysis on differential sensitivity to the successors of selected paths of length 3 (Experiment 1) clearly demonstrate that subjects are sensitive to contingencies they are unable to elaborate in their explicit reports. In other words, we think that awareness of some aspects of the sequential structure of the material emerges as a side effect of processing and plays no significant role in learning itself. As it stands, the SRN model does not address this question directly. Indeed, it incorporates no mechanism for verbalizing knowledge or for detecting regularities in a reportable way. However, the model implements a set of principles that are relevant to the distinction between implicit and explicit processing. For instance, even though the internal representations of the model are structured and reflect information about the sequence, the relevant knowledge is embedded in the connection weights. As such, this knowledge is relatively inaccessible to observation. By contrast, the internal representations of the model may be made available to some other component of the system. This other component of the system may then be able to detect and report on the covariations present in these internal representations, even though it would play but a peripheral role in learning or in processing. Even so, the internal representations of the model may be hard to describe because of their graded and continuously varying nature.

Other aspects of the data support the view that explicit knowledge of the sequence played but a minimal role in this task. For instance, even though the results of the generation task, which followed training in Experiment 2, clearly indicate that subjects were able to use their knowledge of the sequence to predict the location of some grammatical events, overall prediction performance was very poor, particularly when compared with previous results. A. Cohen et al. (1990), for instance, showed that subjects were able to achieve near perfect prediction performance in as little as 100 trials. In stark contrast, our subjects were only able to correctly predict about 25% of the grammatical events after 450 trials of the generation task and 60,000 trials of training. This difference further highlights the complexity of our experimental situation and suggests that the presence of the noise and the number of different possible grammatical subsequences make it very hard to process the material explicitly. This was corroborated by subjects' comments that they had sometimes tried to predict successive events, but had abandoned this strategy because they felt it was detrimental to their performance.

In short, these observations lead us to believe that subjects had very little explicit knowledge of the sequential structure in this situation and that explicit strategies played but a negligible role during learning. One may wonder, however, about the role of explicit recoding strategies in task settings as simple as those used by Lewicki et al. (1988) or A. Cohen et al. (1990). In both these situations, subjects were exposed to extremely simple repeating sequences of no more than six elements in length. But the work of Willingham et al. (1989) has demonstrated that a sizeable proportion of subjects placed in a choice reaction situation involving sequences of 10 elements do become aware of the full sequence. These subjects were also faster in the sequence learning task and more accurate in predicting successive sequence elements in a follow-up generation task. By the same token, a number of subjects also failed to show any declarative knowledge of the task despite good performance during the task. These results highlight the fact that the relationship between implicit and explicit learning is complex and subject to individual differences. Claims that acquisition is entirely implicit in simple sequence learning situations must be taken with caution.

To summarize, although it is likely that some subjects used explicit recoding strategies during learning, the complexity of the material we used—as well as the lack of improvement in the generation task—make it unlikely that they did so in any systemic way. Further experimental work is needed to assess in greater detail the impact of explicit strategies on sequence learning, using a range of material of differing complexity, before simulation models that incorporate these effects can be elaborated.

Learning Mechanisms and Attention

The augmented SRN model provides a detailed, mechanistic, and fairly good account of the data. Although the correspondence is not perfect, the model nevertheless captures much of the variability of human responses.

The model's core learning mechanism implements the notion that sensitivity to the temporal context emerges as the result of optimizing preparation for the next event on the basis of the constraints set by relevant (i.e., predictive) features of the previous sequence. However, this core mechanism alone is not sufficient to account for all aspects of performance. Indeed, as discussed above, our data indicate that in addition to the long-term and progressive facilitation obtained by encoding the sequential structure of the material, responses are also affected by a number of other short-term (repetitive and associative) priming effects. It is interesting to note that the relative contribution of these short-term priming effects tends to diminish with practice. For instance, an ungrammatical but repeated Q that follows an SQ− at Node #1 in Experiment 1 elicits a mean RT of 463 ms over the first 4 sessions of training. This is much faster than the 540 ms elicited by a grammatical X that follows SQ− at the same node. By contrast, this relationship becomes inverted in the last 4 sessions of the experiment: The Q now evokes a mean RT of 421 ms, whereas the response to an X is 412 ms. Thus, through practice, the sequential structure of the material comes to exert a growing influence on response times and tends to become stronger than the short-term priming effects. The augmented SRN model captures this interaction in a simple way: Early in training, the connection weights underlying sensitivity to the sequential structure are very small and can only exert a limited influence on the responses. At this point, responses are quite strongly affected by previous activations and adjustments to the fast weights from preceding trials. Late in training, however, the contribution of these effects in determining the activation of the output units ends up being dominated by the long-term connection weights, which, through training, have been allowed to develop considerably.[4]

With both these short-term and long-term learning mechanisms in place, we found that the augmented SRN model captured key aspects of sequence learning and processing in our task. Furthermore, the model also captured the effects of attention on sequence learning reported by A. Cohen et al. (1990). Even though ambiguous sequences are not processed by separate mechanisms in the SRN model, they are nevertheless harder to learn than unique and hybrid sequences because they require more temporal context information to be integrated. So the basic difference between the three sequence types is produced naturally by the model. Furthermore, when processing is disturbed by means of noise, the model produces an interference pattern very similar to that of the human data. Presumably, a number of different mechanisms could produce this effect. For instance, Jennings and Keele (1990) explored the possibility that the absence of learning of the ambiguous sequence under attentional distraction was the result of impaired "parsing" of the material. They trained a sequential back-propagation network (Jordan, 1986) to predict successive elements of a sequence and measured how the prediction error varied with practice under different conditions and for different types of sequences. The results showed that learning of ambiguous sequences progressed much slower than for unique or hybrid sequences when the input information did not contain any cues as to the structure of the sequences. By contrast, learning of ambiguous sequences progressed at basically the same rate as for the other two types of sequences when the input to the network did contain information about the structure of the sequence, such as the marking of sequence boundaries or an explicit representation of its subparts. If one assumes that attention is required for this explicit parsing of the sequence to take place and that the effects of the secondary task is to prevent such mechanisms from operating, then indeed learning of the ambiguous sequence will be hampered in the dual-task condition. However, the data seem to indicate that learning of the unique and hybrid sequences is also hampered by the presence of the secondary task. One would therefore need to know more about the effects of parsing on learning of the unique and hybrid sequences. Presumably, parsing would also facilitate processing of these kinds of sequences, although to a lesser extent than for ambiguous sequences.

In the case of the SRN model, we found that specifically interfering with processing of the ambiguous sequence by adding noise to the connections from the context units to the hidden units would not produce the observed data. On the contrary, our simulations indicate that the interference produced by the secondary task seems to be best accounted for when noise is allowed to equally affect processing of information coming from the context units and information coming from the input units. Therefore, it appears that there is no a priori need to introduce a theoretical distinction between processing and representation of sequences that have a hierarchical structure and sequences that do not. Naturally, we do not mean to suggest that sequence learning never involves the use of explicit recoding strategies of the kind suggested by A. Cohen et al. (1990) and by Jennings and Keele (1990). As pointed out earlier, it is very likely indeed that many sequence-learning situations do in fact involve both implicit and explicit learning and that recording strategies play a significant role in performance. Further research is needed to address this issue more thoroughly.

Conclusion

Subjects placed in a choice reaction time situation acquire a complex body of procedural knowledge about the sequential structure of the material and gradually come to respond on the basis of the constraints set by the last three elements of the temporal context. It appears that the mechanisms underlying this progressive sensitivity operate in conjunction with short-term and short-lived priming effects. Encoding of the temporal structure seems to be primarily driven by anticipation of the next element of the sequence. A PDP model that incorporates both of these mechanisms in its architecture was described and found to be useful in accounting for key aspects of acquisition and processing. This class of model therefore appears to offer a viable framework for modeling unintentional learning of sequential material.

[4] As Soetens, Boer, and Hueting (1985) have demonstrated, however, short-term priming effects also tend to become weaker through practice even in situations that only involve random material. At this point, the SRN model is simply unable to capture this effect. Doing so would require the use of a training procedure that allows the time course of activation to be assessed (such as cascaded back-propagation; see J. D. Cohen, Dunbar, & McClelland, 1990) and is a matter for further research.

References

Berry, D. C., & Broadbent, D. E. (1984). On the relationship between task performance and associated verbalizable knowledge. *Quarterly Journal of Experimental Psychology, 36A,* 209–231.

Bertelson, P. (1961). Sequential redundancy and speed in a serial two-choice responding task. *Quarterly Journal of Experimental Psychology, 13,* 90–102.

Carmines, E. G., & Zeller, R. A. (1987). *Reliability and validity assessment* (Sage University Paper Series No. 07-017). Newbury Park, CA: Sage.

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

Cohen, A., Ivry, R. I., & Keele, S. W. (1990). Attention and structure in sequence learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16,* 17–30.

Cohen, J. D., Dunbar, K., & McClelland, J. L. (1990). On the control of automatic processes: A parallel distributed account of the Stroop effect. *Psychological Review, 97,* 332–361.

Cohen, J. D., & Servan-Schreiber, D. (1989). *A parallel distributed processing approach to behavior and biology in schizophrenia* (Tech. Rep. No. AIP-100). Pittsburgh, PA: Carnegie Mellon University, Department of Psychology.

Dulany, D. E., Carlson, R. C., & Dewey, G. I. (1984). A case of syntactical learning and judgment: How conscious and how abstract? *Journal of Experimental Psychology: General, 113,* 541–555.

Dulany, D. E., Carlson, R. C., & Dewey, G. I. (1985). On consciousness in syntactical learning and judgment: A reply to Reber, Allen, and Regan. *Journal of Experimental Psychology: General, 114,* 25–32.

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

Elman, J. L. (in press). Representation and structure in connectionist models. In G. Altmann (Ed.), *Computational and psycholinguistic approaches to speech processing.* San Diego, CA: Academic Press.

Estes, W. K. (1976). The cognitive side of probability learning. *Psychological Review, 83,* 37–64.

Falmagne, J. C. (1965). Stochastic models for choice reaction time with application to experimental results. *Journal of Mathematical Psychology, 2,* 77–124.

Hayes, N. A., & Broadbent, D. E. (1988). Two modes of learning for interactive tasks. *Cognition, 28,* 249–276.

Hebb, D. O. (1961). Distinctive features of learning in the higher animal. In A. Fressard, R. W. Gerard, J. Konorsky, & J. F. Delafresnaye (Eds.), *Brain mechanisms and learning* (pp. 37–51). Oxford, England: Blackwell Scientific.

Hinton, G. E., & Plaut, D. C. (1987). Using fast weights to deblur old memories. *Proceedings of the Ninth Annual Conference of the Cognitive Science Society* (pp. 177–186). Hillsdale, NJ: Erlbaum.

Hyman, R. (1953). Stimulus information as a determinant of reaction time. *Journal of Experimental Psychology, 45,* 188–196.

Jennings, P. J., & Keele, S. W. (1990). A computational model of attentional requirements in sequence learning. *Proceedings of the Twelfth Annual Conference of the Cognitive Science Society* (pp. 876–883). Hillsdale, NJ: Erlbaum.

Jordan, M. I. (1986). Attractor dynamics and parallelism in a connectionist sequential machine. *Proceedings of the Eighth Annual Conference of the Cognitive Science Society* (pp. 531–546). Hillsdale, NJ: Erlbaum.

Laming, D. R. J. (1969). Subjective probability in choice-reaction experiments. *Journal of Mathematical Psychology, 6,* 81–120.

Lewicki, P., Czyzewska, M., & Hoffman, H. (1987). Unconsicous acquisition of complex procedural knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 13,* 523–530.

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

Luce, R. D. (1963). Detection and recognition. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology* (Vol. 1, pp. 103–189). New York: Wiley.

Mathews, R. C., Buss, R. R., Stanley, W. B., Blanchard-Fields, F., Cho, J. R., & Druhan, B. (1989). Role of implicit and explicit processes in learning from examples: A synergistic effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15,* 1083–1100.

McClelland, J. L., Cleeremans, A., & Servan-Schreiber, D. (1990). Parallel distributed processing: Bridging the gap between human and machine intelligence. *Journal of the Japanese Society for Artificial Intelligence, 5,* 2–14.

McClelland, J. L., & Rumelhart, D. E. (1985). Distributed memory and the representation of general and specific information. *Journal of Experimental Psychology: General, 114,* 159–188.

Miller, G. A. (1958). Free recall of redundant strings of letters. *Journal of Experimental Psychology, 56,* 485–491.

Millward, R. B., & Reber, A. S. (1968). Event recall in probability learning. *Journal of Verbal Learning and Verbal Behavior, 7,* 980–989.

Millward, R. B., & Reber, A. S. (1972). Probability learning: Contingent-event schedules with lags. *American Journal of Psychology, 85,* 81–98.

Newell, A., & Simon, H. A. (1972). *Human problem solving.* Englewood Cliffs, NJ: Prentice-Hall.

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

Pew, R. W. (1974). Levels of analysis in motor control. *Brain Research, 71,* 393–400.

Reber, A. S. (1967). Implicit learning of artificial grammars. *Journal of Verbal Learning and Verbal Behavior, 6,* 855–863.

Reber, A. S. (1989). Implicit learning and tacit knowledge. *Journal of Experimental Psychology: General, 118,* 219–235.

Reber, A. S., Allen, R., & Regan, S. (1985). Syntactical learning and judgment, still unconscious and still abstract: Comment on Dulany, Carlson, and Dewey. *Journal of Experimental Psychology: General, 114,* 17–24.

Remington, R. J. (1969). Analysis of sequential effects in choice reaction times. *Journal of Experimental Psychology, 82,* 250–257.

Restle, F. (1970). Theory of serial pattern learning: Structural trees. *Psychological Review, 77,* 481–495.

Rumelhart, D. E., Hinton, G., & Williams, R. J. (1986). Learning internal representations by error propagation. In D. E. Rumelhart & J. L. McClelland (Eds.), *Parallel distributed processing: I. Foundations* (pp. 318–362). Cambridge, MA: MIT Press.

Schacter, D. L. (1987). Implicit memory: History and current status. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 13,* 501–518.

Servan-Schreiber, D., Cleeremans, A., & McClelland, J. L. (1988). *Encoding sequential structure in simple recurrent networks* (Tech. Rep. No. CMU-CS-88-183). Pittsburgh, PA: Carnegie Mellon University, Department of Computer Science.

Servan-Schreiber, E., & Anderson, J. R. (1990). Learning artificial grammars with competitive chunking. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16,* 592–608.

Soetens, E., Boer, L. C., & Hueting, J. E. (1985). Expectancy or automatic facilitation?: Separating sequential effects in two-choice reaction time. *Journal of Experimental Psychology: Human Perception and Performance, 11,* 598–616.

Stadler, M. A. (1989). On learning complex procedural knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15,* 1061–1069.

Willingham, D. B., Nissen, M. J., & Bullemer, P. (1989). On the development of procedural knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15,* 1047–1060.

Received June 18, 1990
Revision received November 29, 1990
Accepted December 11, 1990 ■