Statistical Learning in a Serial Reaction Time Task: Access to Separable Statistical Cues by Individual Learners

Ruskin H. Hunt and Richard N. Aslin
University of Rochester

The ability of adult learners to exploit the joint and conditional probabilities in a serial reaction time task containing both deterministic and probabilistic information was investigated. Learners used the statistical information embedded in a continuous input stream to improve their performance for certain transitions by simultaneously exploiting differences in the predictability of 2 or more underlying statistics. Analysis of individual learners revealed that although most acquired the underlying statistical structure veridically, others used an alternate strategy that was partially predictive of the sequences. The findings show that learners possess a robust learning device well suited to exploiting the relative predictability of more than 1 source of statistical information at the same time. This work expands on previous studies of statistical learning, as well as studies of artificial grammar learning and implicit sequence learning.

Nearly all behaviors, from the simplest conditioned reflexes to the most complex motor skills, involve making predictions. Predictions serve to anticipate the next event in a temporal sequence, thereby reducing uncertainty and enabling the effective preparation of a motor response. A classic example of using predictions to guide motor responses is Bryan and Harter’s (1897) study of telegraph operators, who showed both gradual and abrupt improvements in performance as they used increasingly complex predicative information about the structure of English. Predictions are based on extracting one or more statistics from the distribution of inputs and using that information to make a decision. For example, a squirrel foraging for food may return to a particular oak tree because over the past 60 days that location, among all others sampled, had the highest frequency of acorns. In contrast, human adults predict the next word in a sentence not on the basis of simple frequency (e.g., the is the most frequent word in English) but on the basis of the higher order regularities that constitute the grammar of their native language. Lashley (1951; see also Restle, 1970) noted that higher order patterns can be learned using a variety of algorithms.

The goal of the present series of experiments was to explore what types of statistics adult learners use during a sequential-order task whose complexity falls between simple conditioned reflexes and natural languages. This research forms part of a larger program of investigation that attempts to delineate the types of statistics that learners bring to bear on different types of learning tasks involving sequentially ordered information (Aslin, Saffran, & Newport, 1999; Newport & Aslin, 2000). One of the premises of this research program is that learners may be able to take advantage of a number of different statistics present in the learning material and that they may do so in parallel. The particular statistic a learner uses is likely to depend on several factors: (a) which statistics are available in the learning material, (b) which statistics the learner is capable of computing, (c) which statistics of those available in the input provide the most reliable information about the input, and (d) the ease with which those statistics can be computed (i.e., the relative perceptual and computational demands required to compute the different statistics in the learner’s repertoire). To begin to disentangle these issues, we designed a complex stimulus structure that allowed us to simultaneously present different predictive statistics to the learner but that also limited the number of statistics the learner could reasonably use to improve task performance. As a result, we could assess which of the available statistics learners were exploiting as their exposure to the task increased. Finally, the task we used allowed us to present stimuli as a continuous stream so that no explicit cues (e.g., segmentation or grouping cues) to the underlying structural relations would be available to the learner other than those under our control. At the same time, the task provided us with a dependent measure that allowed us to assess learners’ ability to exploit these statistics across time.

The task used in our experiments involves a multichoice, disjunctive, serial reaction time (SRT) paradigm (see Nissen & Bullem, 1987, for prototypical instantiations of this paradigm). In this paradigm, a visual cue is linked to a distinct, spatially-specific motor response. For example, each of five horizontally positioned locations on a computer screen may be linked with the corresponding positions of the five digits of the hand. The participant’s task is to press a button located under each digit as its corresponding visual cue is illuminated on the computer screen. If the sequence of cue lights is random, then participants show an overall improvement in reaction time (RT) over trials, but no differential improvement for any of the five response alternatives. However, if a sequential structure is embedded in the stream of cue lights, then RTs decline below this random baseline level. Thus, the dependent measure of learning in the SRT task is the pattern of RT differences that correlate with the underlying temporal-order statistics.

This research was supported by National Science Foundation Grant SBR-9873477. We gratefully acknowledge the assistance of Elizabeth Gramow for data collection and Carla Hudson, Daehee Lee, and Elissa Newport for helpful comments on an earlier version of this article.

Correspondence concerning this article should be addressed to Ruskin H. Hunt or Richard N. Aslin, Department of Brain and Cognitive Sciences, University of Rochester, Rochester, New York 14627-0268. Electronic mail may be sent to aslin@ivs.rochester.edu or ruskin@bcs.rochester.edu.
SRT experiments typically involve relatively few cue lights whose temporal order is fixed and repeated without interruption (e.g., Cohen, Ivry, & Keele, 1990). For example, three cue lights might appear in the following fixed order: 1–3–2–3–1–2. In this case, each of the cue lights occurs twice in the six-element sequence, and each is followed by one of two other cue lights. Participants quickly learn the exact order of the sequence, despite no overt segmentation of the six-element sequence, and their RTs not only decline but begin to anticipate the next cued response in the sequence. Thus, the sequential structure, with a small number of elements and a fixed sequence of moderate length, can be perfectly predicted by learning the precise ordering and length of the underlying element structure. We call this underlying structure deterministic because it is, in principle, perfectly predictable.

Many applications of the SRT paradigm have focused on investigating the explicit versus implicit nature of learners' knowledge of the underlying structure of the sequence (e.g., Perruchet & Amorim, 1992; Reber & Squire, 1994; Shanks & Johnstone, 1999; Willingham, Nissen, & Bullemer, 1989). However, because deterministic sequences often lead to anticipatory responses to every element, as well as explicit (conscious) knowledge of the sequence, it has not been possible to use steady-state RTs as a dependent measure of structure-dependent learning (i.e., participants demonstrate a floor effect due to explicit representation). Thus, a secondary task (e.g., tone counting) is sometimes added to prevent the formation of explicit representations of the underlying sequence (e.g., Reed & Johnson, 1994; it is also used to investigate the relationship between attention and implicit learning; e.g., Curran & Keele, 1993). Interestingly, learners in this dual-task SRT paradigm are often unable to report the underlying structure of the sequence, even though their RTs indicate that they have acquired implicit knowledge of this structure. After practice with the sequence, learners show faster RTs to the sequence than to brief, untrained modifications to the sequence.

Although the distinction between explicit and implicit learning is interesting and has received extensive study (see Stadler & Frensch, 1998), there are three reasons why this distinction is not the focus of our research. First, the definition of explicit knowledge is problematic. One can use a variety of tests to assess its presence, including open-ended verbal reports, rating scales, completion (generation) tasks, and two-alternative forced-choice (2AFC) judgments, and different tests lead to different results. One of the consequences of this variability is that the distinction between explicit and implicit learning can be quite test dependent (see Shanks & Johnstone, 1998, for a discussion of the assessment of implicit and explicit knowledge). Second, although there are examples in the real world of deterministic sequences (e.g., melodies), most of the events in the real world are probabilistic. Because probabilistic sequences rarely result in explicit knowledge of the underlying structure, the use of a dual-task SRT paradigm to prevent the intrusion of explicit knowledge may not be necessary when probabilistic sequences are used in a laboratory study. Third, recent work (Rah, Reber, & Hsiao, 2000; Schmidtke & Heuer, 1997) has raised concerns about the interpretation of dual-task experiments. Rather than acting to divert attention from the primary task, the secondary task may simply provide information that the learner processes along with information from the primary task in an attempt to find consistent environmental correlations. The decline in performance on the primary task may be due to the absence of reliable correlations between the primary and secondary tasks rather than to interference by the secondary task with the primary task.

A related literature on the learning of probabilistic sequences uses an artificial grammar learning paradigm (Reber, 1993). Reber (1967) presented participants with strings of visual stimuli (i.e., letters) organized according to a finite state grammar. The strings consisted of printed items presented simultaneously, rather than sequentially, with each string generated by a single pass through the grammar. Reber (1967) demonstrated that participants exposed to strings generated by such a finite state grammar can discriminate grammatical from nongrammatical test strings. Note that Reber's (1967) use of a finite-state grammar generated more variability in the strings to be learned than was present in most SRT tasks (e.g., Cohen et al., 1990). However, this variability was still considerably less than that found in natural languages or in the stimuli used in artificial language studies (e.g., Braine, 1963; Morgan, Meier, & Newport, 1987).

Subsequently, Reber (1969) found that changing the specific symbols used in test strings produced little or no interference in a transfer task, whereas changing the rules of the grammar did. He argued that what participants had become sensitive to was not the order of specific sequences of symbols but rather their underlying grammatical structure. Finally, Reber (1976) found that learners who were informed that there was some kind of structure in the input, and were encouraged to find it, performed less well on a grammaticality judgment task than did learners who were given neutral instructions. These results were interpreted as evidence of an interference effect produced by an explicit search for rules, which led to the formation of false hypotheses about the underlying structure of the input.

Cleeremans and McClelland (1991) were the first to marry the SRT task with the probabilistic structure of artificial grammar paradigms. Their sequences of lights were generated by a finite state grammar in which transitions from one light to the next were under probabilistic control. In addition, there was an injection of noise in the form of an occasional random substitution of a grammatical element in the sequence with an ungrammatical element. These factors prevented participants from perfectly anticipating the lights and thus from acquiring explicit knowledge of the sequence. Nevertheless, responses to transitions in the grammar that were legal had shorter RTs than did responses to transitions that violated the underlying structure of the grammar, indicating that implicit knowledge of the grammar had been acquired.

Stadler (1992) investigated the effects of different amounts of statistical structure on SRT performance in a fixed, repeating sequence, as well as the reliability of that statistical structure. Stadler found that fixed sequences that contained greater statistical structure (characterized as larger numbers of embedded two-, three-, and four-element sequences) produced faster improvement in RT than did fixed sequences containing less redundancy. He also found that when he varied the statistical structure from a completely random sequence to a completely deterministic, fixed sequence, RT performance improved with increasing approximation to a fixed sequence. Both of these findings lend support to the idea that learners are sensitive to and can exploit the overall level of both statistical redundancy and statistical regularity in a sequential stimulus stream.
Although many researchers have varied the level of statistical information in a sequential stimulus stream or the reliability of that information (e.g., Cleeremans & McClelland, 1991; Cohen et al., 1990; Stadler, 1992), they typically have relied on aggregate differences in predictability to demonstrate learning. These aggregate differences confound a number of different sources of statistical information. Thus, most applications of the SRT paradigm, whether using deterministic or probabilistic sequences, have not attempted to explore the specific information participants use to learn the underlying temporal structure. With short deterministic sequences one can solve the task by memorizing the exact order of elements, although in dual-task situations it seems clear that participants are unable to report the order information with great precision, suggesting the absence of a memorization strategy. Similarly, such a strategy is unlikely for learning long deterministic sequences, although it may be useful for learning parts of the longer sequence. In contrast, in probabilistic sequences one cannot use a pure memorization strategy because there is considerable variability in the order of elements, suggesting that participants learn local element relations. In the limit, these local relations consist of element pairs (e.g., Light No. 1 is often followed by Light No. 4).

Of course, learners may extract a large variety of statistics (e.g., Reed & Johnson, 1994), including the successive predictability of elements in the sequence (e.g., Cleeremans & McClelland, 1991) or a set of abstract rules (e.g., Reber, 1989). However, there must be constraints on the number and types of statistics that are extracted by the learner. Without constraints, the learning mechanism would face the combinatorial explosion problem: The number of possible statistics characterizing even a fairly simple learning set is infinite (e.g., the relation between element $x_n$ and element $x_{n+1}$ grows exponentially with increasing $n$). Even if learners do have constraints on the statistical hypotheses they are willing to consider in a given learning situation, they may not always initially settle on those that are most suited to the learning task. Thus, one might expect to find subclasses of learners who begin with different initial strategies in attempting to solve a given learning task before settling on the correct solution. In fact, depending on the nature of the input, one might even find learners who fail to settle on the best solution to the problem, which only highlights the question of what kinds and varieties of computations learners bring to different learning situations.

Even for the simple case of two successive elements in a temporal sequence, there are a number of different statistics that characterize the relation between those elements. For example, consider a simple finite-state grammar in which Element A is always followed by Element B, and Element B is followed by Elements C and D with equal probability. If this grammar generated 100 sequences, then Elements A and B would occur 100 times each and Elements C and D would occur 50 times each. Thus, one simple statistic that could be used to generate faster RTs to Element B than to Elements C or D is overall element frequency. That is, one could observe that Element B is more frequent than Elements C and D (although the 2:1 ratio might not be estimated precisely) and use this differential expectancy to respond faster to Element B, independent of its relative position in the temporal sequence.

A somewhat more complex statistic than element frequency is element probability, which normalizes the frequency of each element by the total frequency of elements in the sample. In the foregoing example, each pass through the grammar generates 3 elements, so 100 sequences contain a total of 300 elements. Thus, the element probabilities are .33 for Elements A and B and .17 for Elements C and D. Note that both first-order statistics—element frequency and element probability—provide the same information about the relative expectancy of Elements A and B versus Elements C and D. However, these two statistics involve different underlying computations (counting vs. counting and normalizing by division) and thus may be differentially affected by sample size, internal noise, and storage format.

Another statistic that could be used to respond faster to Element B than to Elements C and D in the foregoing example is bigram frequency. The finite-state grammar in this example generates different frequencies of pairs of elements, with pair (A, B) occurring 100 times and pairs (B, C) and (B, D) occurring 50 times. Thus, participants could learn that Element B follows Element A more frequently than Elements C and D follow Element B. A somewhat more complex statistic than bigram frequency is the joint probability (JP) of a bigram, which normalizes bigram frequency by the total frequency of bigrams in the sample. In the example above, the JP of the ordered pair (A, B) is .50, whereas the JP of (B, C) and of (B, D) is .25. Note that both bigram frequency and JP imply the extraction of temporal-order information.

One potential problem with bigram frequency or JP in learning temporal-order information becomes apparent when the finite state grammar becomes slightly more complex. Consider the foregoing example in which Element A is always followed by Element B, and Elements C and D follow Element B with equal probability. If we add further options to the grammar such that Element X can begin the sequence as often as Element A, and Element X is always followed by Element B, then 100 sequences generated by the grammar would yield 50 bigrams each of element pairs (A, B), (X, B), (B, C), and (B, D). Therefore, if RTs were based solely on bigram frequency (ignoring simple element frequency), they should be equal for Elements B, C, and D. However, if RTs were generated by a different statistic, one that involves predictability, then they should be faster to Element B than to Elements C and D because B always follows A, and B always follows X, but C and D each follow B only half of the time.

A statistic that can be used to compute the predictability of element pairs is conditional probability (CP), which normalizes the JP of a pair of elements by the probability of the first element in the pair. For example, the more complex finite-state grammar described above would yield CPs of 1.00 for Element B given Element A (B|A) and for Element B given Element X (B|X), whereas the CPs for Elements C and D given Element B were .50. Thus, if participants are extracting the predictability of element pairs rather than their frequency or probability of occurrence, then CP should provide a better characterization of the pattern of RTs to the sequence of elements in the SRT paradigm.

---

1 Unless otherwise noted in the text, CP refers to first-order conditional probability (e.g., the likelihood that B will follow A, also known as transitional probability). Similarly, unless otherwise noted, JP refers to bigram joint probability.
The extraction of these various sequential statistics has been explored extensively in the auditory domain by Safran, Newport, and Aslin (1996). They created an artificial language in which an inventory of 11 consonant–vowel syllables were generated with a speech synthesizer and combined to form six trisyllabic nonsense words. Words were then concatenated randomly into a continuous stream of speech, such that the timing of each syllable was identical, regardless of its position in the stream. Thus, there were no acoustic cues to word boundaries, and word segmentation could be performed only by learning the sequential statistics of the syllables. After 21 min of exposure to this speech stream, adults performed significantly above chance on a 2AFC test, indicating that they could distinguish between sequences of sounds that were words and sequences of sounds that were nonwords (trisyllabic strings consisting of novel sequences of familiar syllables). A second set of studies demonstrated that first-grade children, as well as adults, were able to succeed on this task even when the speech stream was presented incidentally and participants were attending to another task (Safran, Newport, Aslin, Tunick, & Barrueco, 1997). These findings demonstrate that adults and children possess statistical learning abilities, minimally at the level of extracting JPs, and that this learning proceeds automatically as a by-product of mere exposure.

A subsequent series of experiments (Safran, Aslin, & Newport, 1996) asked whether human infants can also detect statistical cues to word boundaries from a continuous stream of syllables. Eight-month-old infants were exposed to 2 min of a speech stream generated by a smaller artificial language containing four trisyllabic nonsense words. Although only statistical information was available to extract word boundaries, 8-month-old infants successfully distinguished the familiar words from nonwords and also from part words (trisyllabic strings consisting of familiar but less statistically predictable sequences spanning a word boundary). An additional experiment demonstrated that this learning process was based on the computation of the CPs of successive syllables, rather than a simpler computation of the frequencies or JPs of syllable pairs (Aslin, Safran, & Newport, 1998). All of these results strongly suggest that detection of sequential probabilities plays an important role in the process of word segmentation.

To determine whether the underlying mechanism was specific to linguistic materials, a series of follow-up experiments (Safran, Johnson, Aslin, & Newport, 1999) extended the study of statistical learning from speech to nonspeech (tone) stimuli. When a set of pure tones was substituted for the speech syllables used in the original Safran, Newport, and Aslin (1996) study, adults showed the same ability to learn tone sequences that they had shown for syllable sequences. When pure tones were substituted for the syllables in the artificial language used by Safran, Aslin, and Newport (1996) to test 8-month-old infants, their performance was again identical to their performance with speech stimuli. Thus, the statistical learning abilities in the auditory domain are general enough to operate on linguistic and nonlinguistic stimulus materials. Furthermore, Hauser, Newport, and Aslin (2001) have shown that tamarin monkeys perform the word segmentation task of Safran, Aslin, and Newport (1996), suggesting that this aspect of statistical learning is not unique to humans.

Two further series of experiments by Fiser and Aslin (in press-a, in press-b) have extended these investigations of statistical learning from the auditory to the visual modality. One series of studies substituted 12 shapes for the 12 syllables used in Safran, Aslin, and Newport (1996) and presented them at a rate of 1 per second in a continuous animation. The other series of studies created complex visual scenes from the spatial arrangement of 6 shapes (from the same inventory of 12), with each pair of shapes constrained to a particular relative position. These two series of studies demonstrated that adults are sensitive to the JPs and CPs of shape sequences, as well as to the JPs and CPs of shape positions. Thus, the same statistical learning mechanisms available to adults, children, and infants in the auditory modality appear to be used by adults in the visual modality.

The present series of experiments extends the foregoing studies to the learning of visuomotor sequences. Our goal was to use the statistical structure of a language-like word segmentation task, in which sequences of lights were grouped into units and presented in random order, and to track learning over time using an SRT task. The input was designed to be a mixture of the deterministic (i.e., fixed within-unit structure) and probabilistic sequences (i.e., random ordering of units) used in previous SRT studies and to provide a link to the literature on the segmentation and extraction of speech and tone patterns in the auditory domain, as well as to studies of shape sequences and configurations in the visual domain. The current set of experiments is distinguished from previous work using the SRT paradigm (e.g., Cleermans & McClelland, 1991; Cohen et al., 1990; Stadler, 1992) in that it represents an attempt to identify which statistics learners are exploiting in order to improve their performance, rather than simply characterizing how differences in aggregate predictability produce differences in learning. Our initial experiments provided baseline measures of how variations in a given statistic were reflected across time during learning. Our final experiment was aimed at assessing the ability of learners to simultaneously exploit differences in the predictability of a given statistic when information from other statistics was unavailable, as well as to simultaneously exploit different statistics in the same stimulus stream.

**Experiment 1**

The purpose of Experiment 1 was to establish the utility of the languagelike design (a mixture of deterministic and probabilistic sequential structure) used by Safran, Newport, and Aslin (1996) in the context of the SRT paradigm. One limitation of the fixed familiarization and single posttest paradigm used by Safran, Newport, and Aslin (1996) is that it cannot provide an on-line measure of learning. In contrast, the SRT paradigm provides a measurement of performance (RT) on every trial, thereby allowing us to chart the time course of separation between RTs for sequences that are more or less predictable. Previous studies using a probabilistic SRT paradigm (e.g., Cleermans & McClelland, 1991) have not separated RT performance by the predictability of specific transitions within the finite-state grammar. Rather, RT differences between all of the grammatical sequences and a variety of ungrammatical sequences have been reported. Averaged within these grammatical sequences were individual transitions ranging in predictability from high (e.g., 1.00) to moderate (e.g., .33), and the ungrammatical transitions had predictabilities of zero. Rather than introducing ungrammatical sequences into the SRT task to assess learning, we simply charted the changes over trials (and sessions) in RTs to transitions that varied in predictability. We expected that
transitions with high predictability would eventually result in faster RTs than transitions with low predictability. Experiment 1 contained large differences in the range of predictability for subsets of transitions to maximize the chances of obtaining significant RT differences after a period of learning.

Method

Participants. Data were collected from 10 adults (4 men and 6 women) ages 19 to 28, all of whom were undergraduates at the University of Rochester. Participants were paid $7.50 per hour, regardless of performance.

Apparatus. An on-line measure of sequence learning was obtained from a button box that recorded RT responses to visuospatial sequences. The button box consisted of a 10 × 17 × 2 in. (25.4 × 43.2 × 5.1 cm) aluminum box with eight red button switches, ¾ in. (1.9 cm) in diameter, that could be illuminated under computer control. Seven buttons were arranged in a semicircle with a radius of 5-in. (12.7-cm) (see Figure 1). An eighth button was positioned at the center of the base of the semicircle. This eighth button served as the location for a “home” orienting response prior to the onset of each trial. The button box was interfaced to an Apple Quadra 650 microcomputer using a hardware-software system (designed by James R. Savusich at the Speech Research Laboratory, State University of New York at Buffalo) that illuminated the buttons and recorded the RT responses with millisecond accuracy.

Design. We sought to design a stimulus sequence with a complexity at least as great as that already shown to be learnable by adults in the auditory domain. Because the sequences used by Saffran, Newport, and Aslin (1996) contained six trisyllabic words, we built our SRT sequences from seven groups of 3 elements. If each element were composed of a single button, we would need a total of 21 unique elements. However, we also sought to present participants with a visuomotor task that would be reasonably tractable, and 21 spatially discrete elements did not meet that requirement. Accordingly, we decided to present each element not as a single button but rather as a pair of simultaneously illuminated buttons. Pairs of lights were chosen as the basic element because they afforded a greater degree of complexity while limiting the number of buttons to a manageable spatial distribution. By using pairs of the seven non-home buttons, 21 unique pairs of lights could be presented in the semicircle. To achieve this same level of complexity with 21 single buttons would have resulted in a more complicated spatial array that might have impeded learning.

For convenience, we adopt the following terminology: Element refers to a pair of lights, which is the smallest coherent unit in the stimulus stream, and word refers to a fixed sequence of three pairs of lights. By convention, the buttons in the semicircle were numbered from left to right as shown in the bottom panel of Figure 1. The 21 unique elements (pairs of buttons) were organized into seven unique words (triplets of pairs). Table 1 illustrates the structure of the words, with each row corresponding to a word. For example, Word 1 in Table 1 is composed of Elements A, B, and C, each of which corresponds to a pair of lighted buttons. Table 2 indicates the two possible assignments of lights to the elements and words used in the experiment. Which particular lighted buttons correspond to Elements A, B, and C depends on which button assignment a given participant received. For example, if the participant received Button Assignment 1, then by consulting Table 2 we can see that for this participant Elements A, B, and C correspond to the button pairs (2, 4), (3, 7), and (5, 6), respectively. Each participant received one of these two assignments for all sessions of the experiment. The use of two assignments ensured that not all participants were exposed to the same spatial distribution of buttons. This helped guard against the undesirable scenario in which all participants might be exposed to an idiosyncratic spatial distribution of buttons that was, for some reason, more (or less) easily learned. An example of the physical sequence of lights corresponding to Word 1 from Assignment 1 (Elements A, B, C) is shown in Figure 1.

Within each word no individual button was illuminated more than once (to control for repetition effects). Thus, within any word six of the seven buttons were lit at some point as part of the word’s component elements. In addition, each of the 21 possible elements (button pairs) appeared only once in all of the words. Each stimulus block was formed by randomly sequencing the 7 words. Thus, each word had to occur once within any

Table 1

<table>
<thead>
<tr>
<th>Word</th>
<th>1</th>
<th>2a</th>
<th>3a</th>
<th>→</th>
<th>1 of next word</th>
<th>CP</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A</td>
<td>B</td>
<td>C</td>
<td>→</td>
<td>D, G, J, M, P, S</td>
<td>.17</td>
</tr>
<tr>
<td>2</td>
<td>D</td>
<td>E</td>
<td>F</td>
<td>→</td>
<td>A, G, J, M, P, S</td>
<td>.17</td>
</tr>
<tr>
<td>3</td>
<td>G</td>
<td>H</td>
<td>I</td>
<td>→</td>
<td>A, D, J, M, P, S</td>
<td>.17</td>
</tr>
<tr>
<td>4</td>
<td>J</td>
<td>K</td>
<td>L</td>
<td>→</td>
<td>A, D, G, M, P, S</td>
<td>.17</td>
</tr>
<tr>
<td>5</td>
<td>M</td>
<td>N</td>
<td>O</td>
<td>→</td>
<td>A, D, G, J, P, S</td>
<td>.17</td>
</tr>
<tr>
<td>6</td>
<td>P</td>
<td>Q</td>
<td>R</td>
<td>→</td>
<td>A, D, G, J, M, S</td>
<td>.17</td>
</tr>
<tr>
<td>7</td>
<td>S</td>
<td>T</td>
<td>U</td>
<td>→</td>
<td>A, D, G, J, M, P</td>
<td>.17</td>
</tr>
</tbody>
</table>

*CP = 1.0.

Although we were concerned that the dual-button stimuli might lead to a situation in which learning was based on a bias toward one side of the apparatus (e.g., right-handed learners consistently responding first to the right-most button in button pair stimuli), no consistent evidence was found for such a side-bias among participants in either Experiment 1 or Experiment 2.

Figure 1. Light sequence for Assignment 1, Word 1 (Elements A, B, and C), Experiment 1. Note that elements consist of pairs of lights in the semicircle and that element illumination alternates with an illuminated “home” button.
given block, which balanced both element and word frequency across blocks. Ten independently randomized blocks were then concatenated into a single 70-word session, with the constraint that words could not repeat at the edges of blocks (again, to control for repetition effects). This meant that each word was presented 10 times within each session. Participants took part in 48 independently generated sessions.

Table 3 shows the CPs and JPs of each of the element transitions in Experiment 1. The CPs within words (i.e., from the first to the second element and from the second to the third element) were 1.0. This is because both the identity and order of elements within words were fixed, and hence the last two elements of any word were completely predictable. However, the CPs at the boundaries between words were much lower than those within words because of the random ordering of words within blocks, so that across the input corpus there was a mean probability of .17 that the third element of a word would be followed by the first element of any other word.

The likelihood of encountering any given pair of within-word elements as one made a pass through a randomized block of seven words was 1.0 because each within-word pair was guaranteed to occur once out of the 21 within-block pairs. Thus, within-word JPs were .04762 for each unique, within-word pair. However, the likelihood of encountering a given pair of elements at a word boundary was not guaranteed but instead depended on the order in which the words were randomized. Thus, on average, across blocks, the probability of encountering any given pair of between-word elements was .00794. It is important to note that any observed effects in this experiment cannot be unequivocally attributed to CP or JP alone, because both CP and JP are more predictable for within-word transitions than they are for between-word transitions.

Procedure. Participants were seated in a sound-attenuated booth, with the button box placed on their lap. Participants were instructed to push buttons only with their preferred hand and to steady the box with their other hand. At the start of an experimental session, all of the buttons were illuminated for 1 s to indicate that the session was about to begin, after which all the lights were extinguished except for the home button. Participants were instructed to press and hold the home button until a pair of buttons (corresponding to an element) in the semicircle was illuminated. When the element button pair was illuminated, the home button was simultaneously extinguished.

Participants were informed that the experiment involved RTs, but no mention was made of learning or of patterns embedded in the sequential stimuli. Participants were instructed simply to press the illuminated buttons as rapidly as possible while maintaining high accuracy. They were told that the order in which they pressed the buttons in the button pairs was unimportant but that the buttons should be pressed in succession (to avoid any tendency to simultaneously press two adjacent buttons). Both buttons in the pair remained illuminated until two buttons in the semicircle had been pressed. At that point, regardless of which two buttons had been pressed, the illuminated button pair was extinguished and the home button was simultaneously illuminated. The participant returned to the home button, pressing it and holding it down until the onset of the next stimulus pair (element). After returning to the home button, there was a 100-ms delay before the onset of the next stimulus pair. This delay period was included to prevent participants from leaving the home button prematurely before the onset of the next stimulus pair. If the participant anticipated the next stimulus by releasing the home button before illumination of the stimulus pair, the stimulus pair was not illuminated. In such cases, the participant was required to return to the home button and the delay period was reintroduced. If the participant correctly delayed release of the home button, the same stimulus pair that would have been presented during the previously aborted event was illuminated. No feedback was given to participants at any point during the 6-day training period.

Stimuli were generated independently for each participant at the beginning of each session and were presented as a single continuous sequence of events. Participants took part in eight 70-word sessions during the course of a single, 1-hr-long training period, with 1–2 min rest breaks between sessions. One 8-session training period was conducted each day for 6 consecutive days, yielding a total of 48 sessions for each participant.

Data collection. RT data were collected throughout each session by recording the sequence of all button presses. Both button position (1–7) and RT were recorded for the first response to a stimulus pair (element),

Table 3

<table>
<thead>
<tr>
<th>Position</th>
<th>Element transition</th>
<th>CP</th>
<th>JP</th>
</tr>
</thead>
<tbody>
<tr>
<td>3 → 1</td>
<td>(C, F, I, L, O, R, U) to (A, D, G, J, M, P, S)</td>
<td>0.17</td>
<td>0.00794</td>
</tr>
<tr>
<td>1 → 2</td>
<td>AB, DE, GH, JK, MN, P, ST</td>
<td>1.0</td>
<td>0.04762</td>
</tr>
<tr>
<td>2 → 3</td>
<td>BC, EF, HI, KL, NO, PR, TU</td>
<td>1.0</td>
<td>0.04762</td>
</tr>
</tbody>
</table>

*Not all permutations of transitions from Position 3 to Position 1 were permissible under the experimental design because the same word never appeared twice in immediate succession.

3 JPs were calculated by determining the minimum number of blocks one would need to possibly encounter all possible combinations of transitions in question. Thus, for within-word JPs, only a single block is needed to encounter all possible within-word transitions, and there is a 1/21 (0.4762) chance of encountering any given within-word transition. However, because word-final elements can transition to any of six other elements, the frequencies used in calculating between-word transitions must be based on six blocks. There are a total of 42 possible between-word transitions, each of which would occur once, along with an additional 84 within-word transitions across six consecutive blocks. Thus, the likelihood of encountering any given between-word transition is 1/98 (i.e., 1/(42 + 84) or 0.00794).
whereas only button position was recorded for the second response. RT was not recorded for the second response because the distance between the two buttons of a given stimulus varied and we had no objective way of “normalizing” the resulting variation in RTs. In contrast, first responses were always initiated from the home button, which was equidistant from each of the buttons in the peripheral semicircle. Note that RT was defined as the time from stimulus onset to the first button press. A total of 10,080 RT measurements were possible across the entire 6 days of the experiment for each participant. If the participant pressed an unilluminated button, either as a first or second button press, the response was counted as an error and the RT was discarded. The average error rate throughout the experiment across participants was 1.31% (range = 0.18%–3.36%).

When participants had completed the final (48th) session, but prior to being debriefed about the details of the experiment, they were asked if they had any general impressions about the nature of the task. All participants reported that they thought they occasionally noticed patterns in the way the pairs of lights were presented, but there was no correlation between explicit knowledge of these patterns and performance.

**Results and Discussion**

We hypothesized that with extended training, participants would become sensitive to the structure within the sequence of stimulus elements (button pairs), as indicated by differential RTs. Specifically, we predicted that RTs for first elements, which have low CP and JP, would begin to plateau before RTs for second and third elements, which are maximally predictive. As shown in Figure 2, our results support this hypothesis. Across the first few sessions, there was general improvement in RTs for all three elements, which likely reflects the participants’ adaptation to the general task of making rapid button-press responses from the home location. Subsequently, RTs for the second and third elements continued to decrease, whereas RTs for the first element began to level off. By session 48, the gap between RTs for first versus second and third elements was approximately 65 ms. Note that this RT difference must have been based on some analysis of element pairs because the frequency of all 21 individual elements was equated within each session.

An analysis of variance (ANOVA) with session (1–48) and element position (1, 2, or 3) as repeated measures indicated significant main effects of session, $F_{\text{sess}}(4, 984, 44, 859) = 40.802, p < .001, \text{MSE}_{\text{adj}} = 9, 654, 745$, and of element position, $F_{\text{Elem}}(1.581, 14.225) = 21.347, p < .001, \text{MSE}_{\text{adj}} = 6, 896, 729$, as well as a significant interaction, $F_{\text{sess} \times \text{Elem}}(6.320, 56.881) = 12.082, p < .001, \text{MSE}_{\text{adj}} = 2, 020, 239$. Significant RT differences were also found between Elements 1 and 2, and between Elements 1 and 3, but not between Elements 2 and 3. $F_{1, 9} = 28.346, p = .001, \text{MSE} = 9, 630, 331; F_{1, 9} = 23.341$.

---

4 Because of the two-factor within-subjects experimental design, we were concerned about violating assumptions of sphericity essential to the statistical model of the ANOVA. For that reason, when reporting results from the ANOVA, we adjusted the degrees of freedom using the Huynh-Feldt method to compensate for the possibility of such violations, wherever appropriate. Because it is the adjusted degrees of freedom that were used to evaluate the probability value of the $F$ statistic reported in the text, we report the adjusted degrees of freedom in the text. Thus, for example, although the unadjusted degrees of freedom for the first statistic reported in Experiment 1 are $F_{\text{sess}}(47, 423)$, the adjusted degrees of freedom that we report are $F_{\text{sess}}(4, 984, 44, 859)$. Similarly, we report the adjusted mean square error in the text ($\text{MSE}_{\text{adj}}$). This method of reporting adjusted degrees of freedom and mean square error is used throughout the article, wherever appropriate.
\( p = .001, \text{MSE} = 17,609.363; \text{and } F_{20}(1, 9) = 2.577, p > .05, \text{MSE} = 5,462.388. \)

It is clear from Figure 2 that the rate of improvement for Element 1 is slower than for Elements 2 and 3, and that the improvement for Elements 2 and 3 is roughly equivalent. It is interesting to note that RTs for Element 2 are somewhat slower than for Element 3, but not significantly so. This suggests that there is little or no cumulative advantage in predicting Element 3 from Elements 1 and 2, compared with predicting Element 3 from Element 2 alone. A separate ANOVA conducted on the last eight sessions alone showed no significant RT difference between Element 2 and Element 3, confirming this lack of advantage.

**Experiment 2**

It could be argued that the statistical structure of Experiment 1 was both simple and exaggerated because all the within-word transitions were perfectly predictable. This resulted in a large difference in predictability between the within-word and the between-word button pair transitions (a ratio of 6:1 for both CP and JP). To assess the generality of the results obtained in Experiment 1, we reduced the within-word CPs in Experiment 2 from 1.00 to .50. The between-word CPs remained unchanged at .17. JPs were unaffected by these changes in sequence structure.

**Method**

**Participants.** Data were collected from 10 naive adults (3 men and 6 women), ages 19 to 20, all of whom were undergraduates at the University of Rochester. Participants were paid $7.50 per hour, regardless of performance.

**Apparatus and procedure.** These were identical to Experiment 1. The average error rate throughout the experiment across participants was 1.26% (range = 0.48%–2.73%).

**Design.** The basic structure of the button assignments and button sequences in Experiment 2 was similar to Experiment 1. Seven words consisting of three elements each were constructed, with elements consisting of pairs of lighted buttons. However, in contrast to Experiment 1, in which each element appeared in only one word, some elements in Experiment 2 appeared in more than one word. This reduced the CPs between elements within words from 1.0 to .5.

To lower the within-word CPs, we used the same elements in Experiment Positions 1 and 2 but did not use the same elements in the same words (i.e., elements in Element Position 2 were shifted across words; see Table 4).

**Table 4**

<table>
<thead>
<tr>
<th>Element position</th>
<th>1</th>
<th>2*</th>
<th>3*</th>
<th>( \text{of next word} )</th>
<th>CP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word</td>
<td>A</td>
<td>B</td>
<td>C</td>
<td>H</td>
<td>.17</td>
</tr>
<tr>
<td>1</td>
<td>A</td>
<td>B</td>
<td>C</td>
<td>D</td>
<td>.17</td>
</tr>
<tr>
<td>2</td>
<td>A</td>
<td>B</td>
<td>C</td>
<td>E</td>
<td>.17</td>
</tr>
<tr>
<td>3</td>
<td>A</td>
<td>B</td>
<td>C</td>
<td>F</td>
<td>.17</td>
</tr>
<tr>
<td>4</td>
<td>A</td>
<td>B</td>
<td>C</td>
<td>G</td>
<td>.17</td>
</tr>
<tr>
<td>5</td>
<td>A</td>
<td>B</td>
<td>C</td>
<td>H</td>
<td>.17</td>
</tr>
<tr>
<td>6</td>
<td>A</td>
<td>B</td>
<td>C</td>
<td>D</td>
<td>.17</td>
</tr>
<tr>
<td>7</td>
<td>A</td>
<td>B</td>
<td>C</td>
<td>E</td>
<td>.17</td>
</tr>
</tbody>
</table>

*CP = .5. * Elements in Position 2 were identical to those in Position 1 but were assigned to different words.

This, in combination with a unique third element for each word, resulted in a set of seven unique three-element words. However, it also produced within-word CPs of .5 (because any given element in first or second position has exactly two elements that can legally follow it; e.g., when A is in Position 1 it is followed by E, and when A is in Position 2 it is followed by K) and average between-word CPs of .17 (because any given third element can be followed by any one of six other elements at the beginning of the next word). Note that JPs in this experiment are the same as in Experiment 1 because the factors affecting the JP of each possible element pair have not changed, even though the possible combinations of elements in pairs have changed. Thus, just as in Experiment 1, within-word element pairs occur exactly once in any given block (JPs of .04762), and between-word element pairs occur probabilistically across blocks as a result of randomization (JPs of .00794). The assignment of lights to elements and words is given in Table 5.

Table 6 shows the CPs and JPs of each of the element transitions possible in Experiment 2. Note that it is still not possible to unambiguously attribute any RT differences between Element 1 and Elements 2 or 3 to either CP or JP because both CP and JP for Elements 2 and 3 are greater than for Element 1.

**Results and Discussion**

As in Experiment 1, we hypothesized that participants who become sensitive to the structure within the sequence of stimulus elements (button pairs) should reflect this sensitivity through differentially faster RTs. Specifically, we predicted that RTs for first elements, which have low CP and JP, would begin to plateau before RTs for second and third elements, which are more predictive. As shown in Figure 3, our results support this hypothesis. Across the first few sessions, there was a general improvement in RTs for all three elements, which likely reflects the participants’ adaptation to the general task of making rapid button-press responses from the home location. Subsequently, RTs for the second and third elements continued to decrease, whereas RTs for the first element began to level off. By session 48, the gap between RTs for first versus second and third elements was approximately 40 ms.

An ANOVA with session (1–48) and element position (1, 2, or 3) as repeated measures indicated significant main effects of session, \( F_{20}(7.667, 69.004) = 23.763, p < .001, \text{MSE}_{\text{adj}} = 5,695.324 \), and of element position, \( F_{20}(1.553, 13.976) = 12.835, p < .001, \text{MSE}_{\text{adj}} = 6,094.704 \), as well as a significant interaction, \( F_{20}(7.534, 67.805) = 7.135, p < .001, \text{MSE}_{\text{adj}} = 1,207.628 \). Significant RT differences were also found between Elements 1 and 2, and between Elements 1 and 3, but not between Elements 2 and 3, \( F_{20}(1, 9) = 17.707, p < .005, \text{MSE} = 9,281.416; F_{20}(1, 9) = 13.206, p = .005, \text{MSE} = 15,029.390; \text{and } F_{20}(1, 9) < 1 \).

It is clear from Figure 3 that the rate of improvement for Element 1 is slower than for Elements 2 and 3, and that the improvement for Elements 2 and 3 is roughly equivalent. This suggests that there is little or no cumulative advantage in predicting Element 3 from Elements 1 and 2, compared with predicting Element 3 from Element 2 alone. As in Experiment 1, a separate ANOVA conducted on the last eight sessions alone showed no significant RT difference between Element 2 and Element 3, confirming this lack of advantage.

For comparison, the results from Experiments 1 and 2 are plotted together in Figure 4. It is of interest that in both experiments, RTs for Element 1, which are equally predictable in terms of CP and JP, plateau at approximately the same level. However, RTs for Elements 2 and 3, which have CPs of 1.0 and .5 in
Table 5
Assignments of Elements to Button Pairs, Experiment 2

<table>
<thead>
<tr>
<th>Word</th>
<th>Assignment 1</th>
<th>Assignment 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Element to button pair assignments</td>
<td>Element to button pair assignments</td>
</tr>
<tr>
<td>1</td>
<td>A (1, 7)</td>
<td>A (5, 6)</td>
</tr>
<tr>
<td>2</td>
<td>B (2, 6)</td>
<td>B (2, 7)</td>
</tr>
<tr>
<td>3</td>
<td>C (1, 5)</td>
<td>C (3, 5)</td>
</tr>
<tr>
<td>4</td>
<td>D (3, 6)</td>
<td>D (1, 2)</td>
</tr>
<tr>
<td>5</td>
<td>E (4, 5)</td>
<td>E (3, 4)</td>
</tr>
<tr>
<td>6</td>
<td>F (3, 7)</td>
<td>F (1, 6)</td>
</tr>
<tr>
<td>7</td>
<td>G (2, 4)</td>
<td>G (4, 7)</td>
</tr>
</tbody>
</table>

Experiments 1 and 2, respectively, but the same JPs (.04762), show different trajectories. To test these observations, we conducted post hoc comparisons on the last eight sessions (the terminal performance) of Experiments 1 and 2, comparing each element position between the two experiments. To compensate for the high between-subjects variability due to the small sample size, we first converted the data to Z scores. No difference was found between Experiments 1 and 2 for Element Position 1, F_{Exp1/Exp2} (1, 18) < 1, but comparisons for Element Positions 2 and 3 were just at the level of significance, F_{Exp1/Exp2} (1, 18) = 4.403, p = .05, MSE = 0.592, and F_{Exp1/Exp2} (1, 18) = 4.387, p = .051, MSE = 1.165. The only difference in predictability between Experiments 1 and 2 for Element Positions 2 and 3 is in terms of CP, not JP. Interestingly, the more predictable CPs for Elements 2 and 3 in Experiment 1 are reflected in faster RTs than those in Experiment 2. These results suggest that differences in relative predictability of elements, even when JPs are held constant, may be reflected in differentially rapid RTs. The next experiment provided a more direct test of the ability of participants to simultaneously detect stimulus subsets that differed in relative predictability when either CP or JP was held constant.

Experiment 3

The goal of Experiment 3 was to determine which of two statistics learners would use in an SRT paradigm whose structure was similar to the one used in Experiments 1 and 2. Specifically, in Experiment 3 we attempted to compare results from two subsets of transitions that were equated for JP but differed in CP, as well as two subsets that were equated for CP but differed in JP. This is the same design that was used by Aslin, Saffran, and Newport (1998) in their study of infants’ segmentation of speech streams. If learners rely solely on JPs or CPs, then when that statistic is equated across subsets of transitions, performance should not differ. Alternatively, if learners can make use of whatever statistic provides information about underlying structural differences in event sequences, then they should be able to rely on that informative statistic.

Pilot studies using word and element structures similar to those of Experiments 1 and 2 (i.e., elements composed of button pairs, and seven 3-element words) resulted in considerable variability across participants. We suspected that the task was too difficult for some participants, perhaps because of the stimuli used (elements consisting of two lighted buttons), the response required (motor responses to two distinct locations per element), the complexity of the sequential structure (words consisting of three elements in combination with a greater variety of predictability than in the first two experiments), or some combination of these factors. We were uncertain what the effects of each of these factors (or their interaction) might have been on performance and were unable to deduce the reason for the highly variable results obtained with the first 5 participants. Therefore, we decided to terminate the experiment and redesign it to address the above concerns by simplifying the task. The number of words was reduced from seven to four, each word was reduced from three elements to two elements, and the elements were changed to single buttons rather than button pairs. Additional changes in stimulus structure allowed us to compare results from data subsets that were equated for JP but differed in CP, as well as subsets that were equated for CP but differed in JP.

Method

Participants. Data were collected from 10 naive adults (2 men and 8 women), ages 19 to 26, all of whom were undergraduates at the University.

Table 6
Conditional Probability (CP) and Joint Probability (JP) Combinations, Experiment 2

<table>
<thead>
<tr>
<th>Position</th>
<th>Element transition</th>
<th>CP</th>
<th>JP</th>
</tr>
</thead>
<tbody>
<tr>
<td>3 → 1</td>
<td>(H, I, J, K, L, M, N) to (A, B, C, D, E, F, G)⁵</td>
<td>.17</td>
<td>.00794</td>
</tr>
<tr>
<td>1 → 2</td>
<td>AE, BF, CG, DA, EB, FC, GD</td>
<td>.5</td>
<td>.04762</td>
</tr>
<tr>
<td>2 → 3</td>
<td>AK, BL, CM, DN, EH, FI, GJ</td>
<td>.5</td>
<td>.04762</td>
</tr>
</tbody>
</table>

⁵ Not all permutations of transitions from Position 3 to Position 1 were permissible under the experimental design because the same word never appeared twice in immediate succession.
of Rochester. Participants were paid $7.50 per hour, regardless of performance.

**Apparatus.** A new button box was constructed to allow for eight response buttons in addition to the home button. These eight response buttons were arranged in a semicircle with a radius of 5 in. (12.7 cm), with the ninth, home button positioned at the center of the base of the semicircle. The dimensions of the box and the buttons were the same as in the previous experiments.

**Design.** Three design changes were implemented in Experiment 3. First, each response element now consisted of a single illuminated button rather than a button pair. This was intended to reduce the attentional and processing load imposed by the individual element stimuli, as well as the processing load required to produce a motor response. Second, each word consisted of two rather than three elements. This reduced the number of button presses per word from six in Experiments 1 and 2 to only two in Experiment 3. Finally, the overall design consisted of only four unique words of two elements each, with no light used in more than one word. This resulted in a reduced pool of words from which to structure the stimulus stream, thereby reducing the overall variability possible within the stimulus stream. These changes in experimental design greatly reduced the task demands confronting the participants and, we hoped, would produce more consistent and interpretable results.

The third design change involved altering word frequency and sequencing. In contrast to Experiments 1 and 2, the overall sequence of words did not contain an equal number of each word. Words 3 and 4 occurred twice as frequently as Words 1 and 2 and were further constrained such that one was followed 50% of the time by the other and 50% of the time by Word 1 or Word 2. Table 7 illustrates the design and the within- and between-word CPs, with each row corresponding to a word (e.g., AB is a word composed of the Elements A and B, each of which corresponds to a single lighted button). The assignment of lights to elements is given in Table 8. An example of the sequence of lights corresponding to Word 1 from Assignment 1 (Elements A, B) is shown in Figure 5.

Two additional constraints stipulated that words could not repeat at the edges of blocks and that when Word 3 was followed by Word 4 or Word 4 was followed by Word 3, the subsequent word had to be Word 1 or Word 2. Furthermore, because the number of words was so small, we decided to double the size of the pool within which words were randomized into blocks. Thus, following the above constraints, 12 words were randomly sequenced to form a block, and 10 blocks were then concatenated to form a 120-word session. This design resulted in several subsets of transitions with different combinations of CP and JP. Table 9 shows these CP and JP data subsets as well as the predicted rank ordering of RTs according to increasing predictability. Table 9 also shows a factor labeled position, which refers to the possibility that participants learned the cumulative predictability of the two more frequent words (e.g., that EFG better predicted H than did G alone).

Note that Table 9 predicts two different results for the last data subset, S6, depending on whether participants are able to detect position as a meaningful statistic. If position can be detected and used, then Data Subsets S5 and S6 should be discriminable, with an RT enhancement for S6. On the other hand, if position cannot be detected or used by participants, then one would expect no difference in predictability between S5 and S6 and, hence, no difference in RT.6

---

6 Specifically, because Word 3 is followed by Word 4, and vice versa, 50% of the time, both words can be characterized as sometimes occurring in either Position 1 or Position 2 relative to one another. This is similar to Experiments 1 and 2, in which Elements 2 and 3 had the same CPs and JPs but differed in element position. In those experiments no advantageous effect was found for Position 3 over Position 2. In the current experiment, position is relevant to word position rather than element position. If participants become sensitive to this positional substructure, they should have faster RTs to the second element of words in Position 2 than to those in Position 1. However, given the lack of positional effect found in Experiments 1 and 2, we were uncertain whether such an effect would be found in Experiment 3.
Figure 4. Mean reaction times (RTs) by element position across sessions, Experiments 1 and 2. Exp = Experiment.

Procedure. The total number of RT measurements per participant across the 32 sessions was 7,680. The average error rate throughout the experiment across participants was 0.33% (range = 0.05%–0.86%). Participants took part in eight 120-word sessions during the course of a single, 1-hr-long training period, with 1–2 min rest breaks between sessions. One 8-session training period was conducted each day for 4 consecutive days, yielding a total of 32 sessions for each participant. This resulted in a greater number of words per session than in the previous experiments (120 vs. 70) but, because there were fewer elements per word (two vs. three), only a slightly greater number of responses per day (1,920 vs. 1,680).

Results and Discussion

Group findings. As in Experiments 1 and 2, there are differences in the relative predictability of the different element positions across words. For any given word, Element 2 is on average more predictable than is Element 1 (Element 2 has CP = 1.0 and JP = .08334, whereas Element 1 has CP = .5 and JP = .08334). Thus, as in the first two experiments, we hypothesized that participants who become sensitive to the structure within the sequence of stimulus elements should reflect this sensitivity through differentially faster RTs to more predictable element positions. This is shown in Figure 6. By session 32, the gap between RTs for first versus second elements was approximately 100 ms.

An ANOVA with session (1–32) and element position (1, 2) as repeated measures indicated significant main effects of session, $F_{\text{sess}}(5.843, 52.585) = 43.009, p < .001, MSE_{(adj)} = 5,282.796$, and of element position, $F_{\text{Elem}}(1, 9) = 25.212, p = .001, MSE_{(adj)} = 28,760.198$, as well as a significant interaction, $F_{\text{sess}\times\text{Elem}}(7.369, 66.323) = 17.421, p < .001, MSE_{(adj)} = 873.561$. A significant RT difference was also found between Elements 1 and 2, $F_{1,9}(1, 9) = 25.212, p = .001, MSE = 57,520.396$.

The data were also examined for CP, JP, and position corresponding to the data subsets listed in Table 9. We had predicted, in those cases in which CP or JP was balanced between two subsets, that the subset with the stronger of the unbalanced factors would show differentially faster RTs. In the one case in which both CP and JP were balanced (Data Subset S5 vs. S6), any observed effect must be due to participants’ sensitivity to position. The results, plotted by data subset, are shown in Figure 7. As can be seen, the rank order of the results according to the data subsets was as predicted in Table 9, and there was an apparent effect of position.

An ANOVA with session (1–32) and combinations of CP, JP, and position as repeated measures indicated significant main effects of session, $F_{\text{sess}}(6.202, 55.820) = 43.924, p < .001, MSE_{(adj)} = 14,858.353$, and CP–JP–position, $F_{\text{CP–JP–pos}}$,

Table 7

<table>
<thead>
<tr>
<th>Word</th>
<th>1</th>
<th>2$^a$</th>
<th>→</th>
<th>I of next word</th>
<th>CP</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 A</td>
<td>B</td>
<td>→</td>
<td>E, G</td>
<td>.5</td>
<td></td>
</tr>
<tr>
<td>2 C</td>
<td>D</td>
<td>→</td>
<td>E, G</td>
<td>.5</td>
<td></td>
</tr>
<tr>
<td>3 E</td>
<td>F</td>
<td>→</td>
<td>A, C</td>
<td>.25</td>
<td></td>
</tr>
<tr>
<td></td>
<td>E</td>
<td>→</td>
<td>G</td>
<td>.5</td>
<td></td>
</tr>
<tr>
<td>4 G</td>
<td>H</td>
<td>→</td>
<td>A, C</td>
<td>.25</td>
<td></td>
</tr>
<tr>
<td></td>
<td>G</td>
<td>→</td>
<td>E</td>
<td>.5</td>
<td></td>
</tr>
</tbody>
</table>

$^a$ CP = 1.0.
(2.323, 20.911) = 25.180, p < .001, MSE(adj) = 57,499.988, as well as a significant interaction, \( F_{\text{main}} \times CP \times JP \times set = (18.406, 165.656) = 10.876, p < .001, \text{MSE(adj)} = 2,626.351. \) Significant RT differences were also found between each of the rank-ordered data subsets from Table 9 (plotted in Figure 7), except between the fourth and fifth subsets (which differ only in JP), and the fifth and sixth subsets (which differ only in position), \( F_{\text{adj}}(1, 9) = 16.270, p < .005, \text{MSE} = 55,256.113; F_{\text{adj}}(1, 9) = 16.289, p < .005, \text{MSE} = 18,978.367; F_{\text{adj}}(1, 9) = 5.866, p < .05, \text{MSE} = 23,298.105; F_{\text{adj}}(1, 9) = 4.880, p = .055, \text{MSE} = 15,792.198. \)

A stepwise regression analysis revealed that session, CP, and JP were significant factors affecting performance, but position was not. Together, the significant factors accounted for 58.4% of the total variance, with sessions accounting for 33.3%, CP accounting for 24.7%, and JP accounting for 0.4% of the total variance. Although the largest factor aside from that of session (which is a motor practice effect) is that of CP, it is important to note that CP and JP are highly intercorrelated (\( r = 0.808 \)). Thus, although at first glance it seems that CP alone accounts for most of the variance in performance, the fact that JP is nonetheless a significant factor suggests that there may be an effect of JP that is largely masked by its correlation with CP. 

Identification of individual differences. The last comparison in the ANOVA (S5 vs. S6) was marginally significant, suggesting that some participants may have been able to detect and use position as a meaningful statistic. Earlier pilot work had also indicated that there were likely to be large differences in individual performance. Accordingly, we decided to investigate the performance of individual participants. A simple way to assess data subset comparisons was needed without reproducing a full set of RT curves for each individual. Figure 8 shows the results of an analysis in which the grand mean was calculated for each participant and used as the basis for calculating averaged Z scores based on the data from each day’s training. This allowed us to derive Z scores for different subsets of a given participant’s data that could then be used to trace changes in the relative difference between those data subsets. As an example, consider the changes over time in RT to Element 1 as opposed to Element 2. In general, the gap between Element 1 and Element 2 increases gradually with increasing training for all participants. The Z score difference between Element 1 and Element 2 in Figure 8 is a measure of the relative size of that gap (indicated by the amplitude of the difference), as well as its direction (indicated by the sign, plus or minus, of the difference, which shows whether Element 1 RTs are faster or slower than Element 2 RTs). As can be seen in Figure 8, the difference between Element 1 and Element 2 produces positive values as well as a generally increasing positive trend across days for each participant. This indicates that RTs to Element 2 are faster than to Element 1 and that this difference increases with training (see also Figure 6). This is what one would predict if Element 2 was generally more predictable than Element 1.

Similar increasing trends of positive values are expected for data subsets that differ in predictability along only a single dimension.

---

It should be noted that although the predicted results cannot be explained by the frequency of single elements alone (which are balanced within each block throughout Experiment 1 and which produce predictions contrary to the observed results from Experiments 2 and 3), one might ask whether a number of statistics other than CP and JP could possibly account for the results of any of our experiments. One such statistic that has been widely considered in the literature is second-order conditionalism (SOC, the ability to predict C given that one has already encountered A followed by B). Another statistic that might be reasonable to consider is the frequency or JP of trigrams. It turns out that by itself SOC produces a different set of predictions than the observed results in all of our experiments and, hence, is not considered in detail. Although the results of Experiment 3 are most parsimoniously accounted for by combinations of CP and JP, the results of Experiments 1 and 2 could be accounted for by trigram JP. Furthermore, it remains moot that some additional, more complex statistic or combination of greater numbers of statistics could yield predictions that are in accord with our results. However, we appeal to parsimony and processing simplicity as legitimate arguments for winnowing the hypothesis space of possible mechanisms brought to bear in solving the tasks set before our participants. Thus, given the three reasonable alternatives that can account for the results of Experiments 1 and 2 (i.e., CP, bigram JP, and trigram JP), we propose that in the absence of biases, the human information processor will likely seek the simplest and most efficient solution to the problem. This allows us to discard trigram JP as a likely mechanism for our results simply because it requires the participant to keep track of information that occurs over a wider temporal and spatial bin width than do statistics based on bigrams. Although this does not allow us to determine which of the bigram statistics participants might preferentially rely on to solve the learning task of Experiments 1 and 2, we emphasize that in Experiment 3, the observed data can only be accounted for by simultaneously and differentially exploiting the more predictive of either CP or JP.
Table 9
Conditional Probability (CP), Joint Probability (JP), and Position, Experiment 3

<table>
<thead>
<tr>
<th>Data subset</th>
<th>Element transition</th>
<th>CP</th>
<th>JP</th>
<th>Position</th>
<th>Predicted rank order</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>(H, F) to (A, C)</td>
<td>0.25</td>
<td>0.4167</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>S2</td>
<td>(B, D) to (E, G)</td>
<td>0.5</td>
<td>0.4167</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>S3</td>
<td>FG, HE</td>
<td>0.5</td>
<td>0.08334</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>S4</td>
<td>AB, CD</td>
<td>1.0</td>
<td>0.08334</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>S5</td>
<td>EF, GH</td>
<td>1.0</td>
<td>0.16667</td>
<td>1 or 2</td>
<td>5 or 6</td>
</tr>
<tr>
<td>S6</td>
<td>EF, GH</td>
<td>1.0</td>
<td>0.16667</td>
<td>1</td>
<td>5</td>
</tr>
</tbody>
</table>

* The experimental design applies certain restrictions to these transitions. See text for explanation. b The rank order of Data Subset S6 depends on whether participants are capable of detecting and using information about the relative positions of EF and GH. Fifty percent of the time Word 3 is followed by Word 4, producing EFGH, and, similarly, 50% of the time Word 4 is followed by Word 3, producing GHEF. Hence, both EF and GH can occur in either first or second position relative to one another. If participants become sensitive to this aspect of the input, it should be reflected in faster reaction times when EF and GH occur in the second position.

of either CP or JP. For illustration, the difference of one pair of data subsets equated for CP (Data Subsets S2 and S3) and that of another equated for JP (Data Subsets S3 and S4) are plotted in Figure 8. In both cases, the subset predicted to be faster is subtracted from that predicted to be slower, so that the sign of the data points should be positive. The CP-equated data showed a generally increasing trend across days of training. This reflects a sustained improvement in RT for the more predictable of the two CP-equated data subsets with continued training (i.e., an effect of JP). However, the overall pattern for the JP-equated data subsets is not as uniform across participants. Three participants in particular (Participants 3, 5, and 7) stand out as having more than one negative score, indicating that for a large portion of training they responded more slowly to the more predictable transition. Indeed, although Participant 5 showed positive values for the last 2 days of training, Participant 3 never attained a positive score, and Participant 7 actually showed increasingly negative values across days. Thus, there appears to be a subset of participants who are performing differently than the other participants, even though all participants are receiving the same underlying structure and showing overall improvement in global task performance (Element 1 slower than Element 2; see Figure 8, dashed lines). This led us to ask how this subset of participants might be processing the input in order to produce such results.

Characterization of individual differences. To address this question, we examined the sequential pattern of average RTs for a given element transition within each session. If it was faster, this implied that the participant was relatively more certain about the subsequent element in the sequence of stimuli and was able to anticipate the correct transition and produce a faster response. Conversely, if the RT was slower, this implied that the participant was less certain of the subsequent stimulus element, resulting in a slower response. Consider, for example, the transition from A to B in a given session. If the average RT for A in that session is 390 ms and the average RT for B is 370 ms, the RT for B is faster than the RT for A, and this difference in RT would be assigned a "-". Conversely, if the average RT for B is 410 ms, the RT for B is slower than for A, and this difference would be assigned a "+".

One way to analyze the data is to look at RTs as they occur across multiple transitions. For example, sequences such as EAB or FCD (each of which has two transitions) should generate a "+ -" pattern of RT differences because the first transition (between words) is relatively unpredictable but the second transition (within word) is highly predictable. A sequence of repeating "+ -" patterns of RTs implies that participants are processing the button sequences as bigrams. However, if we wish to determine whether some participants are processing the sequences as trigrams, then we must consider sequences of RTs that span three transitions (e.g., EABG or GHEF). The entire stimulus sequence of Experiment 3 can be subdivided into six smaller sequences of element transitions corresponding to the data subsets listed in Table 9. All of the possible element transitions correspond to the following sequence of these data subsets: S1–S4–S2–S5–S3–S6 (see bottom right panel of Figure 9). For convenience, we analyze this six-transition sequence in two smaller, three-transition parts: S1–S4–S2 and S5–S3–S6. All of the possible stimulus strings (element sequences) that correspond to these two data subset sequences are listed in Table 10, and these strings exhaust the combinations possible in this experiment. There are eight possible combinations of "+" and "-" values that can be generated by this three-transition analysis, each corresponding to a particular pattern of expected predictability on the part of the participant. Each of these combinations is shown in the key in Figure 9.

Note that all of the elements in a given sequence position (e.g., S1, S3, etc.) in Table 10 have the same CP–JP combinations because they all correspond to the same data subset. For each participant in each session, we produced a mean RT for each sequence position (column). Then we determined for each sequence position whether the mean RT to that position was faster (--) or slower (++) than the RT that preceded it. In this way, we were able to determine each participant's expectations for the predictiveness of each of the different data subsets as they occurred in sequence, as well as how their expectations changed with increased training. On the basis of the embedded predictability of the data subsets, one would expect a "+ - +" pattern of RT differences for the S1–S4–S2 sequence and a "- - -" pattern for the S5–S3–S6 sequence.

As an example, consider the element sequences that contribute to the S1–S4–S2 sequence of data subsets. One such sequence is

---

8 Analyzing all six data subset transitions as a single sequence would have produced 64 different possible combinations of plus and minus values, which would have been impossible to present in a figure. Using two sequences of three transitions allowed us to look for bigram, trigram, and quadgram processing patterns and also to present the data compactly.
HABE (see Element Sequence 1 in Table 10). Transitions into the first element (H) of this particular sequence must be from Element G when GH occurs in Position 2 (this is because GH is a word, and only when GH is in Position 2 does it transition into Word AB). Table 9 shows that the data subset corresponding to this transition is S6, which has a CP of 1.0 and a JP of .16667. Both CP and JP are highly predictive in this case (relative to the transitions for other data subsets), and the RT for this transition should be relatively fast. However, the element following H in this example is A. By consulting Table 9, one can see that the transition from H to A corresponds to the data subset S1, which has a CP of .25 and a JP of .04167, both of which are less predictive than were the CP and JP for Element H. Thus, one would expect the RT to Element A to be slower than the RT to Element H, resulting in the assignment of a “+” to the difference in RT between the first and second elements in the sequence. The next element in the sequence is B, and the transition from A to B corresponds to Data Subset S4, which has a CP of 1.0 and a JP of .08334. In this case, the CP and JP associated with the transition to Element B are more predictive than those associated with the transition to Element A, resulting in the assignment of a “−” to the difference in RT between the second and third elements in the sequence. The last element in the sequence is E, and the transition from B to E corresponds to Data Subset S2, which has a CP of .5 and a JP of .04167. In this case, the CP and JP associated with the transition to Element E are less predictive than those associated with the transition to Element B, resulting in the assignment of a “+” to the difference in RT between the third and fourth elements in the sequence. When all of the element transitions in the sequence are considered together, the pattern of changes in RT one would predict based on the relative CP and JP values of the individual element transitions is “+−−”, and this same pattern is predicted for each of the individual sequences labeled as S1–S4–S2. Similar reasoning applies to the “−−−” pattern predicted for the S5–S3–S6 sequences.

The top two panels of Figure 9 display the results of this analysis for individual participants, and as can be seen, a somewhat varied collection of patterns is evident. Although all participants showed some degree of variation from the predicted pattern of responses (usually in the earlier sessions), the majority of participants eventually settled into the predicted patterns after the first day of training (after Session 8). However, Participants 3, 5, and 7 showed persisting differences (solid black cells) throughout Days 2 and 3 and into Day 4. The question to be asked is what these differences can tell us about how these 3 participants were processing the input differently from the other participants.

The answer to this question can be found by examining the number of transitions over which facilitative RTs are present. Thus, for the majority of participants, the predicted patterns of “+−−” for transitions S1–S4–S2 and “−−−” for transitions S5–S3–S6 involve RTs that alternate between relatively fast (within-word transitions) and relatively slow (between-word transitions). These participants were classified as “bigrain” processors because they appear to have been extracting information from the stimulus stream that led them to develop expectations about the predictiveness of the sequence that spans two elements. This pattern of expectation corresponds exactly to what one would need in order to segment the continuous input stream into 2-element words (see bottom right panel of Figure 9).

In contrast, Participants 3, 5, and 7 showed a “+−−” pattern for many sessions. This sometimes occurred for S1–S4–S2 or S5–S3–S6 sequences alone but often occurred for both sequences in the same session. This “+−−” pattern indicates that these partic-
Participants showed faster RTs for two adjacent transitions (spanning three adjacent elements) after they encountered a slower transition. This suggests that these participants were processing at least part of the stimulus stream as triplets of elements (see bottom right panel of Figure 9). Accordingly, they were classified as "trigram" processors. Note that this pattern of facilitative RTs would lead to a segmentation of the input stream that would not correspond to words.

Although we classified Participants 3, 5, and 7 as trigram processors, it should be noted that they each demonstrated a different distribution of trigram processing over the course of the 4 days of training. Participant 5 showed trigram processing primarily in the S1–S4–S2 sequence, with only three sessions in which trigram processing occurred simultaneously in both the S1–S4–S2 and the S5–S3–S6 sequences. In comparison, Participant 3 showed a much greater degree of simultaneous trigram processing, with that processing persisting through Days 2 and 3. Both of these participants, however, appear to have shifted their processing strategy to bigrams for both S1–S4–S2 and S5–S3–S6 sequences during Day 4. This is in strong contrast to Participant 7, who showed the opposite trend, with increasing trigram processing as training increased. In fact, for most of Day 4, Participant 7 showed trigram processing for both S1–S4–S2 and S5–S3–S6 sequences.

To provide a more quantitative representation of this analysis, the RT differences (rather than just their plus–minus direction) of the S1–S4–S2 and S5–S3–S6 subsets for each session from the last 3 days (Sessions 9–32) were transformed and replotted in polar coordinates (see Figure 10). Major differences can be seen in the performance of bigram and trigram processors (left and right plots, respectively). Note that bigram processors had the majority of their values in the "+−−" and "−−−" sectors, both of which correspond to the predicted patterns for bigram processing. In contrast, trigram processors had a large number of values in the "+−−" sector, which corresponds to the predicted pattern for trigram processing. In addition, neither bigram processors nor trigram processors had any values located in the "−−−" sector, which indicates that none of the participants processed sequences EFGE or GHEF as quadgrams.

The fact that none of the participants showed evidence of quadgram processing suggests that for bigram processors, relative position of the more frequent bigrams was not a factor that they were able to extract from the stimulus sequence. Accordingly, when the data subset RTs are replotted by processor type, there should be no difference for bigram processors between the plots.

---

9 Only the last 3 days were plotted because much of the variation found in the 1st day is likely due to adaptation to the task and would obscure the patterns more clearly present in later learning. Three axes are represented on each polar plot in Figure 10. These axes correspond to the three values that can be attributed to a sequence of three data subsets. The first axis corresponds to the first value of the data subset sequence and is represented on the polar plot as magnitude. Because all values of the data subsets can be positive or negative, the magnitude axis has been offset such that zero is located on the inner circle. Positive values along this axis increase toward the periphery, and negative values increase toward the center (see the key, Figure 10). The other two axes correspond to the second and third values of the data subset sequence, which have been transformed from a two-dimensional Cartesian space into polar coordinates. Thus, second values in the data subset sequence that are positive are located in the right semicircle, and negative values are located in the left semicircle. Third values in the data subset sequence that are positive are located in the upper semicircle, and negative values are located in the lower semicircle.
Figure 8. Z-score differences for each participant, Experiment 3. Z scores were calculated for each participant separately. The grand mean of a participant’s performance across all 4 days was used as the basis for calculating Z scores of all reaction time (RT) data. RT Z scores were then averaged by day for each of the different data subsets. Differences between these data-subset Z-score averages were then calculated and plotted in the figure. JPs = joint probabilities; CPs = conditional probabilities.

for Data Subsets S5 and S6. However, there should be an effect of position for trigram processors because at some point they all showed a “+--” pattern for S5–S3–S6 sequences, indicating that RTs to Data Subset S5 (Position 1) were slower than to Data Subset S6 (Position 2). As can be seen in Figures 11 and 12, the replotted data support these predictions.10

Because 3 of the 10 participants were not consistently sensitive to the underlying bigram structure of the sequence, we conducted a separate analysis on the 7 bigram processors. No statistics were calculated for the trigram processors due to the small number of participants. ANOVA on the data from these bigram processors (N = 7) with sessions (1–32) and combinations of CP, JP, and position as repeated measures indicated significant main effects of session, F(31, 233) = 3.599, p < .001, MSE(adj) = 14,901.901, and CP–JP–position, F(31, 11980) = 31.956, p < .001, MSE(adj) = 52,537.729, as well as a significant interaction, F(31, 11980) = 16.249, 97.496) = 10.436, p < .001, MSE(adj) = 2,782.442. Significant RT differences were also found between each of the adjacent rank-ordered data subsets from Table 9 (plotted in Figure 11), F(1, 6) = 7.210, p < .05, MSE = 67,812.094; F(1, 6) = 17.444, p < .01, MSE = 19,484.484, F(1, 6) = 42.468, p = .001, MSE = 6,572.005; F(1, 6) = 18.783, p = .005, MSE = 1,687.671; and F(1, 6) = 6.090, p < .05, MSE = 674.857.

Inspection of Figure 11 suggests that the significant difference between Data Subsets S5 and S6 (supporting an effect of position) was probably due to an effect seen early in training, before RTs begin to plateau. When a post hoc ANOVA was conducted using only data from the last eight sessions (final day) of bigram processors, significant differences were found between each of the adjacent, rank-ordered data subsets from Table 9, except subsets S5 and S6: for all comparisons, p < .05, except F(1, 6) = 1.349, p > .05, MSE(adj) = 608.039. The lack of advantage

10 It should be noted that all of the participants who appear to have processed the stimuli as trigrams were also given Button Assignment 1, raising the possibility that their performance may have been due to some idiosyncrasy in Assignment 1 that biased them to process the input as trigrams rather than as bigrams. To test this hypothesis, we conducted an unplanned comparison between groups based on button assignments (Assignment 1 vs. Assignment 2). No significant effect of assignment was found, F(8, 63) = 1.643, p > .05, MSE = 18,375.666, suggesting that the tendency to process the input as bigrams or trigrams was independent of the particular button assignment participants received. However, an estimate of effect size, partial omega squared, suggests that the effect was of medium size, \( \omega^2_{\text{partial}} = .0604 \). Therefore, it is possible that, because of the small sample size, there may not have been enough power to detect a meaningful effect.
Figure 9. N-gram analyses for each participant (rows) by session (columns), Experiment 3. The top left panel shows results for the S1–S4–S2 sequence of data subsets. The top right panel shows results for the S5–S3–S6 sequence of data subsets. In the key and the bottom right panel, a plus sign indicates a slower reaction time (RT) to a given element than to the previous element. A minus sign indicates a faster RT. Bigram processors are predicted to show a "+-+" pattern (gray) for the top left panel and a "-+-" pattern (white) for the top right panel. Trigram processors are predicted to show a "++-" pattern (black) in either panel. The bottom right panel shows an idealized representation of the predicted RT patterns for bigram processors (left graph) and trigram processors (right graph) as a function of the repeating data subset sequence S1–S4–S2–S5–S3–S6.
Table 10
Element Sequences Corresponding to the Data Subset Sequence, Experiment 3

<table>
<thead>
<tr>
<th>Element sequence</th>
<th>Data subset sequence:</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Part 1</td>
</tr>
<tr>
<td></td>
<td>(S6) S1 S4 S2</td>
</tr>
<tr>
<td>1</td>
<td>H A B E</td>
</tr>
<tr>
<td>2</td>
<td>H A B G</td>
</tr>
<tr>
<td>3</td>
<td>F A B E</td>
</tr>
<tr>
<td>4</td>
<td>F A B G</td>
</tr>
<tr>
<td>5</td>
<td>H C D E</td>
</tr>
<tr>
<td>6</td>
<td>H C D G</td>
</tr>
<tr>
<td>7</td>
<td>F C D E</td>
</tr>
<tr>
<td>8</td>
<td>F C D G</td>
</tr>
<tr>
<td></td>
<td>Part 2</td>
</tr>
<tr>
<td></td>
<td>(S2) S5 S3 S6</td>
</tr>
<tr>
<td>9</td>
<td>E F G H</td>
</tr>
<tr>
<td>10</td>
<td>G H E F</td>
</tr>
</tbody>
</table>

for Position 2 over Position 1 suggests that participants were not able to process Words 3 and 4 or Words 4 and 3 as larger units.11

When a separate stepwise regression analysis was conducted using data from the last eight sessions (final day) of bigram processors (N = 7), session, CP, and JP were once again found to be significant factors affecting performance (position was not included in the regression analysis because it was found to be a significant factor in the corresponding ANOVA).12 The amount of total variance accounted for by these factors was 71.1%, with CP accounting for 68.5% and JP accounting for 2.6%. Session did not account for sufficient variance to enter the regression. Note that the primary effect of removing the trigram processors from the analysis and focusing on terminal performance was to greatly reduce the effect of session. Although it appears that, as bigram processors approach terminal performance, the largest effect is that of CP, it is important to note that CP and JP are highly intercorrelated (r = .761), again raising the possibility that there is an effect of JP that is largely masked by its correlation with the effect of CP.

In summary, Experiment 3 identified two different classes of processor: those who preferentially process the element stream as bigrams and those who do so as trigrams. The results from all participants taken together support the rank ordering of RT results predicted in Table 9, which indicates that participants were able to exploit both CP and JP at the same time and also supports the hypothesis that participants were able to use information regarding the relative position of element bigrams. However, results from the terminal performance of bigram processors, whose processing conforms to the underlying structure of the input, indicate that although bigram processors exploited both CP and JP at the bigram level, they were not able to use relative bigram position as a third source of statistical information. Thus, Experiment 3 shows that learners were able not only to simultaneously use information based on two different statistics (CP and JP) but also to exploit the different levels of predictability within each of these statistics.

General Discussion

The present series of experiments has extended the SRT paradigm to a task analogous to the statistically based extraction of words from fluent speech studied by Saffran, Newport, and Aslin (1996). In our SRT task, several pairs or triplets of elements (analogous to syllables in words) were randomly concatenated into a continuous sequence (roughly analogous to a continuous stream of unsegmented words in speech). RTs were used to assess participants’ learning of the predictability of each element in the sequence. The statistical structure of these sequences differed from previous SRT studies in that there was neither a single deterministic sequence of elements nor a probabilistic sequence that was compared only with a random sequence. Our results showed that RTs to low-predictability elements were reliably slower than to high-predictability elements, even when both low- and high-predictability element sequences occurred in the exposure stream.

It was also possible with the designs used in our studies to determine that participants were capable of exploiting more than one statistic in order to learn the predictiveness of elements in the sequences. The simplest statistic one could use for making predictions is the frequency of occurrence of elements. That is, if some elements occur more frequently than other elements, this could trigger higher expectancies for those elements, which would then correlate with shorter RTs. In Experiment 1, the frequency of elements was equated and was therefore not available as a predictor of RTs. In Experiment 2, element frequency predicted results that were not observed. Thus, the faster RTs in Experiments 1 and 2 could not have been based on element frequency and must have been based instead on learning the temporal order of pairs (or triplets) of elements. At least two temporal-order statistics were both available to learners in Experiments 1 and 2: JPs of element pairs and CPs of element pairs. Either one (or both) of these statistics could have enabled learners in Experiments 1 and 2 to show shorter RTs to the more predictable element transitions.13

---

11 Close comparison of Figures 11 and 12 reveals that the rank-ordering of the data subsets for bigram and trigram processors is not the same. Bigram processors demonstrate the ordering of RTs that one would expect if participants were not sensitive to the predictive value of position (see Table 9). Trigram processors show the predicted rank ordering for Data Subsets S1, S2, S3, and S6, but Data Subsets S4 and S5 appear to be shifted upward, indicating slower responses than predicted. Data Subset S5 corresponds to Position 1 and has already been accounted for as a consequence of trigram processing. Data Subset S4, however, cannot be accounted for by trigram processing. A separate analysis revealed that some of the RTs contributing to Data Subset S4 (transitions from C to D) were responsible for shifting the plot of Data Subset S4 upward. The performance of Participant 3 was especially skewed for this portion of the data, and close analysis suggests that during the latter half of training he was attempting to process the transitions CDEF and CDGH as quadrags, alternating between one and the other with a “~”~“~” pattern of responding. We have no explanation for this performance aside from individual participant differences and the fact that only 3 participants contributed to the plots in Figure 12.

12 Note that because bigram processors did not become sensitive to position, Data Subsets S5 and S6 were averaged into a single data subset.

13 In addition to CP and JP, two other statistics, SGCs and trigram JPs, were available to learners in Experiments 1 and 2. However, only trigram
In Experiment 3, JPs were held constant whereas CPs were allowed to vary, and vice versa. As in studies of auditory sequence learning (Aslin et al., 1998), visual sequence learning (Fiser & Aslin, in press-a), and spatial pattern learning (Fiser & Aslin, in press-b), learners in Experiment 3 were able to rely on differences in CPs when JPs were held constant and to rely on differences in JPs when CPs were held constant. Although the overall variability of RTs in Experiment 3 was largely accounted for by CPs, the fact that CP and JP were highly intercorrelated raises the possibility that the observed effect of JP was largely a consequence of participants' having computed CP. That is, it could be the case that participants computed CP and, by virtue of having done so, had some knowledge of the relative frequencies of occurrence of the different bigrams constituting different CP-equated data subsets. The current design does not allow us to determine if participants were computing JP per se.

An important feature of our results and analyses was the characterization of the learning styles of individual participants in Experiment 3. Seven of the 10 participants clearly learned the element pair structure of the sequences. In contrast, 3 of the 10 participants in Experiment 3, despite showing overall RTs that were faster for the more predictable Element 2 in the element pairs than for Element 1, did not show consistent learning of the element pair (bigram) structure of the sequences. A detailed analysis of the performance of these 3 participants suggested that they processed
at least some of the element pairs as element triplets. Thus, the SRT paradigm can be “learned” at several different levels of structure: a global level that conforms to the most fundamental predictiveness of elements in the sequence and a local level that is characterized by clusters of elements. This same detailed analysis of the 7 participants who consistently processed the elements as bigrams showed that they did not process sequences of the more frequent element pairs (EF and GH) as quadgrams. However, it is possible that with further exposure to the statistics of the sequence, such a hierarchical processing of elements (some as pairs and some as pairs of pairs) would have emerged.

The results of the present series of experiments are similar to those of Cleeremans and McClelland (1991), who used a finite-state grammar to generate the element sequences. However, Cleeremans and McClelland assessed learning by randomly interspersing ungrammatical element transitions within the overall grammatical sequence. Although they did examine whether a series of elements was a better predictor than a single element, and whether one element could predict another element separated by intervening elements, they did not determine which of a variety of statistics best predicted the RTs to elements in the sequence. Clearly, human learners can keep track of simple element frequency (Hasher & Zacks, 1984). However, like other work in the animal learning literature (e.g., Rescorla & Wagner, 1972), our results suggest that participants in an SRT task use a conditionalized statistic (i.e., CP) as the basis for estimating predictability. Note that we are not claiming that CP is always used in preference to other simpler statistics. It seems likely that differences in the frequency or JP of elements or element pairs are used in temporal-order tasks, especially when the differences are large (e.g., a factor of 2 or more). However, there are many situations in which these frequency-based statistics can lead to errors in estimating predictability (e.g., when the frequency of some elements is very low). Under these circumstances, CPs provide a more reliable estimate of predictability than either frequency or JP.

Are there simpler computational mechanisms than those described here that could enable learners to show the same pattern of RTs in these experiments? Perruchet and Vinter (1998) argued that adult, child, and infant performance in the studies of Saffran and colleagues (Saffran, Aslin, & Newport, 1996; Saffran, Newport, & Aslin, 1996; Saffran et al., 1997) can be accounted for by a computational model that does not calculate CPs. In their model, the learner sequentially extracts from the input sequence one to three element-chunks that conform to potential parsing units held in memory. These newly extracted chunks are themselves combined and stored in memory as new parsing units and either maintained as active parsing units, by encountering identical chunks later in the sequence, or shifted to the status of potential parsing units, which may be discarded if identical chunks are never (or rarely) encountered again. The underlying structure of the three studies of Saffran and colleagues provided learners with more frequent within-word element sequences than element sequences that spanned a word boundary. The Perruchet and Vinter model settled to a state that only had words or combinations of words as potential parsing units, and correctly gave the greatest weight to chunks corresponding to words, thereby “solving” the word segmentation task in the studies of Saffran and colleagues.

Perruchet and Vinter (1998) did not model the follow-up experiment by Aslin et al. (1998) in which the frequency of some within-word element chunks was identical to the frequency of some between-word element chunks. This frequency-balanced condition, which was also used in Experiment 3 of the present

Figure 11. Mean reaction times (RTs) by combinations of conditional probability (CP), joint probability (JP), and position (POS) across sessions. Experiment 3, bigram processors only.
Figure 12. Mean reaction times (RTs) by combinations of conditional probability (CP), joint probability (JP), and position (POS) across sessions. Experiment 3, trigram processors only.

A series of studies, can be learned by infants and by adults. It remains to be seen whether Perruchet and Vinter’s model is successful at this frequency-balanced learning task. More important, however, their model illustrates the fact that a variety of algorithms can be used to learn the structure of sequential patterns, particularly when more than one source of information is available in the input. For example, in the studies of Saffran and colleagues (Saffran, Aslin, & Newport, 1996; Saffran, Newport, & Aslin, 1996; Saffran et al., 1997) both differential co-occurrence frequency (and JP) and CP provided information that could be used to discriminate between words and part words. However, element frequency (and element probability) did not provide such differential information.

Ultimately, the goal of our research program is to identify the mechanisms that learners bring to bear on certain classes of problems (e.g., complex, sequentially ordered input such as natural language) and the conditions under which they successfully extract information about the structure of that input (Newport & Aslin, 2000). Just as Chomsky (1955/1975, 1957, 1981) has proposed in the domain of language acquisition, learners in any purely induction-based learning situation must be constrained in the number and kinds of learning strategies they are willing to consider. Presumably, there are many hypotheses a learner could entertain when attempting to extract regularities from an information stream, most of which would be largely if not totally uninformative. In fact, as the length of the input under consideration increases linearly, the number of potentially hypotheses open to consideration explodes exponentially, eventually leading to a situation in which the learner would be incapable of evaluating the available hypotheses in a reasonable amount of time. Without constraints, the learner would attempt to extract an infinite number of statistics, thereby facing a computational explosion that requires an infinite amount of processing resources. Thus, it seems that learners must bring a rather more limited set of hypotheses to this type of task, effectively restricting the dimensionality of the hypothesis space they are willing to entertain as plausible.14

What source of information do learners make use of, and how do they access it? One possibility is that learners deploy, in serial order, a predetermined set of learning algorithms, moving from one to the next if the initially deployed algorithm cannot detect information in a sample of the input. Another possibility is that learners deploy a set of learning algorithms in parallel, with performance determined either by a weighted average of algorithms that detect information in the input or by the single algorithm that detects the most salient information. Regardless of whether information processing proceeds by a serial or parallel process, there must be constraints on which algorithms are applied to the input (and hence what statistics are learned). Although it is unclear whether participants are computing JP per se using an independent algorithm, the results of these experiments suggest that CP (or a similar conditionalized statistic) is in fact computed and that it provides a good account of much of the information learners are extracting in the current experiments when more than one source is available in the input. Note that there is no a priori reason why either infant or adult learners would necessarily be

14 Note that we are making no claims as to the specific factors that constrain the set of potential hypotheses or whether such factors are innate, learned, task specific, or species specific. A number of factors seem likely to contribute to limiting the hypotheses in the learner’s repertoire, including, perhaps, limitations on short-term memory, long-term memory, attention, perception, and computational capacity.
limited to a single statistical-learning algorithm. As shown by the frequency-balanced conditions in Aslin et al. (1998) and in Experiment 3 of the present series of experiments, learners are clearly capable of exploiting information in the input stream that can be formally characterized by different statistics. The fact that JP cannot be completely disentangled from CP (because CP is in part based on the frequency of bigrams) renders us incapable of determining whether JP could be computed independently of CP. Nonetheless, because the SRT paradigm—in contrast to the fixed familiarization and posttest paradigm used by Saffran, Newport, and Aslin (1996) and Saffran et al. (1997)—can assess learning as exposure to the input unfolds, it may prove particularly powerful in future studies that directly manipulate the relative informativeness of various statistics in the input.

In summary, the work presented in this article made use of the SRT paradigm to mimic certain aspects of the structure of natural language input, even though the specific response required of the learner is quite different from what occurs in natural language learning. Because the SRT paradigm allows us to control the statistics to which the learner is exposed, and also allows us to track learning of those statistics over time, we were able to determine that learners are capable of simultaneously exploiting different levels of at least two specific statistics, CP and JP, from the input. In addition, careful investigation of individual RT patterns revealed that participants did not constitute a homogeneous learning group but that there were two types of learners, those who preferentially processed the input as bigrams and those who preferentially processed the input as trigrams. This is important because it indicates that although all participants learned something about the structure of the input, as indicated by improved RTs at the level of within- and between-word element position, not all learners acquired the same information about the more subtle variations in predictability that were available in the input. Future research should examine how these individual differences in the kinds of statistics and processing styles implemented during learning affect the efficiency, accuracy, and time course of sequence learning in different domains.

References


Reed, J., & Johnson, P. (1994). Assessing implicit learning with indirect tests: Determining what is learned about sequence structure. Journal of...


Received October 31, 2000
Revision received February 1, 2001
Accepted February 1, 2001