COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 48 (2024) e13404

ISSN: 1551-6709 online
DOI: 10.1111/cogs.13404

The Keys to the Future? An Examination of Statistical Versus Discriminative Accounts of Serial Pattern Learning

Fabian Tomaschek,[a,b] Michael Ramscar,[c] Jessie S. Nixon[d]

[a] *Quantitative Linguistics Group, Eberhard Karls University of Tübingen*
[b] *Institut für Germanistik, Universität Bern*
[c] *Department of Psychology, Eberhard Karls University of Tübingen*
[d] *Linguistics and Cultural Studies, Carl von Ossietzky University Oldenburg*

Received 28 June 2022; received in revised form 8 December 2023; accepted 19 December 2023

Abstract

Sequence learning is fundamental to a wide range of cognitive functions. Explaining how sequences—and the relations between the elements they comprise—are learned is a fundamental challenge to cognitive science. However, although hundreds of articles addressing this question are published each year, the actual learning mechanisms involved in the learning of sequences are rarely investigated. We present three experiments that seek to examine these mechanisms during a typing task. Experiments 1 and 2 tested learning during typing single letters on each trial. Experiment 3 tested for "chunking" of these letters into "words." The results of these experiments were used to examine the mechanisms that could best account for them, with a focus on two particular proposals: statistical transitional probability learning and discriminative error-driven learning. Experiments 1 and 2 showed that error-driven learning was a better predictor of response latencies than either n-gram frequencies or transitional probabilities. No evidence for chunking was found in Experiment 3, probably due to interspersing visual cues with the motor response. In addition, learning occurred across a greater distance in Experiment 1 than Experiment 2, suggesting that the greater predictability that comes with increased structure leads to greater learnability. These results shed new light on the mechanism responsible for sequence learning. Despite the widely held assumption that transitional probability learning

Correspondence should be sent to Fabian Tomaschek, Quantitative Linguistics Group, Eberhard Karls University of Tübingen, 72074 Tübingen, Germany and Institut für Germanistik, Universität Bern, 3012 Bern, Schweiz. E-mail: fabian.tomaschek@unibe.ch or Jessie S. Nixon, Carl von Ossietzky University of Oldenburg, Building A10, Ammerländer Heerstraße, 26129 Oldenburg, Germany. Email: jessie.nixon@uni-oldenburg.de

is essential to this process, the present results suggest instead that the sequences are learned through a process of discriminative learning, involving prediction and feedback from prediction error.

Keywords: Sequence learning; Discriminative learning; Error-driven learning; Statistical learning; Pattern learning

1. Introduction

Speech production and writing can be described as processes during which sequences of articulatory events are produced. Sentences are made up of words that in turn comprise serial patterns of articulatory events. Accordingly, explaining how these serial patterns—and the grammars governing them—are learned, and identifying the mechanisms that give rise to their learning are important to our understanding of language production. The present study addresses the second of these questions in three experiments in which participants were required to learn serial patterns. The goal was to examine two different learning mechanisms that have been put forward in the literature, and to establish which best accounts for the pattern of data observed.

The first learning mechanism is based on the statistical and distributional properties of events, such as frequencies and conditional probabilities, and in particular, transitional probabilities (Ellis, 2006, Saffran, Aslin, & Newport, 1996a). We refer to this mechanism as "statistical learning" (see also Rebuschat & Monaghan, 2019, for a special issue on that topic). Statistical learning models are generative models: the goal of learning is to uncover the underlying statistical distribution in the relevant environment. We contrast this with a second, error-driven mechanism, which we refer to as "discriminative learning" (Nixon, 2018, 2020, Nixon & Tomaschek, 2021, Ramscar, 2021, Ramscar, Dye, & McCauley, 2013b, Ramscar, Yarlett, Dye, Denny, & Thorpe, 2010, Ramscar & Dye, 2009, Ramscar & Yarlett, 2007).

Error-driven learning uses a process of prediction and feedback from prediction error to reapportion the values of a set of cues to future outcomes or events. While statistical learning measures such as transitional probability focus on co-occurrence of stimuli, error-driven learning models *discriminate* between *cues* that occur temporally prior to later *outcomes*, which the cues serve to predict (Hoppe, van Rij, Hendriks, & Ramscar, 2020, Nixon, 2020, Ramscar et al., 2010). In learning, the degree to which any cue comes to predict an outcome over time depends on both occurrences and nonoccurrences of a relevant outcome given the cue in question. Importantly, because learning about an outcome is a function of the degree to which it is cued by prior learning, the amount of learning that occurs at any given time depends on the history (or experience) of the learner up to that point in time.

This last point is particularly relevant to the learning of serial patterns in language, where learning in experimental paradigms tends to be overshadowed by the knowledge that language users already have about serial patterns in the languages they speak (see, e.g., Siegelman, Bogaerts, Christiansen, & Frost, 2017). This problem is further confounded by the fact that serial patterns in language are not only predicted by preceding items in a sequence, but also higher-level information, such as semantics or morphology (Arnon & Ramscar, 2012).

Accordingly, to empirically investigate the mechanisms underlying the learning of serial patterns, it is often helpful to employ tasks that do not compete with or explicitly involve linguistic knowledge.

Since it is widely accepted that the learning mechanisms that are responsible for language learning are also responsible for the learning of serial patterns in a nonlinguistic domain (Saffran, Johnson, Aslin, & Newport, 1999), in the experiments we report below, participants were simply presented with sequences of single letters on a screen and had to press the corresponding key on a keyboard. The serial pattern of these sequences was structured following a probabilistic distribution, and we hypothesized that as participants implicitly learned these sequences—such that upcoming letters became more predictable given preceding letters—their key presses would become faster. This hypothesis was supported, replicating existing findings in the literature on learning of serial patterns (e.g., Cleeremans & McClelland, 1991) and typing (e.g., Bertram, Tønnessen, Strömqvist, Hyönä, & Niemi, 2015). Moreover, we found that changes in response latencies were better predicted by discriminative learning than by statistical learning.

In what follows, we first review previous findings in the literature on serial pattern learning. We then describe the two proposed learning algorithms in more detail, before presenting the experiments we conducted to test our hypothesis. We conclude by discussing the implications of our results for understanding the mechanisms that support the learning of serial patterns in language.

2. Background

2.1. *Serial pattern learning*

Serial pattern learning in nonlinguistic domains has traditionally provided a fruitful method for investigating the mechanisms and processes that support language learning. For example, Nissen and Bullemer (1987) used a task during which participants were presented asterisks in one of four positions, and had to respond by pressing a corresponding key, finding that distracting participants with a secondary task significantly increased key pressing latencies. Critically, Nissen and Bullemer (1987) discovered that as the experiment progressed, the speed of participants' key presses increased systematically, revealing a strong effect of learning during the experiment itself. Using the same task, Howard, Mutter, and Howard (1992) showed that serial patterns could be also learned by simple observation, even when an overt response was not produced.

Serial pattern tasks have also been used to investigate whether learners required some form of explicit instruction to acquire the underlying patterns —or whether they are capable of learning them implicitly, that is, without any instructions about underlying rules (see Ellis, 1994, for a review). For example, Reber (1967) asked participants to reproduce serial patterns comprising five different letters (e.g., A, B, C, D, E). Feedback was provided to indicate whether the patterns participants reproduced were correct or incorrect, but did not contain any information about the specific errors they had made. Participants were nevertheless able to learn the patterns, and even able to apply the “grammars” implicit in these patterns to new stimuli that had not been presented during the training phase. Additionally, Reber (1967) examined the effect of presenting participants patterns in two contrasting conditions: “grammatical,” that is, patterns that were based on a simple finite-state grammar that created a fixed set of patterns, and “nongrammatical,” that is, patterns that were random and did not match the finite-state grammar. Participants presented with grammatical sequences produced fewer errors than participants presented with nongrammatical sequences. Taken together, Reber’s (1967) results provided clear evidence that people are capable of learning the implicit serial patterns present in these tasks.

Cleeremans and McClelland (1991) further tested the learning of grammatical versus nongrammatical patterns using a flashing-asterisk task, similar to the one used by Nissen and Bullemer (1987) and Howard et al. (1992). Cleeremans and McClelland (1991) created sets of grammatical (i.e., predictable) patterns by following a finite-state grammar, while nongrammatical patterns were created by presenting random sequences. These grammatical and nongrammatical patterns were then concatenated into a sequences of 60,000 signals, which were used to train six participants (in a total of 20 training sessions!). To measure learning, Cleeremans and Mcclelland examined changes in response latencies, hypothesizing that response latencies would be shorter in the grammatical than in the nongrammatical condition. This prediction was supported by the results.

Cleeremans and McClelland (1991) also examined the degree to which participants were able to predict upcoming signals on the basis of preceding signals. They predicted that if participants had learned an underlying probabilistic serial pattern, they would respond more quickly to likely continuations of a pattern than to less likely continuations. In other words, the patterns of their response latencies should reflect the patterns of the probabilities of continuations at any point in a sequence. This hypothesis was tested by modeling learning in the task using a recurrent neural network. The network contained one input layer, one hidden layer, one output layer, and one memory layer, which learned the signals that preceded the presented signal. Weight adjustment was accomplished by the back-propagation of prediction error in training. Once the network had been trained on the grammatical and nongrammatical patterns, Cleeremans and Mcclelland took the activations of upcoming signal elements in the model to be a measure of their empirical predictability, and found that higher activations of the presented signal correlated with shorter response latencies of the participants. Since the activation in the network was also correlated with the distribution of conditional probabilities in the sequences, Cleeremans and McClelland (1991) suggested that participants were implicitly learning the serial patterns, and that it was this that allowed them to predict upcoming signals and respond more rapidly.

2.2. *Statistical learning*

The work of Saffran, Aslin, and Newport in the 1990s (Saffran et al., 1996a, Saffran, Newport, & Aslin, 1996b, Saffran et al., 1999) has had an enormous influence in the field of language acquisition and other fields of cognitive science (see Frost, Armstrong, & Christiansen, 2019, for a review). In a series of experiments, Saffran and colleagues passively exposed children and adults to continuous sequences of syllables. Some of these syllables always occurred in sequence (i.e., *a* was always followed by *b*); however, others had lower transitional probabilities (i.e., *a* was followed by *b* 50% of the time). Saffran et al. defined syllables that reliably co-occurred as "words" and lower-probability transitions as word boundaries. After being trained using a passive exposure paradigm, participants responses to different syllable sequences were tested. Test items were defined as "learned" words if they contained sequences that did not cross word boundaries, and as "new" words if they contained sequences that crossed word boundaries. Among the many findings from the passive exposure paradigm, Saffran and colleagues showed that babies were able to discriminate between learned words and new words, and that adult participants showed higher recognition accuracy for learned than new words. Based on these results, Saffran and colleagues argued that listeners learned to exploit the differences in the transitional probabilities between syllables, suggesting that dips in transitional probabilities indicate word boundaries. Listeners use this information—which they learn from exposure to language—to segment running speech into individual words.

At a time when many linguists and psychologists assumed that the bulk of people's linguistic knowledge must be innate because the statistics associated with language seemed to be far too complex to be learned (Miller & Chomsky, 1963), this finding had a major impact. It suggested that statistical learning might indeed provide a means by which language could be acquired from the input. Saffran et al. (1996a) continue to receive hundreds of citations each year and the experimental paradigm has a been tested with neuroscientific methods. It should be noted that most psychological theories of learning are in some sense statistical (see, e.g., Daw, Courville, & Dayan, 2008, Hull, 1943, Kruschke, 2006, Rescorla & Wagner, 1972, Thorndike, 1898). Nevertheless, Saffran et al.'s simple operationalization of statistical learning— that is, estimating the probabilities of items and their transitional probabilities using a finite order Markov chain model—has become the most frequently used in linguistic studies (see, e.g., Aylett & Turk, 2006, Cohen Priva, 2015). Thus, we term this specific formalization "statistical learning." In this regard, it is notable that despite the huge body of work following up on this landmark study, little attention has been devoted to understanding whether this operationalization of learning actually fully accounts for human learning behavior.

2.3. Discriminative learning

However, a large amount of evidence indicates that learning a cue's probability does not capture the full picture of the human learning mechanism. One of the key insights of learning theory is that cues compete for relevance in predicting outcomes. That is, if an outcome event is predicted by multiple cues, the contribution of each individual cue is not independent of the other cues in the event. Learning involves changes in the degree to which cues predict outcomes. Importantly, these changes in cue weighting are *shared* between all the cues available for predicting that event. This observation that learning involves cue competition is not captured by the conditional probability measure.

The importance of cue competition was recently demonstrated in a study of second-language speech acquisition. In an artificial language learning experiment, Nixon (2020) had participants learn to use two non-native speech cues (tone and nasality) as predictors of morphological outcomes (diminutive). One group of participants (the “blocking group”) learned one critical cue (e.g., tone) as a strong predictor during a pretraining phase; a second group (the control group) learned a control cue. In the following training phase, both groups had identical exposure to both the tone and the nasal cue. Finally, participants were tested on the second cue (nasality). Results showed significantly lower accuracy in the blocking group, compared to the control group. It is important to note that the conditional probability of the tested cue with the outcome was identical between conditions. But because the first cue (tone) became a strong predictor during the pretraining phase, there was little uncertainty left to drive learning. That is, learning of the second cue was “blocked” due to cue competition. The blocking effect was originally demonstrated in animal learning research (Kamin, 1968) and has had a significant effect on learning theory.

A second important consideration is that perception is a linear time-dependent process, which in turn defines the temporal structure of the learning process. Perceived cues are used to predict upcoming outcomes, not vice versa (Bröker & Ramscar, 2020, Hoppe et al., 2020, Nixon, 2018, Nixon, 2020, Nixon & Tomaschek, 2020, Nixon & Tomaschek, 2021, Ramscar et al., 2010). Combining these two aspects—cue competition and order effects—implies that the informativity of perceived cues about upcoming outcomes depends on more than just conditional probability. Learning occurs within a network of interrelated events. Furthermore, in many cases, cues will be completely disassociated from the outcomes they co-occurred with in training, and the predictive value of cues is rarely proportional to co-occurrence counts (Nixon, 2020, Ramscar, Dye, & Klein, 2013a).

The specific value of these changes depends on two aspects. First, the value of change is a function of what has already been learned, and decreases in proportion to the degree to which outcomes are predicted by prior learning. Second, the number of cues that are taken into account when predicting upcoming outcomes modulates the strength of the change - in the predictive strength of the individual cue (see, e.g., Eimas, 1969, Nixon, 2020, for learning in phonetics). Specifically, the more cues that are present during an event, the less predictive the individual cue is. The effect of prior experience is such that when outcomes are fully predicted, no learning occurs. With a few modifications, these empirical features of learning can also be described by the Bayes’ Theorem, which was originally proposed as a means for updating the probability of a hypothesis as new evidence comes to light (Daw et al., 2008, Kruschke, 2006).

These factors force cues to compete for predictive value as part of a fully connected system of cues and outcomes. This competition typically results in the formation of strong positive weights from cues that produce little or no error for a given outcome in training, and strong negative weights from cues that predict the nonoccurrence of an outcome. This means that in practice, rather than simply learning to associate cues with outcomes, this method of learning results in a network of link values that *discriminates* in favor of more reliable cues and against less reliable cues in predicting the experienced outcomes.

The discriminative learning process thus captures incremental, implicit learning, rather than explicit changes in perception or beliefs based on logic or reasoning (see, e.g., Nixon, Poelstra, & van Rij, 2022, Ramscar et al., 2013a). Cognitively, these aspects of learning can be taken to describe a *discriminative theory of learning* (Nixon & Tomaschek, 2020, Ramscar et al., 2010, Ramscar et al., 2013b, Ramscar, 2021), which can be formalized by training neural networks using virtually any computational learning algorithm that takes into account error during learning in the ways described above. In the present study, we use the error-driven learning rule proposed by Rescorla and Wagner (1972), where association strength V_i^{t+1} between a cue C_i and an outcome O at time $t + 1$ as calculated in Eq. 1:

$$V_i^{t+1} = V_i^t + \Delta V_i^t \tag{1}$$

The change in association strength between the cues and outcomes is weighted by the equations in 2.

$$\Delta V_i^t = \begin{cases} 0 & if\ Absent(C_i, t) \\ \alpha_i \beta_i (\lambda - \sum_{Present(C_j,t)} V_j) & if\ Present(O, t) \\ \alpha_i \beta_i (0 - \sum_{Present(C_j,t)} V_j) & if\ Absent(O, t) \end{cases} \tag{2}$$

Because models that employ the error-driven learning rules are easily interpreted, a number of recent psycholinguistic studies have made use of the learning equations proposed by Rescorla and Wagner (1972) (see also Widrow & Hoff, 1960; Rosenblatt, 1962) to formalize learning problems and derive predictions about participants' learning behavior. Various models, including the dual-path model (Chang, Dell, & Bock, 2006, Chang, 2002) and reinforcement learning (e.g., Harmon, Idemaru, & Kapatsinski, 2019), as well as the Rescorla–Wagner model, have also been used to successfully predict learning behavior (see Nixon & Tomaschek, 2023, and the contributions therein for recent developments in error-driven learning of language using various types of models). This work has included studies in domains such as: word learning (Nixon, 2020, Ramscar et al., 2010; Ramscar, Dye, Popick, & O'Donnell-McCarthy, 2011, Ramscar et al., 2013a), morphological structure learning in children (Ramscar et al., 2010, Ramscar et al., 2011, Ramscar et al., 2013b, Ramscar et al., 2013a) and adults (Arnon & Ramscar, 2012, Nixon, 2020, Ramscar, 2013, Vujović, Ramscar, & Wonnacott, 2021); morphological processing in adults (Baayen, Milin, Durdevic, Hendrix, & Marelli, 2011, Nieder, Tomaschek, Cohrs, & de Vijver, 2021, Seyfarth & Myslin, 2014, Tomaschek, Plag, Ernestus, & Baayen, 2019); lexical selection in production (Harmon & Kapatsinski, 2020); auditory comprehension and recognition (Arnold, Tomaschek, Sering, Lopez, & Baayen, 2017, Shafaei-Bajestan & Baayen, 2018); effects of the lexicon on articulation (Schmitz, Plag, Baer-Henney, & Stein, 2021, Stein & Plag, 2021, Tucker, Sims, & Baayen, 2019, Tomaschek et al., 2019, Tomaschek & Ramscar, 2022); phonetic learning (Olejarczuk, Kapatsinski, & Baayen, 2018); trial-by-trial changes in the neural signal that occur with learning (Lentz, Nixon, & Rij, 2021); speech sound acquisition in infants (Nixon & Tomaschek, 2020, Nixon & Tomaschek, 2021); and second language speech sound acquisition (Nixon, 2018, 2020).

2.4. *Window of integration*

Another important factor in characterizing serial pattern learning, in addition to the learning mechanism, is the distance from which participants learn to predict upcoming items. How many trials before the target trial are taken into account? Previous work has yielded somewhat divergent results. Cleeremans and McClelland (1991) demonstrated that response latencies to the target trial were well predicted by conditional probabilities calculated from a sequence of four trials before the target. Yet, the optimum predictability was obtained by taking into account just the two trials before the target trials. In another study, Jiménez, Paz, and Cleeremans (1996) demonstrated that using increasingly longer sequences (up to three trials) to predict the target yielded higher predictability of response latencies. In light of these different findings, the present study will also investigate the size of the window in which information is integrated when predicting upcoming outcomes in a serial pattern.

2.5. *The present study*

As we have argued above, the different ways of operationalizing learning —either by a finite order Markov chain model for statistical learning, or by means of an error-driven learning mechanism to operationalize discriminative learning—focus on different types of information present during learning. To investigate which learning mechanism—statistical learning or discriminative learning—better explains the learning of serial patterns, we conducted a series of empirical studies of typing, and examined which mechanism best accounted for the behavior observed empirically. Participants were presented with a single letter on the screen on each trial and had to press the equivalent key on a keyboard. In Experiment 1, the trial sequence was presented in a partly predetermined order, which was not made known to the participants. The order mimicked morphological patterns. In order to investigate whether learners also learn to predict upcoming events in the absence of a linguistic—and thus grammatical—structure, the order was rendered random in Experiment 2. By using a random serial pattern, we extend the work of Cleeremans and McClelland (1991), who used partially predetermined sequences. Following Cleeremans and Mcclelland's results in serial pattern learning, we hypothesize that more predictable letters would yield overall faster key presses. Moreover, given the discrepancy in results regarding how many trials are taken into account (Cleeremans & McClelland, 1991, Jiménez et al., 1996), we also tested how many trials before the target trial affect participant response latencies.

3. Experiment 1

3.1. *Participants*

The experiment was performed with ethics approval of the Psychology Department of the University of Tübingen, Germany. Sixteen female and five male participants (mean age 22.6 years [sd = 2.8]) were paid e10 for their participation. Before the experiment was conducted, they provided written informed consent. In addition, they completed a brief questionnaire about their handedness, native language, how many fingers they typically use for typing, and how frequently they use a keyboard.

3.2. Stimulus design

Natural languages are based on structurally predictable sequences. At the lowest level, typically referred to as phonetics and phonology, a set of articulatory gestures (or inferred gestures) form larger entities referred to as words. In morphologically complex languages, word bases and affixes are sequenced in a predictable structure. Finally, words form predictable sequences which can be subsumed as syntax. In Experiment 1, we created a pseudo-language that mirrored this type of predictive patterns in natural languages. While the sequences in our pseudo-language may technically refer to any level, we will use the terminology established in morphology to refer to the structurally predictive parts of our pseudo-language.

In the following paragraphs, we will describe the process that we used to construct four-letter sequences for our pseudo-language. For the sake of convenience, we will call these four-letter sequences "words." During the experiment, words were constructed using the letters <S, D, F, J, K, L>. In our description, for better clarity, we will use the letters <A, B, C, D, X, Y>. For each participant, we randomly selected two letters that served as pronouns (e.g., <A> and <D>) and randomly selected two additional letters that served as their respective suffixes (e.g., <X> and <Y>). In order to create words with the pronoun <A> and the suffix <X>, the remaining letters (<B, C, D>) served as the context of pronoun + suffix and were used to construct the "base" of the word. In order to obtain the two remaining letters for the four-gram word, a "bag of letters" which contained <B> with a probability of 0.43, <C> with a probability of 0.43, and <D> with a probability of 0.14. From this bag, we draw two letters (with replacement). In natural languages, a common pattern is for pronouns to occur at the beginning of a sequence, while suffixes occur at the end, such as in the English sentences "They walked" or "It finally started." However, constraining our words to the sequence <A__X> would produce only six different kinds of words (<ACBX, ADBX, ADCX, ABDX, ACDX, ABCX>). We also allowed the pronoun to occur at different places of the "word" (<_A_X>: <DACX, BACX, DABX, CABX, CADX, BADX><__AX>: <BCAX, DCAX, BDAX, DBAX, CDAX, CBAX>). This pattern was chosen in order to increase the number of potential words to a total of 18 for each pronoun-suffix combination. In the final set of stimuli, since we draw a letter twice from our "bag of letters" to obtain letters for the base, and since we created a total of 258 words, for the pronoun <A>, the probability of <B> was 0.81, <C> was 0.81, and <D> was 0.38. We repeated this process also for the pronoun-suffix combination <D> and <Y> producing another set of 18 words (<DBCY, DABY, DCBY, DCAY, DBAY, DACY, BDCY, ADCY, BDAY, CDBY, CDAY, ADBY, ABDY, ACDY, BCDY, BADY, CBDY, CADY>). Taken together, we created a total of 36 different words. Note that the letters we selected to serve as a pronoun for one word (e.g., <A>) also could serve as a base for the other word. This created ambiguity that is well known in natural languages and that was only resolved on encountering the suffix.

To approximate the distributional properties of pronouns and verbs in natural languages, words occurred with different frequencies during the experiment.

Fig. 1 demonstrates, for a random participant, the frequency distribution of the four-letter words in the pseudo-language. Since words were created for each participant anew, the frequency with which they occurred also varied. Across all participants, the least frequent item

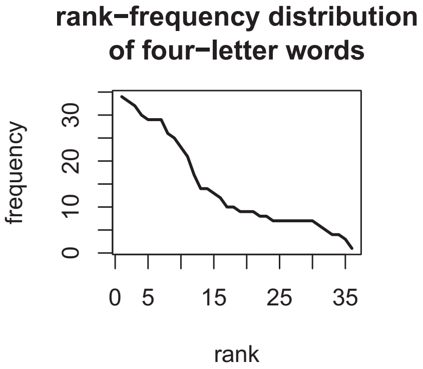

Fig. 1. Rank-frequency distribution of four-letter words in Experiment 1.

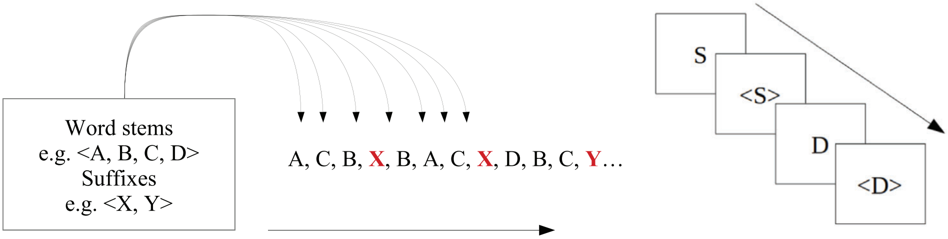

Fig. 2. Left: Creating a serial pattern for Experiment 1. Red letters represent the suffix letter, which occurred on every fourth trial. Right: Trial procedure. Letters in pointy brackets represent key presses on the keyboard.

occurred on average 3.3 times (sd = 1.2), the most frequent item occurred on average 40.1 times (sd = 4.7).

3.3. *Experimental set up*

To create the stimuli for the experiment, the order of the words was randomized and then words were concatenated to a long sequence of letters (Fig. 2, left). This means that the suffix letter occurred every four letters, whereas the pronoun letter varied in its location. This created a sequence of 2044 letters, one for each trial. The 2044 trials were divided into 10 roughly equal blocks.

The experiment was implemented in Python in a custom made script (available in the Supplementary Materials) and performed on a Lenovo laptop. On each trial, one letter of the sequence was displayed on the screen and participants had to press the corresponding key on the keyboard (Fig. 2, right). We will refer to the letter that participants had to respond to as the "presented letter"; letters that preceded this letter will be called "preceding letters." Participants were seated comfortably in a sound shielded room. They were given written instructions to type the letter displayed in the center of the screen as quickly and as accurately as possible. Participants were required to press the key [F] with the left index finger, [D]

Table 1
Windows of integration for measures used to predict key pressing latency

size of window	n-gram frequency	conditional probability	activation
1	F(−1, 0)	P(0|−1)	A(−1 →0)
2	F(−2,−1, 0)	P(0|−2, −1)	A(−2, −1 →0)
3	F(−3, −2, −1, 0)	P(0|−3, −2, −1)	A(−3, −2, −1 →0)

Note. "–3, –2, –1, 0" represents a series of four letters anywhere during the experiment, with "0" representing the presented letter. Activations are formulated in the form A(cues → outcome). Cues were trained in the form of bigrams.

with the left middle finger, [S] with the left ring finger; [J] with the right index finger, [K] with the right middle finger, and [L] with the left ring finger (from now on, displayed letters will be indicated by pointy brackets < >, keys on the key board by square brackets []). Each experiment began with five practice trials to get accustomed to the set up. Practice trials were taken into account in the modeling of the different learning mechanisms, but excluded from the analysis of response latencies. No feedback about performance was given.

3.4. Calculation of learning measures

In addition to investigating which learning mechanism accounts best for the participant's key pressing latencies, we also investigated the size of the integration window that is taken into account in predicting the presented letter. This was accomplished by calculating learning measures in differently sized sliding windows that integrated the cues of the serial pattern presented to the participants. We restricted ourselves to a maximal sequence of four letters. For illustrative purposes, we use the numbers "–3, –2, –1, 0" as a placeholder for a sequence of four letters at any point during the experiment, where position "0" represents the presented letter, that is, the letter to which participants had to respond by pressing the equivalent key; "–1" represents the letter directly preceding the key press; "–2" represents the second preceding letter preceding the key press, and so on. Any of the positions "–3, –2, –1, 0" could be filled with the letters <S, D, F, J, K, L>.

In the following, we will describe how we calculated the statistical measures and discriminative measures used to predict key press latencies. We calculated two types of statistical measures. The first was the frequency of occurrence of bigrams, trigrams, and four-grams in the serial pattern (column "frequency" in Table 1). This measure was obtained by counting at each trial how often an n-gram had already occurred during the experiment. This measure yields thus a *trial-dependent frequency* of bigrams, trigrams, and so on. The second was the conditional probability of the presented letter "0" given the preceding letter, the two preceding letters, and the three preceding letters (column "conditional probability" in Table 1). Conditional probability was obtained by dividing the trial-dependent frequency of the whole n-gram (i.e., the preceding letters plus the target; the size of which depended on the integration window) by the trial-dependent frequency of the preceding letters.

To obtain discriminative measures, we trained a two-layer neural network to predict presented letters on the basis of the cues (the structure of the cues will be discussed below).

Training, and thus the estimation of connection weights between cues and outcomes, was accomplished using the Rescorla–Wagner learning function provided in the NDL package (Arppe et al., 2018) implemented in R (Version 3.5.3). The function was customized to the needs of the present study such that it provided a weight matrix in each learning step (initialized with zeros by default). This provided us with cue-outcome weights calculated for each trial, that is, trial-dependent weights, which allowed predictions about upcoming letters to be made on the basis of the learning history. Recall that a new sequence of letters was created for each participant. Accordingly, we trained an individual network for each participant according to the trial order they received.

The cue structure contained bigrams of the letters preceding the target letter, which served as the outcome of the learning trial. In neural network terminology, preceding letters at each trial represented the input to the network, the output of the network was the target letter to be typed at each trial. For example, when the sequence was <CD> with <C> preceding the presented letter <D>, then the cues were {#C, C#}, with the hashtag # representing the edge of the provided cue structure. Clearly, this structure did not have any cue competition and network weights between cues and outcomes simply mirrored the co-occurrence frequency of letters. Cue competition arises when a larger window of integration is taken into account, for example, in the sequence <BCD>. In this sequence, {#B, BC, C#} constituted the cues. Since {#B, C#} could also occur in other sequences, for example, <BAD> or <ACD>, these cues competed for informativity about the outcome <D>. Likewise, the network was trained on the four-letter sequence <ABCD>. During training, weights between the cues and the outcomes were adjusted on each trial using the Rescorla–Wagner learning equation. The strength with which the preceding cues predicted the target letter in each trial was operationalized by means of “activation,” which was computed by summing trial-dependent weights between the cues (= preceding letters) and the outcome (= presented letter). Column “activation” in Table 1 illustrates the activation in all three windows of integration.

3.5. Development of learning measures

The way that the different learning measures develop across the experiment is visualized in Fig. 3. The plots show the development of each learning measure for one randomly selected participant, calculated for the largest window of integration “–2, –1, 0.” The x-axis represents the trial number, the y-axis represents the strength of the measure. The left column illustrates the measure when the letter sequence was <FDK>.

The top row in the left column illustrates how the frequency of the n-gram <FDK> steadily increases across the experiment. The second row demonstrates that conditional probability P(K | F, D) jumps quickly to a relatively high value and then remains steady across the experiment. Note that the conditional probability plot does not visualize all changes in conditional probability. Specifically, some decreases in conditional probability are not visible in the plot, as the plot only shows the conditional probability measure on trials in which the target letter K occurred. The third row shows activation. Relative to conditional probability, activation A(F,D → K) has a slow initial increase, followed by a steady phase, interspersed with decreases

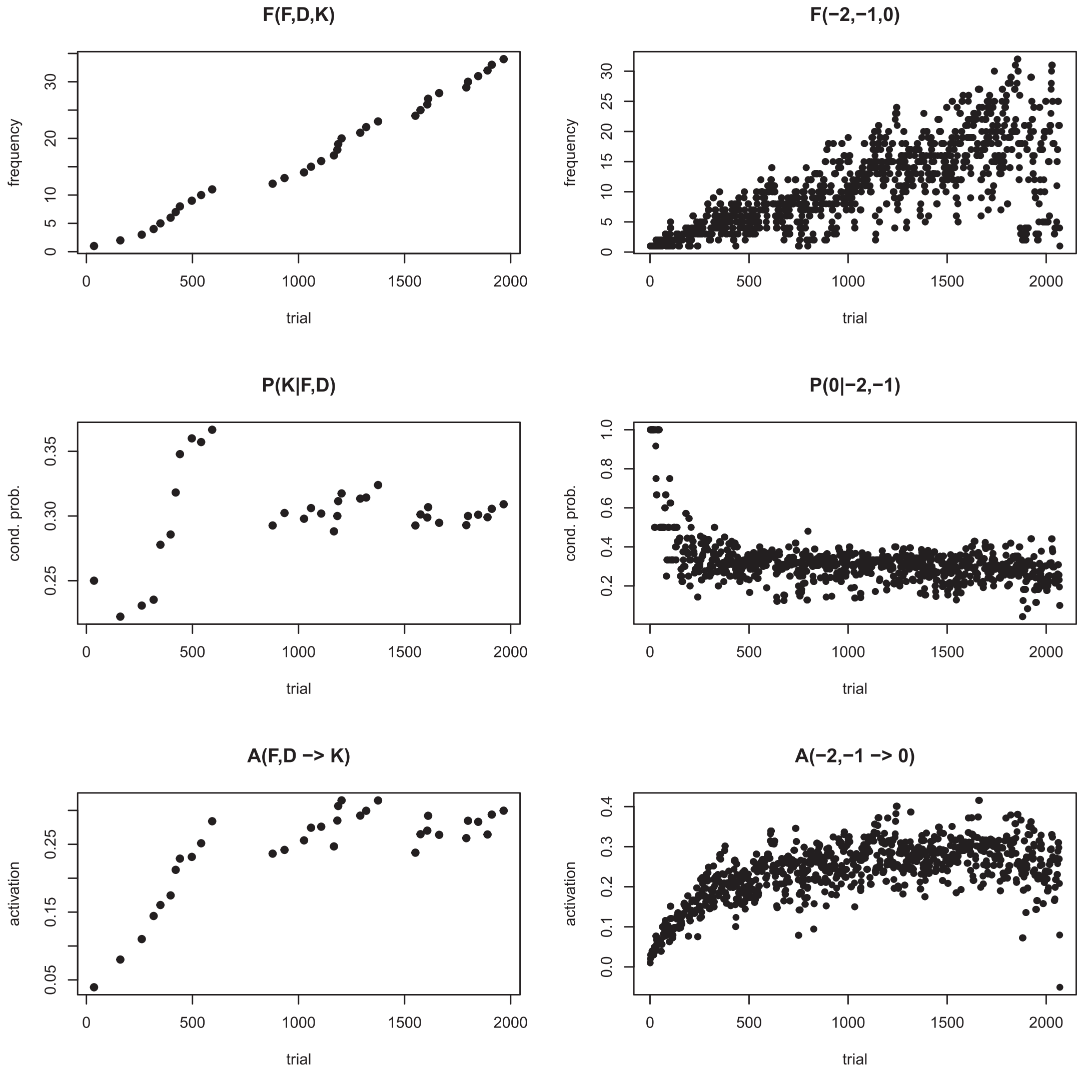

Fig. 3. Left column: Development of learning measures for the sequence "FDK" for one participant. The x-axis represents the trial number, the y-axis represents the strength of the measure. Right column: Development of learning measures across all trials in all contexts.

whenever the bigram cues are followed by letters other than <K>. The activation plot is similar to the conditional probability plot in that it does not show decreases in activation when the cues occur but the target does not occur. The plot only shows the activation on trials in which the target letter K occurred. The right column illustrates the trial-by-trial development of the learning measures independently of the trigram sequence. The three figures illustrate that the participant is faced with ongoing variability across the experiment. While there is a seemingly steady rise in frequency, there are large jumps depending on the frequency

Table 2
Correlations among learning measures and letter frequency

	letter frequency	F(−1,0)	P(0|−1)
F(−1,0)	0.88		
P(0|−1)	−0.09	0.01	
A(−1 →0)	0.59	0.67	0.22
	letter frequency	F(−1,0)	P(0|−2, −1)
F(−2, −1,0)	0.66		
P(0|−2, −1)	−0.28	0.10	
A(−2, −1 →0)	0.55	0.78	0.15
	letter frequency	F(−3, −2, −1,0)	P(0|−3, −2, −1)
F(−3, −2, −1,0)	0.36		
P(0|−3, −2, −1)	−0.38	0.39	
A(−3, −2, −1 →0)	0.4	0.81	0.35

of the individual trigrams (n-gram frequency, top row). Even though conditional probability P(0|–2,–1) settles to roughly 0.3, it continues to vary (second row). Variability is also observable in activation A(–2,–1 → 0), which mirrors the classic learning curve: a steep slope at the beginning of the experiment and an attenuation toward the end of the experiment (third row).

The development of activation and conditional probability, as depicted in the second and bottom rows of Fig. 3, suggests that there might be a positive correlation between n-gram frequency and activation, and a negative correlation between n-gram frequency, activation, and conditional probability. We tested this by performing pair-wise Pearson's product-moment correlations among the measures in all windows of integration. The findings of this correlation analysis are illustrated in Table 2. Unsurprisingly, letter frequency has a positive correlation with n-gram frequencies F(–1,0), F(–2,–1,0) and F(–3,–2,–1,0), whose strength decreases with the size of the integration window. This is also the case when letter frequency is correlated with activation. By contrast, letter frequency has a negative correlation with conditional probabilities, which increases with the size of the integration window. N-gram frequency has a negligible correlation with conditional probability in the smallest window of integration, which increases to a medium correlation in the largest window of integration. N-gram frequency has a relatively strong correlation with activation across all windows of integration. The degree of correlation could differ if either the connection weights of the learning model or the assumed prior frequency of the N-grams was not initially set to 0 to represent prior learning (Harmon & Kapatsinski, 2020). Finally, conditional probability has a small to medium correlation with activation across all windows of integration. In conclusion, these correlations show that our learning measures are related to each other, which is why we expect them to yield similar predictions about key press latency.

3.6. Data processing

Participants pressed the wrong key in 6% of the cases. These responses were excluded from the analysis. In addition, latencies shorter than 250 ms and longer than 2500 ms were also excluded, resulting in a total loss of 6.5% of the collected data. Key press latencies were log-transformed to obtain normally distributed data. Frequencies and conditional probabilities were squared to obtain normal distribution. Activation was normally distributed and did not require any transformation.

3.7. Analysis

To test what learning measure best predicted key press latencies during the experiment, we fitted multiple Generalized Additive Mixed-Effects Models (GAM, Version 1.8-31 Hastie & Tibshirani, 1986; Wood, 2006), implemented in the R package to the log-transformed key press latencies (from now on *latencies*), with participant and letter identity as random effects. GAMs fit nonlinear relations between dependent variables and predictors.[1]

In order to operationalize simple kinematic learning during the experiment, we calculated a letter's frequency of occurrence during the experiment (*letter frequency*). We fitted a baseline model that tested how latencies changed as a function of letter frequency. The model (and all following models) included random intercepts for participants and letter identity. The following model illustrates the model structure:

$$g(\mu_1) = \beta 0 + f_1(\textit{letter frequency}) + \epsilon \tag{3}$$

To examine how latencies were further modulated by the degree to which the presented letter was predicted by the preceding context, we added a learning measure in interaction with letter frequency to the model using tensor product smooths. We fitted one model for each combination of learning measure type by integration window. The following code illustrates the model structure testing the interaction:

$$g(\mu_2) = \beta 0 + f_1(\textit{letter frequency}, \textit{learning measure}) + \epsilon \tag{4}$$

There is evidence that key press latencies in typing tasks are systematically slower at morphological boundaries (Bertram et al., 2015; Weingarten, Nottbusch, & Will, 2004)— as they are actually in an environment of greater uncertainty. To test this effect in the present study, we also fitted a model in which we fitted latencies with a factorial predictor (*morphological boundary*) that specified whether the key press was in response to the suffix letter or in response to other letters.[2]

Finally, we tested whether a further series of control variables were predictive of key pressing latencies. These control variables were age, sex, how many fingers participants typically use for typing, and how often they use the keyboard. However, none of these predictors were significant.

Model comparisons and visualization were carried out using helper functions from the `itsadug` package (van Rij, Wieling, Baayen, & van Rijn, 2015). The data and analysis scripts for Experiment 1 and all following experiments can be found in the Supplementary Materials available on osf.[3]

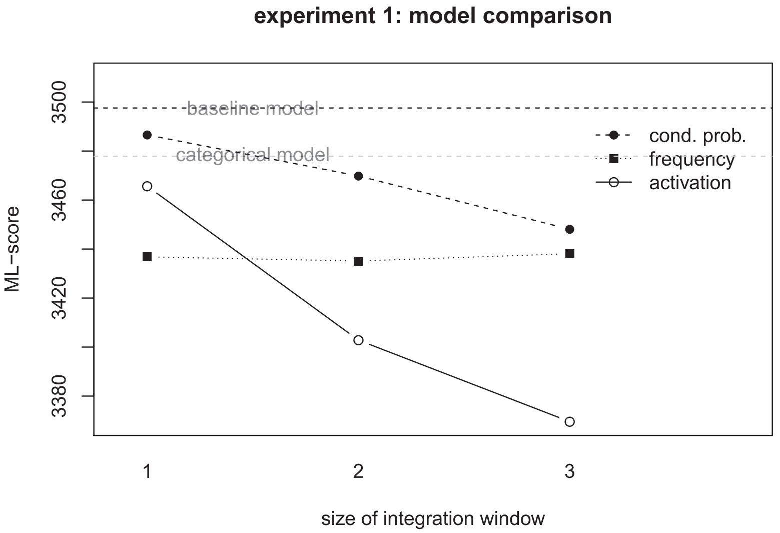

Fig. 4. Model comparison in Experiment 1 (morphology-like structure). Y-axis represents ML-scores for the different GAM models (lower ML-scores indicate better model fit); x-axis represents the size of the integration windows as shown in Table 1. Line type represents the different types of learning measures. The dashed horizontal lines represent the ML-score for the baseline model and categorical model.

3.8. *Results*

Apart from the baseline model, all models had 8 degrees of freedom, so the model comparison can be conducted purely on the basis of the models' maximum likelihood values (ML-score) when fitting different learning measures to key pressing latencies.[4] A lower ML-score thus indicates a better goodness of fit. Fig. 4 illustrates the ML-scores (y-axis) as a function of the size of the integration window (x-axis) and the type of learning measure (line type). As can be seen by means of the dashed horizontal lines at the top of the plot, the baseline model had a systematically higher ML-score than all models that either contained learning measures or the factorial predictor *morphological boundary* (i.e., the categorical model). In other words, the inclusion of either the factorial predictor or any of the learning measures significantly improved model fit.

However, the categorical model (gray horizontal line) outperformed only one learning measure, namely, conditional probability in the smallest integration window (–1, 0), while all other learning measures provided a better model fit than the categorical predictor. This indicates that key pressing latencies were codetermined by gradient, rather than categorical, representations of the structure in the serial pattern.

We now turn our attention to the way that the learning measures interacted with the size of the integration window. In the case of conditional probability, although goodness of fit improved when larger windows of integration were taken into account, conditional probability produced a poorer model fit to the data than both n-gram frequency and activation in all integration windows. Activation also produced a poorer model fit than n-gram

Table 3
Summary tables of nonlinear effects for n-gram frequency, conditional probability, and activations in Experiment 1

n-gram frequency	edf	Ref.df	*F*	*p*-value
te(Freq.Stimuli, F(−1, 0))	6.91	7.85	412.88	< .001
te(Freq.Stimuli, F(−2, −1, 0))	11.14	13.03	250.36	< .001
te(Freq.Stimuli, F(−3, −2, −1,0))	9.27	10.85	298.76	< .001
cond. prob.	edf	Ref.df	*F*	*p*-value
te(Freq.Stimuli, P(0|−1)	15.42	16.92	187.74	< .001
te(Freq.Stimuli, P(0|−2, −1)	14.74	17.05	187.69	< .001
te(Freq.Stimuli, P(0|−3, −2, −1)	9.71	11.04	293.31	< .001
activation	edf	Ref.df	*F*	*p*-value
te(letter frequency, A(−1 → 0))	11.63	14.03	227.76	< .001
te(letter frequency, A(−2, −1 → 0))	12.99	15.49	215.27	< .001
te(letter frequency, A(−3, −2, −1 → 0))	14.20	16.59	205.95	< .001

Note. Full summaries can be found in the Supplementary Materials.*Note*.

frequency in the smallest integration window (1). However, as the size of the integration window increased (2 and 3), activation provided an increasingly better model fit than n-gram frequency. Note that the goodness of fit for n-gram frequency stayed the same for integration windows (1) and (2) and even decreased slightly for the largest integration window (3).

These results indicate that activation performs better than conditional probability and n-gram frequency for larger integration windows. It is, however, possible that this finding could result not from the measures themselves, but from their correlation with letter frequency. To rule out this possibility, we conducted a correlation test. N-gram frequencies have a correlation of $r = .72$, activations $r = .55$, and conditional probabilities $r = –.24$ in the mid-sized window of integration. Thus, n-gram frequencies are the most highly correlated with letter frequency, but they are neither the best- nor the poorest-performing predictor. This leads us to the conclusion that activations yield a better model fit because they yield a better prediction of response latencies, not because of differences in correlation with letter frequency.

The model summaries are shown in Table 3 and the tensor product smooths are illustrated in Fig. 5, in which key press latency is color coded. In Fig. 5, yellow indicates long latencies, blue indicates short latencies. Black points represent the raw data points. Estimated key press latencies were back-transformed to seconds for illustration. Independently of the learning measure, we find that latencies to presented letters were slow at the beginning of the experiment (roughly 0.85 s) but decreased as letter frequency increased (roughly 0.6 s).

Moreover, the pattern of this variation is nonlinear, and it changes as a function of the interaction between letter frequency (x-axis) and the learning measures (y-axis) computed for different sizes of integration windows. Overall, the more frequent the n-gram sequence was (top row), the higher the conditional probability (mid row) or the higher the activation (bottom row), the faster the key press. This means that all of the learning measures are capturing learning to some extent. Beyond this general pattern, there are systematic differences between the measures and windows of integration. For the smallest window of integration (P(0|-1)), most of the variance comes from letter frequency, with relatively little effect of conditional probability. Furthermore, changes in latencies as a function of letter frequency and activation have a stronger nonlinear effect for measures obtained in larger integration windows than for the short integration window.

3.9. Preliminary discussion

In Experiment 1, we investigated which learning mechanism—statistical learning or discriminative learning—underlies learning of serial patterns of letters (a task that we took to be analogous to learning morphological structures). Two types of statistical learning measures were tested, n-gram frequency and conditional probability. We demonstrated that measures of discriminative learning using a long integration window outperform both statistical learning measures in integration windows of at least two letters. This finding replicates the finding by Cleeremans and McClelland (1991) who showed that predictability of upcoming sequences affects response times in a serial learning task. It also extends this previous work by distinguishing which of several candidate learning mechanisms were likely to underlie this learning: the modeling results indicate that participants used cue competition to learn the serial patterns hidden in the sequence of letters, and that they used this information to predict the upcoming letter.

As with the sequential stimuli used in Cleeremans and Mcclelland's (1991) experiment, the sequential stimuli employed Experiment 1 contained implicit serial patterns that resembled grammatical structures. However, the variance in sequential stimuli can be random, in which case there may be no underlying structure to be learned. The question, therefore, arises whether such grammatical structures are necessary for learning to occur—especially given that random, or at least apparently random, sequences are ubiquitous in nature. For example, consider a learner observing the flight of a bird between treetops, followed by a beetle making its way across a rock, followed by a distant hum of traffic. There may not be an underlying structure that dictates that these three events should occur in this rather than a different temporal order. Another example that may be relevant to readers of this paper is laboratory experiments in linguistics and psychology, in which a set of stimuli is randomized. In particular, consider a situation in which there is repetition in the experiments, such as a set of stimuli or a set of conditions which are repeated throughout the experiment, but the specific order of which is randomized. If the stimuli or conditions are repeated throughout the experiment, then certain sets of stimuli or conditions will co-occur more than others, due to the random order.

In other words, even in a random sequence, certain items may become predictive cues for certain following items. Experiment 2 addresses the question of whether participants learn to predict upcoming events, even in a sequence that has no predetermined structure.

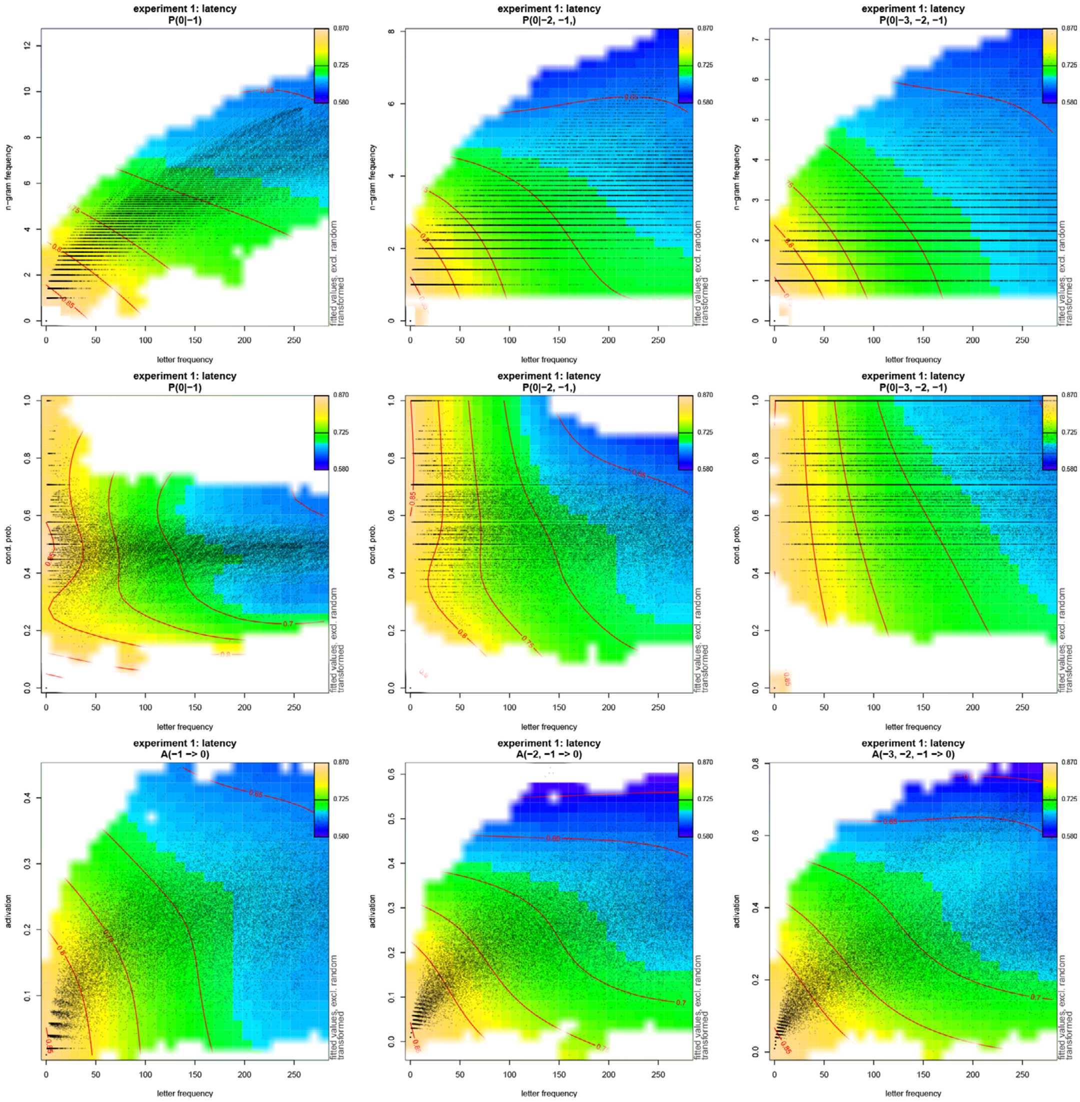

Fig. 5. Estimated key press latency in **Experiment 1** (morphology-like structure) as a function of the interaction between letter frequency (x-axis) and learning measure (y-axis). Top row: n-gram frequency (square-root transformed); mid row: conditional probability; bottom row: activation. Columns represent different sizes of integration windows. Color represents the key press latency: Yellow indicates long latencies, blue indicates short latencies. Black dots represent raw data points. Estimates were back-transformed to seconds for illustration. Estimates exclude random effects.

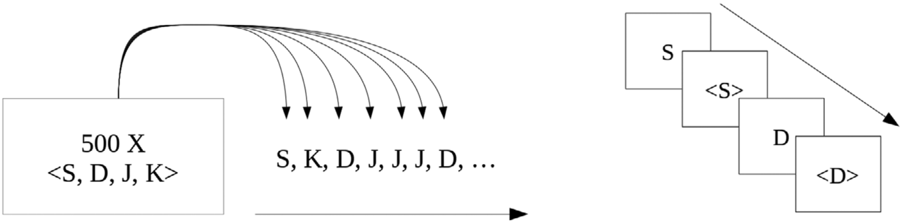

Fig. 6. Left: Creating a serial pattern for Experiment 2. Right: Trial procedure. Letters in pointy brackets represent key presses on the keyboard.

4. Experiment 2

4.1. Participants

Eight female and three male participants were paid e10 for their participation (mean age 23.5 years; sd = 2.4). Before the experiment was conducted, they provided informed consent in written form. In addition, they were asked to provide information about their handedness, native language, how many fingers they typically use for typing, and how frequently they use a keyboard.

4.2. Stimuli and experimental set up

To create the serial pattern for Experiment 2, we generated a random sequence of 2000 letters in which each of the four letters <S, D, J, K> were presented 500 times in total. For each participant, a new serial pattern was created. Fig. 6 (left) illustrates the procedure. The experimental set up in Experiment 2 was the same as in Experiment 1 (cf. Fig. 6, right).

Repetition of items in a sequence leads to a statistical distribution of item co-occurrences, even when the order is random. A straightforward way to visualize this distribution is to inspect n-gram frequencies. This is demonstrated in Fig. 7, which shows the rank-frequency distributions of bigrams and trigrams in Experiment 2 for one randomly selected participant. Clearly, there are some bigrams and trigrams that occur more often than others. In the case of trigrams, the most frequent trigram is twice as frequent as the least frequent trigram. We thus expect participants to learn some sequences better than others, which will be reflected in differences in response latencies similar to Experiment 1. However, as in Experiment 1, differences in learning may not necessarily depend directly on the n-gram frequencies, per se. We also tested whether learning occurred as a function of conditional probability, or due to the discriminative structure in the presentation sequence.

4.3. Analysis

The analysis of Experiment 2 followed the same procedure as Experiment 1. First, the preprocessing of the experimental data was performed in the same way as in Experiment 1. Participants pressed the wrong key in 7.5% of the cases. These responses were excluded from the analysis. In addition, latencies shorter than 250 ms and longer than 2500 ms were also

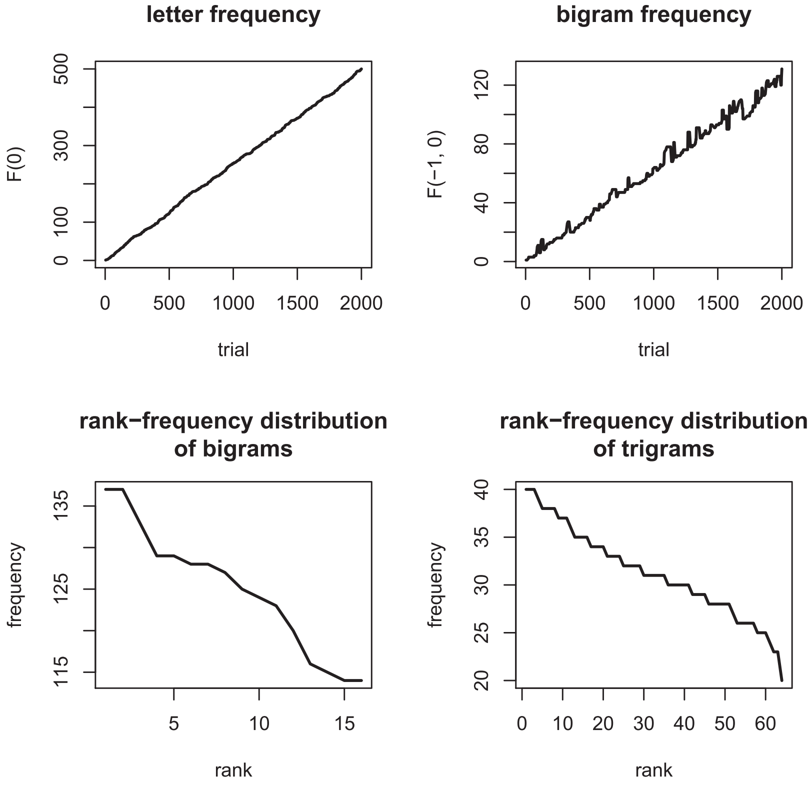

Fig. 7. Rank-frequency distribution of bigrams and trigrams at the end of Experiment 2.

excluded. In total, 9.2% of the collected data were excluded from the analysis. Second, the learning measures for serial patterns in Experiment 2 were calculated in the same way as in Experiment 1 (cf. Table 1). Finally, the statistical analysis in Experiment 2 was performed in the same way as in Experiment 1. Log-transformed key pressing latencies were fitted with an interaction between frequency of letters and one learning measure. Using ML-scores, the resulting nine models were compared to each other and to a baseline model in which latencies were fitted only with frequency of letters.

4.4. Results

Fig. 8 shows the goodness of model fit for the baseline model and each of the nine learning models. The figure illustrates that the goodness of model fit worsened as the size of the integration window increased (reflected in higher ML-scores). This is in direct opposition to the findings in Experiment 1 (cf. Fig. 4), where measures from a larger windows of integration provided a better goodness of fit (except for n-gram frequency, where it had almost no effect). This result is interesting, given the difference in cue informativity between the two experiments. Recall that the probabilistic structure in Experiment 1 was such that within a sequence of three letters there was one cue (the “pronoun”) that was often highly informative about

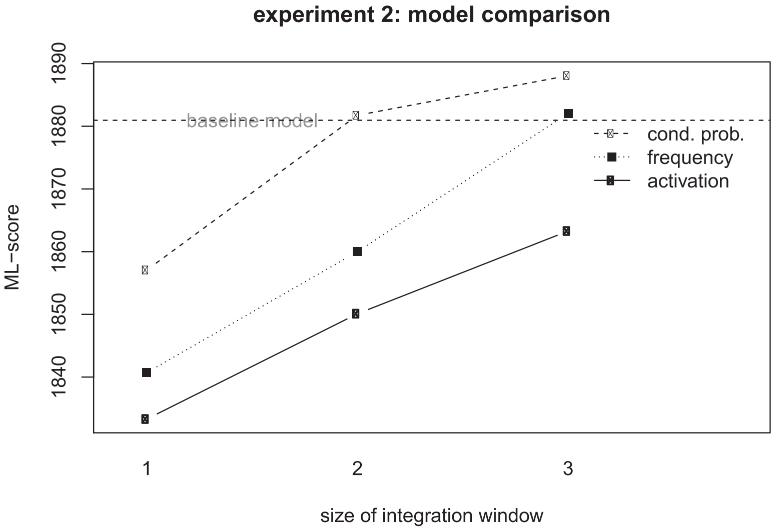

Fig. 8. Model comparison in Experiment 2 (random sequence). Y-axis represents ML-scores for the different GAMM models (lower ML-scores indicate better model fit); x-axis represents the size of the integration window as shown in Table 1. Line type represents the different types of learning measures. The dashed horizontal line represents the ML-score for the baseline model.

the fourth upcoming letter. In contrast, Experiment 2 used a random sequence, in which the informativity of cues is much lower in general. The finding that the integration window differed between experiments demonstrates that the degree of structure (and hence predictability) in the data affects the distance at which participants make predictions about upcoming trials.

Turning our attention to the type of learning measure, we see that conditional probabilities yielded a significantly poorer model fit than the other models across all sizes of integration, while activation yielded the best model fit. This indicates that learning was not based on simple counts of the letter sequences (n-grams) or the derived conditional probabilities. It was the error-driven informativity about the upcoming presented letter that drove learning even in a random serial pattern.

Fig. 9 illustrates the effect of activation obtained in all three integration windows in Experiment 2. (For the sake of space, we focus only on activation in Experiment 2 as we have established in Experiment 1 that the other two learning measures yield comparable changes in typing latencies.) The summary tables in Table 4 show that all interactions as fitted with the tensor product smooth were significantly nonlinear.

In the best model, the effect of A(–1 → 0), we observe that latencies decrease as the frequency of the presented letter increases. Yet, this effect is modulated by activation: there was a negative correlation between latency and activation. The more expected the letter, the faster the response.

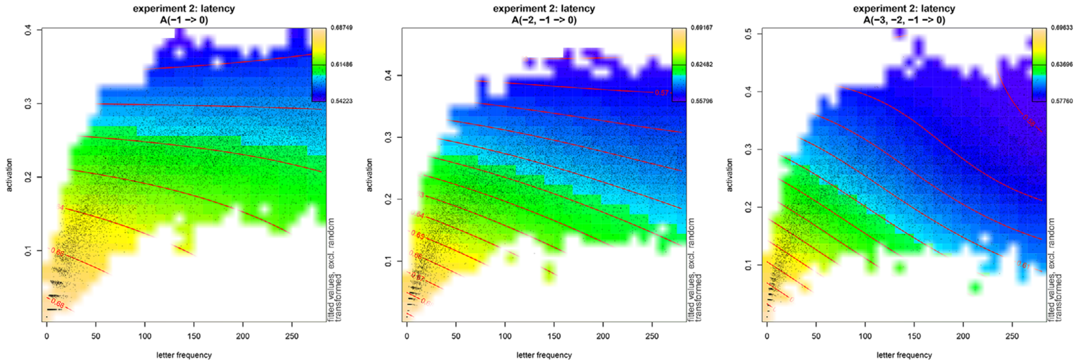

Fig. 9. Estimated key press latency in **Experiment 2** (random pattern) as a function of letter frequency (x-axis) and the activation measures (y-axis). Columns represent different sizes of integration windows. Color represents latency: Yellow indicates long latencies, blue indicates short latencies. Black dots represent raw data points. Estimates were back-transformed to seconds for illustration. Estimates exclude random effects.

Table 4
Summary tables of nonlinear effects for Experiment 2

	edf	Ref.df	*F*	*p*-value
te(letter frequency, A(−1 → 0))	4.04	4.43	104.51	.00
te(letter frequency, A(−2, −1 → 0))	4.47	5.12	83.18	.00
te(letter frequency, A(−3, −2, −1 → 0))	5.54	6.35	63.25	.00

Note. Full summary tables are available in the Supplementary Materials.

4.5. Discussion

In Experiment 2, participants learned to predict upcoming presented letters even in a random sequence, in the absence of morphology-like probabilistic structure. In terms of the learning mechanism, results showed that activation provided a better fit than the statistical measures, consistent with Experiment 1. This suggests that in sequence learning, humans learn from prediction and prediction error, whether the sequence is structured or random. Interestingly, though, the amount of structure in the sequence did affect the size of the integration window. The increased structure in Experiment 1 led to a wider window of integration, compared to the random sequence in Experiment 2, demonstrating that increased predictability leads to increased learnability. That is, the greater differentiation between informative and uninformative cues in the cue-outcome structure means that the learner is better able to *discriminate* cues that are relevant for predicting upcoming outcomes from irrelevant cues. Overall, increased predictability leads to lower uncertainty about upcoming outcome events.

Even though activation obtained from a larger window of integration window provided a poorer model fit than the single trial window, activation was still negatively correlated with response latency in the larger windows. There are two possible explanations for this finding. Either participants learned to use cues from larger integration windows, even though these cues were not as good predictors as the directly preceding one. Or that the negative correlation between activation calculated for cues in the larger integration window is simply due to the predictive power of the cues closest to the target window, while the distant cues contribute less or even nothing to predict the upcoming letter. We leave this question to future research.

5. Experiment 3

Recall that Saffran and colleagues (Saffran et al., 1996a; Saffran et al., 1996b; Saffran et al., 1999) found that listeners learned to group syllables together into clusters that had relatively higher co-occurrence frequency when listening to a stream of syllables. We were interested in whether this kind of process also occurred in Experiment 2 such that certain sequences led to representations of larger clusters in the form of bigrams or trigrams. If this were the case, typing latencies of n-grams should be predicted by their frequency, co-occurrence probability, or activation. We tested this hypothesis in Experiment 3. Specifically, we used the by-participant trial sequences of Experiment 2 to generate n-gram frequency, co-occurrence probability, and activation measures for all possible bigram and trigram combinations. We then tested these measures against the participants' bigram and trigram typing speed.

5.1. Method

Participants from Experiment 2 took part in Experiment 3 directly afterward. We created 16 bigrams and 64 trigrams on the basis of the letters <S, D, J, K> by combining different letters (e.g., <DJ>, <SKJ>) but also the same letter (e.g., <DD>, <JJJ>). Each bigram and each trigram was presented five times in random order, totaling in 400 trials. With 11 participants, a total of 4400 data points were recorded.

5.2. Variables and analysis

We tested three dependent variables: (1) the time between the presentation of the n-gram on the monitor and the first key press; (2) the time between the first key press and the second key press; and (3) for trigrams, the time between the second key press and the third one. For the analysis, we excluded 5.6% of data points for which the stimuli were incorrectly typed. In the data set with only correct responses, we excluded data points for which dependent variables were larger than 2.5 standard deviations away from the mean (for each of the dependent variables 2.1%, 2.3%, and 1.6%, respectively).

To calculate the activation measure, we trained two additional networks on the trial sequence presented to the participants. The first network was trained to discriminate bigrams on the basis of their bigram cues (e.g., the cues {#C, CD, D#} for the bigram <CD>), the second network was trained to discriminate trigrams on the basis of their bigram cues (e.g., the cues {#B, BC, CD, D#} for the trigram <BCD>). To investigate whether discriminative learning predicts key press latencies, we calculated the *activation* for each n-gram by summing the weights between their cues and the associated outcome. Higher activation was assumed to represent stronger support of an outcome, which we expect to be correlated with shorter latencies.

Table 5
Summary tables for the first, second, and third key press in Experiment 3

first key press	Estimate	Std. Error	*t* value
(Intercept)	−0.21	0.07	−3.22
Repetition	−0.01	< 0.001	−3.08
Changing hand = two hands	0.18	0.03	5.10
Word length = trigrams	0.04	0.04	1.09
Frequency	−0.00	< 0.001	−0.34
Activation	−0.00	< 0.001	−0.34
second key press	**Estimate**	**Std. Error**	***t* value**
(Intercept)	−1.67	0.09	−18.82
Repetition	−0.03	0.01	−5.22
Changing hand = two hands	0.07	0.02	3.21
Word length = trigrams	0.14	0.03	5.16
Frequency	0.01	0.01	1.15
Activation	0.01	0.01	1.08
third key press	**Estimate**	**Std. Error**	***t* value**
(Intercept)	−1.60	0.08	−19.98
Repetition	−0.03	0.01	−5.61
Changing hand = two hands	0.06	0.03	2.16
Frequency	0.01	0.01	2.28
Activation	< 0.001	0.01	0.32

Frequency and activation each showed a bimodal distribution with one peak representing bigrams the other trigrams. To allow for a simultaneous analysis of bigrams and trigrams, frequency and activation were z-normalized depending on n-gram length. After z-normalization, predictors formed a monomodal distribution.

In addition to the predictors of interest, we controlled for the following conditions. *Repetition*: how often participants had already typed the n-gram within Experiment 3; *Changing hand*: whether the n-gram was typed with one hand or two hands (e.g., <DF> vs. <DJ>); *condition*, that is, bigrams or trigrams. We used participants and n-gram identity as random effects. As we did not expect any nonlinear effects, we performed the analysis using linear mixed-effect regression (Bates & Sarkar, 2005).

5.3. Results

Before discussing our predictors of interest, we briefly report the effects of the control variables illustrated in Table 5. We find for all three dependent variables that faster key presses were associated with increased repetition and when typed with only one hand. The first key press did not significantly differ between the bigrams and trigrams. The second key press was significantly slower for trigrams. We also tested interactions between *Word length* (bigrams vs. trigrams) and the learning measure (frequency, activation) but they were not significant. Table 5 also reports the estimated coefficients for frequency and activation when added to the baseline model. Counter to our expectations, neither frequency nor activation turned out to significantly predict the latencies of the first or second key press. Surprisingly, the third key press when typing trigrams was significantly slower with higher frequency.

5.4. Discussion

The results of Experiment 3 indicate that, unlike in the work of Saffran and colleagues, our participants did not learn to form sequences into longer clusters. This is likely due to aspects of the present task that systematically differed from that of Saffran et al. In their task, auditory syllables formed an uninterrupted auditory stream. In our experiment, sensory input was interspersed with motor output and vice versa. The cues and outcomes experienced by the participants had a systematically different structure between Experiments 2 and 3. In Experiment 2, a single keystroke motor outcome followed presentation of each single visual cue. In Experiment 3, typing of n-grams required coordination of multiple movements, which had not been learned in Experiment 2. Perhaps if participants were trained with longer random sequences that were presented simultaneously on the screen, they would have learned to “chunk” these sequences. This hypothesis remains to be tested.

6. General discussion

The present study sheds new light on the mechanisms of serial pattern learning. Three experiments investigated the learning mechanisms, the effect of predictive structure, the distance at which learning occurs between items, and whether item sequences led to “chunking.” Three hypothesized operationalizations of learning mechanisms were tested; namely, n-gram frequency, co-occurrence probability between consecutive letters, and a measure of expectation derived from error-driven learning, *activation*. Learning was tested by measuring key press latency in stimulus sequences that represent morphological patterns, similar to pronoun-verb matching in German (Experiment 1) and a random sequence of letters (Experiment 2).

With respect to the underlying learning mechanism, results showed that for both Experiments 1 and 2, activation captured more variance in response latencies than either n-gram frequency or co-occurrence probability. In addition, the size of the window in which predictions were successfully made differed between experiments. When there was more structure in the input (Experiment 1), increasing window size led to better model fit; when the sequence was random (Experiment 2), decreasing window sizes obtained a better fit. This suggests that the more structured, and thus the more predictable a sequence is, the more learnable it is. It also shows that learners cast their nets wide for cues to take into account in their predictions; as long as the predictive structure is there, the cues do not necessarily need to be closely temporally related to the expected events.

Finally, Experiment 3 investigated the factors affecting chunking—the integration of letters into something like a “word.” Chunking was tested with typing onset latency of each letter in the n-gram (bigram or trigram). There was no effect of activation on typing speed for any of the measures. Curiously, there was a significant effect of trigram frequency in the opposite-to-expected direction for the third letter of the trigrams, that is, increased frequency was associated with slower typing speed. We do not have a good explanation for why typing speed should be slower for the last letter of high-frequency trigrams. Given that this occurred only in trigrams and only for the third letter, this may be a spurious result, due to the small number of data points in Experiment 3. Overall, the results of Experiment 3 suggest that the participants did not learn "chunks" in this experiment. Although higher activation led to faster responses on individual trials during Experiments 1 and 2, higher activation did not lead to more rapid typing of whole sequences in Experiment 3. This is probably due to the design of the experiments: in Experiments 1 and 2, a single key press was required on each trial, interspersed with visual presentation of the target letter, leading to an alternation between single visual cues and motor responses. In contrast, Experiment 3 presented a series of two or three visual cues, which required one continuous motor response. It seems that the motor response required for a series of letters was not learned when the letters were presented on screen individually.

6.1. Learning mechanisms

Serial pattern learning is a widely used paradigm in psycholinguistics and other areas of cognitive science. As of March 2022, the Saffran et al. (1996a) paper has over 6000 citations.[5] Although the scope of processes under investigation has expanded—from learning syllables, to other linguistic and cognitive tasks—it remains the case that a quarter of a century after its publication, few studies have investigated the mechanisms underlying this learning. It is often implicitly or explicitly assumed that tracking the co-occurrence probability is the learning mechanism involved. However, the assumption is rarely tested. To take just one example, the majority of presentations at the Interdisciplinary Advances in Statistical Learning Conference (2019) used some version of this paradigm with the analyses often based on frequency or co-occurrence measures. The present results show that while these statistical measures capture a large amount of variance in the data, discriminative learning measures improve on the statistical models, capturing more of the data.

Frost et al.'s (2019) review of two decades worth of statistical learning experiments found that the majority of studies were based on the sequence learning paradigm of Saffran et al. (1996a) and investigated transitional probabilities, not just in language, but in a variety of domains. Frost et al. (2019) concluded that the field of statistical learning has made significant contributions. However, they also argue that it also has some significant shortcomings, such as a lack of specificity in the assumptions of statistical learning research. Frost et al. also found that the statistical learning studies often tested learning only with a forced-choice task after training, which meant that studies often failed to provide online learning data, or information about the learning trajectories. The present study addressed both of these concerns. The error-driven learning models produce a precise numeric prediction, activation, which also provides an estimate of the trial-by-trial learning trajectory.

Frost et al. (2019) also pointed out that there has been relatively little focus on the learning mechanisms, with researchers tending to treat the statistical learning as a "black box."

A key contribution of the present study has been to address the question of what learning mechanism underlies learning of sequences. It has often been assumed in previous work that learning results from transitional probabilities. In the present study, by comparing our error-driven learning measure, activation, to predictions based on frequency and transitional probabilities, we have shown that error-driven learning captures more variance in the data than these other highly frequently used measures. Theories of sequence learning in language and other domains should take this into account.

This learning mechanism is not restricted to the learning of sequences, however. Recent work in phonetic learning also suggests a role of error-driven learning in first and second language acquisition. Nixon (2020) showed that error-driven learning better accounted for learning of second language speech cues than statistical learning models. In first language acquisition, Nixon and Tomaschek (2020, 2021) present a computational model of early infants' learning of speech by using the incoming acoustic signal to predict upcoming acoustic signal. Apart from phonetic learning, error-driven learning has also been found to play a role in word learning (Ramscar et al., 2013a; Ramscar et al., 2010; Ramscar et al., 2011), morphological learning (Hoppe et al., 2020; Ramscar & Yarlett, 2007; Ramscar et al., 2013b; Tomaschek et al., 2019), lexical decision (Heitmeier, Chuang, & Baayen, 2023), speech production (Tucker et al., 2019; Tomaschek et al., 2019; Tomaschek & Ramscar, 2022), and speech perception (Arnold et al., 2017; Shafaei-Bajestan & Baayen, 2018). There is also evidence for trial-by-trial error-driven learning in the brain (Lentz et al., 2021).

6.2. Predictive structure and window size

Cleeremans and McClelland (1991) found that increased structure in the input led in turn to shorter latencies in responses. They explained this finding by suggesting that participants made predictions about upcoming trials based on the preceding trials. The present results provide further confirmation of this hypothesis. When there was (morphological) structure in the data (Experiment 1), participants made predictions based on information from a larger window size compared to when the sequence was random (Experiment 2).

7. Conclusion

In summary, the present study investigated the underlying mechanisms involved in sequence learning, and the distance from which such learning occurs. The results showed that error-driven learning predicted typing latencies better than statistical learning measures, whether n-gram frequency or transitional probability. That is, participants did not just keep track of statistical probabilities, but developed expectations about upcoming events, which were adjusted trial-by-trial through feedback from prediction error. The distance between the predictions and the predicted events—the window of integration—depends on the structure in the sequence: when the predictive structure was greater, the window of integration also increased. This suggests that the net is cast relatively wide for potential cues: if the predictive structure is there, it is not necessary for there to be a close temporal relation between the cues and the predicted event.

Acknowledgments

A version of the present study was presented at Linguistic Evidence 2019. This research was supported by a collaborative grant from the Deutsche Forschungsgemeinschaft (German Research Foundation; Research Unit FOR2373 "Spoken Morphology," Project "Articulation of morphologically complex words" BA 3080/3-1 and BA 3080/3-2). We would like to thank three anonymous reviewers and Vsevolod Kapatsinski for very helpful feedback which greatly improved the manuscript.

Open access funding enabled and organized by Projekt DEAL.

Notes

[1] See Nixon, van Rij, Mok, Baayen, and Chen (2016), Tomaschek, Tucker, Fasiolo, and Baayen (2018), Tomaschek et al. (2020), Baayen and Linke (2019) for more information on GAMs.

[2] In response to a reviewer comment, we also checked the pronoun boundary. The pronoun boundary results confirmed that of the morpheme boundary; namely, there was a slower typing latency following the boundary compared to the other letters.

[3] https://osf.io/rxgm6/

[4] See van Rij's `compareML()` function in the `itsadug` R package (van Rij et al., 2015) for an explanation of GAMM model comparison.

[5] Google Scholar.

References

Arnold, D., Tomaschek, F., Sering, K., Lopez, F., & Baayen, R. H. (2017). Words from spontaneous conversational speech can be recognized with human-like accuracy by an error-driven learning algorithm that discriminates between meanings straight from smart acoustic features, bypassing the phoneme as recognition unit. *PLoS ONE*, *12*, e0174623.

Arnon, I., & Ramscar, M. (2012). Granularity and the acquisition of grammatical gender: How order-of-acquisition affects what gets learned. *Cognition*, *122*, 292–305.

Arppe, A., Hendrix, P., Milin, P., Baayen, R. H., Sering, T., & Shaoul, C. (2018). ndl: Naive Discriminative Learning. https://CRAN.R-project.org/package=ndl.

Aylett, M., & Turk, A. (2006). Language redundancy predicts syllabic duration and the spectral characteristics of vocalic syllable nuclei. *Journal of the Acoustical Society of America*, *119*, 3048–3058.

Baayen, R. H., & Linke, M. (2019). An introduction to the generalized additive model. In *A practical handbook of corpus linguistics*. Berlin: Springer.

Baayen, R. H., Milin, P., Durdevic, D., Hendrix, P., & Marelli, M. (2011). An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. *Psychological Review*, *118*, 438–481.

Bates, D. M., & Sarkar, D. (2005). The lme4 library. http://lib.stat.cmu.edu/R/CRAN/.

Bertram, R., Tønnessen, F. E., Strömqvist, S., Hyönä, J., & Niemi, P. (2015). Cascaded processing in written compound word production. *Frontiers in Human Neuroscience*, *9*, 207.

Bröker, F., & Ramscar, M. (2020). Representing absence of evidence: Why algorithms and representations matter in models of language and cognition. *Language, Cognition and Neuroscience*, *38*, 297–620.

Chang, F. (2002). Symbolically speaking: A connectionist model of sentence production. *Cognitive Science*, *26*, 609–651.

Chang, F., Dell, G. S., & Bock, K. (2006). Becoming syntactic. *Psychological Review*, *113*, 234.

Cleeremans, A., & McClelland, J. (1991). Learning the structure of event sequences. *Journal of Experimental Psychology. General*, *120*, 235–253.

Cohen Priva, U. (2015). Informativity affects consonant duration and deletion rates. *Laboratory Phonology*, *6*, 243–278.

Daw, N. D., Courville, A. C., & Dayan, P. (2008). Semi-rational models of conditioning: The case of trial order. In N. Chater & M. Oaksford (Eds.), *The probabilistic mind* (pp. 431–452). Oxford: Oxford University Press.

Eimas, P. D. (1969). Multiple-cue discrimination learning in children. *Psychological Record*, *19*, 417–424.

Ellis, N. C. (1994). *Implicit and explicit learning of languages*. San Diego, CA: Academic Press.

Ellis, N. C. (2006). Language acquisition as rational contingency learning. *Applied Linguistics*, *27*, 1–24.

Frost, R., Armstrong, B. C., & Christiansen, M. H. (2019). Statistical learning research: A critical review and possible new directions. *Psychological Bulletin*, *145*, 1128.

Harmon, Z., Idemaru, K., & Kapatsinski, V. (2019). Learning mechanisms in cue reweighting. *Cognition*, *189*, 76–88.

Harmon, Z., & Kapatsinski, V. (2020). The best-laid plans of mice and men: Competition between top-down and preceding-item cues in plan execution. In: CogSci.

Hastie, T., & Tibshirani, R. (1986). Generalized additive models (with discussion). *Statistical Science*, *1*(3), 297–318.

Heitmeier, M., Chuang, Y. Y., & Baayen, R. H. (2023). How trial-to-trial learning shapes mappings in the mental lexicon: Modelling lexical decision with linear discriminative learning. *Cognitive Psychology*, *146*, 101598.

Hoppe, D. B., van Rij, J., Hendriks, P., & Ramscar, M. (2020). Order matters! Influences of linear order on linguistic category learning. *Cognitive Science*, *44*, 1–43.

Howard, J. H., Mutter, S. A., & Howard, D. V. (1992). Serial pattern learning by event observation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*, 1029–1039.

Hull, C. L. (1943). *Principles of behavior: An introduction to behavior theory*. Appleton-Century.

Jiménez, L., Paz, C., & Cleeremans, A. (1996). Comparing direct and indirect measures of sequence learning. *Journal of Experimental Psychology Learning Memory and Cognition*, *22*, 948–969.

Kamin, L. J. (1968). Attention-like processes in classical conditioning. In: M. R. Jones (Ed.), *Miami Symposium on the Prediction of Behavior* (pp. 9–31). Miami, FL: Miami University Press.

Kruschke, J. K. (2006). Locally Bayesian learning with applications to retrospective revaluation and highlighting. *Psychological Review*, *113*, 677.

Lentz, T., Nixon, J. S., & Rij, J. v. (2021). Temporal response modelling uncovers electrophysiological correlates of trial-by-trial error-driven learning. https://psyarxiv.com/dg5mw/.

Miller, G., & Chomsky, N. (1963). Finitary models of language users. In: R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology* (pp. 419–492). John Wiley.

Nieder, J., Tomaschek, F., Cohrs, E., & de Vijver, R. v. (2021). Modelling Maltese noun plural classes without morphemes. *Language, Cognition and Neuroscience*, *37*, 1–22.

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

Nixon, J. S. (2018). Effective acoustic cue learning is not just statistical, it is discriminative. *Proceedings of Interspeech 2018* (pp. 1447–1451).

Nixon, J. S. (2020). Of mice and men: Speech sound acquisition as discriminative learning from prediction error, not just statistical tracking. *Cognition*, *197*, 104081.

Nixon, J. S., Poelstra, S., & van Rij, J. (2022). Does error-driven learning occur in the absence of cues? Examination of the effects of updating connection weights to absent cues. In: *Proceedings of the 44th Annual Meeting of the Cognitive Science Society*.

Nixon, J. S., van Rij, J., Mok, P., Baayen, R. H., & Chen, Y. (2016). The temporal dynamics of perceptual uncertainty: Eye movement evidence from Cantonese segment and tone perception. *Journal of Memory and Language*, *90*, 103–125.

Nixon, J. S., & Tomaschek, F. (2020). Learning from the acoustic signal: Error-driven learning of low-level acoustics discriminates vowel and consonant pairs. In: *Proceedings of the 42nd Annual Conference of the Cognitive Science Society* (pp. 585–591).

Nixon, J. S., & Tomaschek, F. (2021). Prediction and error in early infant speech learning: A speech acquisition model. *Cognition*, *212*, 104697.

Nixon, J. S., & Tomaschek, F. (2023). Introduction to the Special Issue: Emergence of speech and language from prediction error: Error-driven language models. *Language, Cognition and Neuroscience*, *38*, 411–418.

Olejarczuk, P., Kapatsinski, V., & Baayen, R. H. (2018). Distributional learning is error-driven: The role of surprise in the acquisition of phonetic categories. *Linguistics Vanguard*, *4*, 1–9.

Ramscar, M. (2013). Suffixing, prefixing, and the functional order of regularities in meaningful strings. *Psihologija*, *46*, 377–396.

Ramscar, M. (2021). A discriminative account of the learning, representation and processing of inflection systems. *Language, Cognition and Neuroscience*.

Ramscar, M., & Dye, M. (2009). Error and expectation in language learning: An inquiry into the many curious incidences of ‘mouses’ in adult speech. In: *Proceedings of the 31st Annual Conference of the Cognitive Science Society* (pp. 485–490). Amsterdam, The Netherlands.

Ramscar, M., Dye, M., & Klein, J. (2013a). Children value informativity over logic in word learning. *Psychological Science*, *24*, 1017–1023.

Ramscar, M., Dye, M., & McCauley, S. (2013b). Error and expectation in language learning: The curious absence of ‘mouses’ in adult speech. *Language*, *89*, 760–793.

Ramscar, M., Dye, M., Popick, H., & O’Donnell-McCarthy, F. (2011). The enigma of number: Why children find the meanings of even small number words hard to learn and how we can help them do better. *PLoS One*, *6*, e22501.

Ramscar, M., & Yarlett, D. (2007). Linguistic self-correction in the absence of feedback: A new approach to the logical problem of language acquisition. *Cognitive Science*, *31*, 927–960.

Ramscar, M., Yarlett, D., Dye, M., Denny, K., & Thorpe, K. (2010). The effects of feature-label-order and their implications for symbolic learning. *Cognitive Science*, *34*, 909–957.

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

Rebuschat, P., & Monaghan, P. (2019). Editors’ introduction: Aligning implicit learning and statistical learning: Two approaches, one phenomenon. *Topics in Cognitive Science*, *11*, 459–467.

Rescorla, R., & Wagner, A. (1972). A theory of pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In: A. H. Black, & W. Prokasy (Eds.), *Classical conditioning II: Current research and theory (pp. 64–69)*. New York: Appleton Century Crofts.

van Rij, J., Wieling, M., Baayen, R. H., & van Rijn, H. (2015). itsadug: Interpreting Time Series, Autocorrelated Data Using GAMMs. https://cran.r-project.org/web/packages/itsadug/index.html.

Rosenblatt, F. (1962). *Principles of neurodynamics; Perceptrons and the theory of brain mechanisms*. Washington, DC: Spartan Books.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996a). Statistical learning by 8-month-old infants. *Psycholinguistics: Critical Concepts in Psychology*, *4*, 1926–1928.

Saffran, J. R., Johnson, E. K., Aslin, R. N., & Newport, E. L. (1999). Statistical learning of tone sequences by human infants and adults. *Cognition*, *70*, 27–52.

Saffran, J. R., Newport, E. L., & Aslin, R. N. (1996b). Word segmentation: The role of distributional cues. *Journal of Memory and Language*, *35*, 606–621.

Schmitz, D., Plag, I., Baer-Henney, D., & Stein, S. D. (2021). Durational differences of word-final /s/ emerge from the lexicon: Modelling morpho-phonetic effects in pseudowords with linear discriminative learning. *Frontiers in Psychology*, *12*, 2983.

Seyfarth, S., & Myslin, M. (2014). Discriminative learning predicts human recognition of English blend sources. In: *Proceedings of the Annual Meeting of the Cognitive Science Society*.

Shafaei-Bajestan, E., & Baayen, R. H. (2018). Wide learning for auditory comprehension. In: Interspeech (pp. 966–970).

Siegelman, N., Bogaerts, L., Christiansen, M. H., & Frost, R. (2017). Towards a theory of individual differences in statistical learning. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*, 20160059.

Stein, S. D., & Plag, I. (2021). Morpho-phonetic effects in speech production: Modeling the acoustic duration of English derived words with linear discriminative learning. *Frontiers in Psychology*, *12*, 3060.

Thorndike, E. L. (1898). Animal intelligence: An experimental study of the associative processes in animals. *Psychological Review: Monograph Supplements, 2*, New York: The MacMillan Company.

Tomaschek, F., Arnold, D., Sering, K., van Rij, J., Tucker, B. V., & Ramscar, M. (2020). Articulatory variability is reduced by repetition and predictability. *Language and Speech*, *64*, 654–680.

Tomaschek, F., Plag, I., Ernestus, M., & Baayen, R. H. (2019). Phonetic effects of morphology and context: Modeling the duration of word-final S in English with naïve discriminative learning. *Journal of Linguistics*, *57*, 123–161.

Tomaschek, F., & Ramscar, M. (2022). Understanding the phonetic characteristics of speech under uncertainty – implications of the representation of linguistic knowledge in learning and processing. *Frontiers in Psychology*, *13*, 1–20.

Tomaschek, F., Tucker, B. V., Fasiolo, M., & Baayen, R. H. (2018). Practice makes perfect: The consequences of lexical proficiency for articulation. *Linguistics Vanguard*, *4*, 1–13.

Tucker, B. V., Sims, M., & Baayen, R. H. (2019). Opposing forces on acoustic duration. Technical Report. psyarxiv.com/jc97w.

Vujović, M., Ramscar, M., & Wonnacott, E. (2021). Language learning as uncertainty reduction: The role of prediction error in linguistic generalization and item-learning. *Journal of Memory and Language*, *119*, 104231.

Weingarten, R., Nottbusch, G., & Will, U. (2004). Morphemes, syllables, and graphemes in written word production. *Trends in Linguistics Studies and Monographs*, *157*, 529–572.

Widrow, B., & Hoff, M. E. (1960). Adaptive switching circuits. In: 1960 WESCON Convention Record Part IV (pp. 96–104). New York.

Wood, S. N. (2006). *Generalized additive models*. New York: Chapman & Hall/CRC.