Hebbian learning can explain rhythmic neural entrainment to statistical regularities

Ansgar D. Endress, City, University of London

Abstract

In many domains, continuous sequences are composed of discrete recurring units. Learners need to extract these recurring units. A prime example is word learning. Fluent speech in unknown language is perceived as a continuous signal. Learners need to extract (and then memorize) its underlying words. One prominent candidate mechanism involves tracking how predictive syllables (or other items) are of one another, a strategy formalized as Transitional Probabilities (TPs). Syllables within the same word are more predictive of one another than syllables straddling word boundaries. Behaviorally, it is controversial whether such statistical learning abilities allow learners to extract and store the underlying units in memory, or whether they just track pairwise associations among syllables. The strongest evidence for the former view comes from electrophysiological results, showing evoked responses time-locked to word boundaries (e.g., N400's) as well as rhythmic activity with a periodicity of the unit duration. Here, I show that a simple Hebbian network can explain such results. When exposed to statistically structured syllable sequences, the network activation is rhythmic with a periodicity of the word duration and an activation maximum on word-final syllables. This is because word-final syllables receive more excitatory input from earlier syllables with which they are associated than syllables that occur earlier in words (and are less predictable). Hebbian learning can thus account for rhythmic neural activity in statistical learning task in the absence of memory representations for words, and previous reports of N400s time-locked to word boundaries might index the reduced predictability of word-initial syllables rather than word-onsets. I also suggest that learners might need to rely on other cues to extract and learn the words of their native language.

1 Introduction

During language acquisition, word learning is challenging even when the phonological form of words is known [Gillette et al., 1999, Medina et al., 2011]. However, speech in unknown languages often appears as a continuous signal with few cues to word onsets and offsets (but see [Brentari et al., 2011, Christophe et al., 2001, Endress and Hauser, 2010, Johnson and Jusczyk, 2001, Johnson and Seidl, 2009, Pilon, 1981, Shukla et al., 2007, 2011]). As a result, learners first need to discover where words start and where they end before than can commit any phonological word form to memory [Aslin et al., 1998, Saffran et al., 1996a,b] (and hopefully link it to some meaning). This challenge is called the segmentation problem.

Learners might solve the segmentation problem using co-occurrence statistics tracking the predictability of syllables. For example, a syllable following "the" is much harder to predict than a syllable following "whis". After all, "the" can precede any noun, but there are very few words starting with "whis" (e.g., whiskey, whisker, ...).

The most prominent version of such co-occurrence statistics involves Transitional Probabilities (TPs), i.e., the conditional probability of a syllable σ_2 following another syllable σ_1 $P(\sigma_2|\sigma_1)$. Infants, newborns and non-human animals are all sensitive to TPs in a variety of modalities [Aslin et al., 1998, Chen and Ten Cate, 2015, Creel et al., 2004, Endress, 2010, Endress and Wood, 2011, Fiser and Aslin, 2002b, 2005, Fló et al., 2022, Glicksohn and Cohen, 2011, Hauser et al., 2001, Kirkham et al., 2002, Saffran et al., 1996b,a, 1999, Saffran and Griepentrog, 2001, Sohail and Johnson, 2016, Toro et al., 2005, Turk-Browne et al., 2005, Turk-Browne and Scholl, 2009].

Following [Aslin et al., 1998, Saffran et al., 1996a,b], participants in a typical behavioral statistical learning

experiment in the auditory modality are first familiarized with a statistically structured speech stream (or a sequence in another modality such as auditory tones or visual symbols). The speech stream is a random concatenation of triplets of non-sense syllables (hereafter "words"). Syllables within words are thus more predictive of one another than syllable across word-boundaries. For example, if *ABC*, *DEF*, *GHJ* and *KLM* are "words" (where each letter represents a syllable), the *C* at the end of *ABC* can be followed by the word-initial syllables of any of the other words, while syllables within words predict each other with certainty.

A sensitivity to TPs is then tested by measuring a preference between high-TP items (i.e., words) and low-TP items created by taking either the final syllable of one word and the first two syllables from another word (e.g., *CDE*) or by taking the last two syllables of one word and the first syllable of the next word (e.g., *BCD*); the low-TP items are called part-words. Participants (adults, infants or other animals) usually discriminate between words and part-words, suggesting that they are sensitive to TPs. In humans, such a sensitivity to TPs might be the first step towards word learning.

1.1 Does statistical learning help learners memorizing words?

While many authors propose that tracking TPs leads to the addition of words to the mental lexicon (and thus to storage of word candidates in declarative long-term memory, LTM) [Erickson et al., 2014, Graf-Estes et al., 2007, Hay et al., 2011, Isbilen et al., 2020, Karaman and Hay, 2018, Perruchet, 2019, Shoaib et al., 2018], the extent to which a sensitivity to TPs really supports word learning is debated, and the results supporting this view often have alternative explanations that do not involve declarative LTM (see [Endress et al., 2020, Endress and de Seyssel, under review] for critical reviews). For example, while high-TP items are sometimes easier to memorize than low-TP items [Graf-Estes et al., 2007, Hay et al., 2011, Isbilen et al., 2020, Karaman and Hay, 2018], it is unclear if any LTM representation have been formed during statistical learning, or whether statistical associations simply facilitate subsequent associations. Likewise, while incomplete high-TP items are sometimes harder to recognize than entire items, such results can be explained by memory-less Hebbian learning mechanisms, and other attentional accounts [Endress and de Seyssel, under review].

Critically, to the extent that a sensitivity to TPs relies on implicit learning mechanisms [Christiansen, 2018, Perruchet and Pacton, 2006], statistical learning might be dissociable from explicit, declarative memory ([Cohen and Squire, 1980, Finn et al., 2016, Graf and Mandler, 1984, Knowlton et al., 1996, Poldrack et al., 2001, Sherman and Turk-Browne, 2020, Squire, 1992]; though different memory mechanisms can certainly interact during consolidation [Robertson, 2022]). In fact, there is evidence that a sensitivity to TPs is *not* diagnostic of the addition of items to the mental lexicon. For example, observers sometimes prefer high-TP items to low-TP items when they have never encountered either of them (when the items are played backwards compared to the familiarization stream; [Endress and Wood, 2011, Turk-Browne and Scholl, 2009, Jones and Pashler, 2007]), and sometimes prefer high-TP items they have never encountered over low-TP items they have heard or seen [Endress and Langus, 2017, Endress and Mehler, 2009b]. In such cases, a preference for high-TP items does not indicate that the high-TP items are stored in the mental lexicon, simply because learners have never encountered these items. Further, when learners have to repeat back the items they have encountered during a familiarization stream with as few as four items, they are unable to do so [Endress and de Seyssel, under review].

On the other hand, there is a straightforward explanation of such results that does not involve declarative LTM: a sensitivity to TPs might reflect Hebbian learning [Endress, 2010, Endress and Johnson, 2021]. After all, the representations of syllables (or other elements in a stream) presumably does not cease to be active as soon as the syllable ends. As a result, multiple syllables can be active together and can can thus form Hebbian associations. [Endress and Johnson, 2021] showed that such a network can account for a number of statistical learning results (see below).

However, there is another class of studies that seems to provide strong evidence for the possibility that statistical learning leads to the extraction of coherent units, and that seems to be inconsistent with a mere Hebbian interpretation of statistical learning results: electrophysiological responses to statistically structured sequences. I will now turn to this literature.

1.2 Electrophysiological correlates of statistical learning

In one of the earliest electrophysiological studies of statistical learning, [Sanders et al., 2002] first presented participants with a speech stream composed of non-sense words. Following this, they presented these words in isolation, and finally another speech stream with the same words. When they compared electrical brain responses to the second presentation of the stream and its first presentation, they observed increased N100 and N400 responses. That is, they showed increased negativities around 100 ms and 400 ms after word onsets (see also [Abla et al., 2008] for similar study with tones rather than syllables as stimuli). [Cunillera et al., 2006] showed that N400 effects can also be obtained without explicitly training participants on the words, and similar results obtain even in newborns [Kudo et al., 2011, Teinonen et al., 2009].

Following [Buiatti et al., 2009], electrophysiological investigations of statistical learning focused on rhythmic entrainment to the speech streams rather than event-related responses such as the N400. Specifically, if listeners learn the statistical structure of the speech stream, they should perceive the speech stream as a sequence of trisyllabic units (given that most statistical learning experiments tend to use tri-syllabic units or their equivalents in other domains, but see [Johnson and Tyler, 2010]), and thus perceive a rhythm with a periodicity of three syllable durations. If so, they should also show a *neural* rhythm with the same periodicity. While [Buiatti et al., 2009] detected such a rhythm only when words were separated by brief silences, later investigations found such rhythms in continuous sequences in adults [Batterink and Paller, 2017] (see also [Moser et al., 2021] for an magneto-encephalography study), children [Moreau et al., 2022], infants [Kabdebon et al., 2015] and even newborns [Fló et al., 2022].

Such results seem to strongly suggest that statistical learning creates integrated units that can (presumably) be stored in memory, and different authors have reached this conclusion [Batterink and Paller, 2017, Fló et al., 2022, Sanders et al., 2002, Teinonen et al., 2009]. However, the original interpretation of the N400 component suggests a different interpretation. In fact, the electrical N400 component is thought to reflect (semantically) surprising and thus unpredictable stimuli [Kutas and Federmeier, 2000]. However, in speech streams such as those used in statistical learning tasks, word onsets are always unpredictable, given that words are randomly concatenated, and word onsets remain unpredictable even when participants learn the statistical structure of the speech stream. As a result, it is unclear why N400-like responses should be time-locked to word *onsets*. In contrast, the last syllable of each word is predictable based on the statistical structure of the streams, but only after learning. As a result, electrophysiological responses might not so much index word onsets as reflect the increased predictability of word-final syllables (or the decreased relative predictability of word-initial syllables) as learning progresses.

A similar conclusion follows from simple associative considerations. For example, after a word ABC is learned (where each letter stands for a syllable), each syllable predicts subsequent syllables. As a result, the C syllable does not only receive (external) bottom-up excitation when it is encountered, but receives additional associative excitation from the preceding A and B syllables (that predict the C syllable). As a result, one would expect a neural rhythm with a period of three syllable durations, and a maximum following the onset of the *word-final* syllable even if no word has been stored in memory.

2 The current study

Here, I provide computational evidence for this idea, and show that such electrophysiological results can be explained in a simple, memory-less Hebbian network that has been used to account for a variety of statistical learning results [Endress and Johnson, 2021]. The network is a fairly generic saliency map [Bays et al., 2010, Endress and Szabó, 2020, Gottlieb, 2007, Roggeman et al., 2010, Sengupta et al., 2014] augmented by a Hebbian learning component. The network comprises units representing populations of neurons encoding syllables (or other items). All units are fully connected with both excitatory and inhibitory connections. Excitatory connections change according to a Hebbian learning rule, while inhibitory connections do not undergo learning. Additionally, activation decays exponentially in all units. Further details of the model can be found in Supplementary Information XXX.

Such an architecture can explain statistical learning results in a relatively intuitive way. For example, if each syllable is represented by some population of neurons, and learners listen to some sequence $ABCD\ldots$,

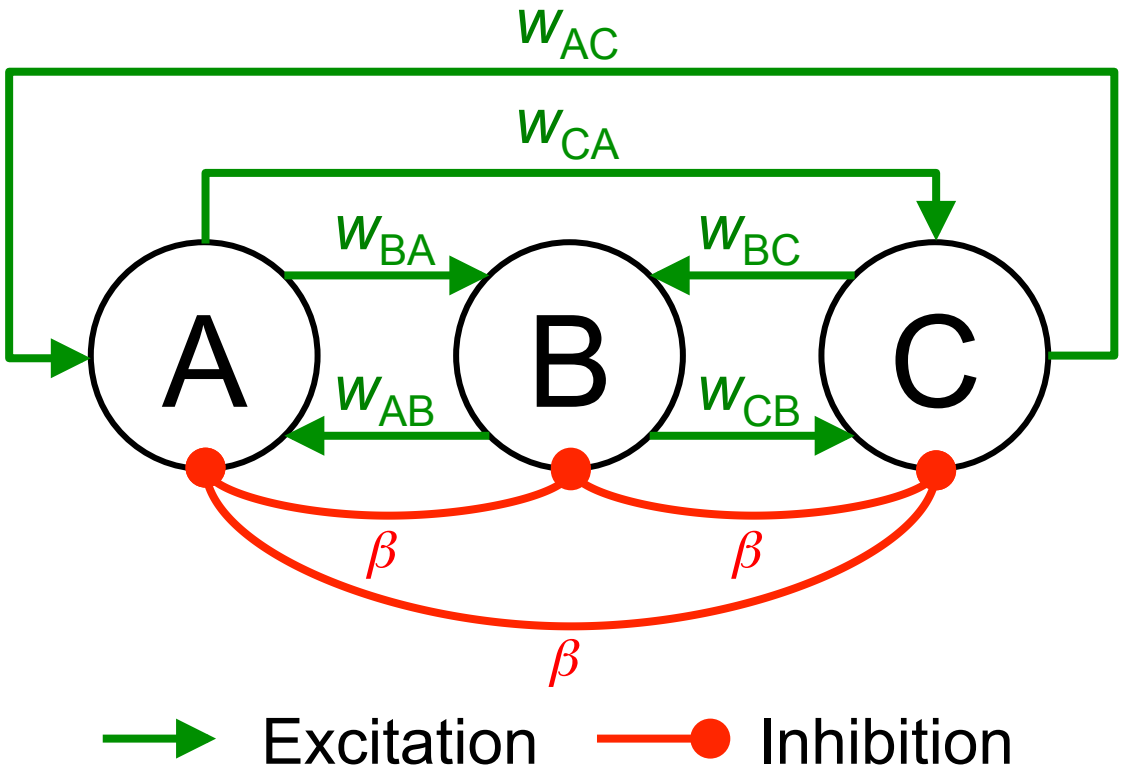

Figure 1: Illustration of the network architecture with only three units A, B and C. Here, these units encode syllable. All units mutually exite and inhibit one another. Excitatory connections undergo Hebbian learning. For example, unit A excites unit B with a tunable weight of w_{BA} as well as unit C with a weight of w_{CA}. In contrast, the inhibitory weight does not undergo learning. In addition to excitation and inhibition, all units undergo forgetting.

associations should form between adjacent and non-adjacent syllables depending on the decay rate. If activation decay is slower than a syllable duration, the representations of two adjacent syllables will be active at the same time, and thus form an association. For example, if a neuron representing A is still active while B is presented, these neurons can form an association. Similarly, if a neuron representing A is still active when C is presented, an association between these neurons will ensue although the corresponding syllables are not adjacent.

Further, this learning rule is non-directional. As a result, the network should be sensitive to associations irrespective of whether items are played in their original order (e.g., ABC) or in reverse order (e.g., BCA). [Endress and Johnson, 2021] showed that such a model can account for a number of statistical learning results (as long as the decay rate was set to a reasonable level) - in the absence of a dedicated memory store. Hence, statistical learning results can be explained even when participants do not create lexical entries for high-TP items.

However, the neural entrainment results above seem to suggest that learners do more than merely computing associations among syllables, and extract statistically coherent units [Batterink and Paller, 2017, Fló et al., 2022, Sanders et al., 2002, Teinonen et al., 2009]. Here, I argue that this simple Hebbian network can also account for the periodic activity found in electrophysiological recordings. Intuitively, if a high-TP item such as ABC is presented, A mostly receives external stimulation, but B receives external stimulation – as well as excitatory input from A, while C receives external stimulation as well as excitatory input from both A and B. As a result, the network activation should increase towards the end of a word, with a maximum on the third syllable, leading to periodic activity with a period of a word duration (though the presence of inhibitory connections might make the exact results more complex). If so, and as mentioned above, previous reports of N400 near a word boundary would not so much indicate the onset of a "word", but rather the onset of a "surprising" syllable.

I tested this idea in [Endress and Johnson, 2021] model. I exposed the network to a continuous sequence inspired by [Saffran et al., 1996a] Experiment 2. The sequence consisted of 4 distinct words of 3 syllables each. The familiarization sequence was a random concatenations of these words, with each word occurring 100 times. During the test phase, I recorded the total network activation as each of the test-items (see below) is presented, and assume that this activation reflects the network's familiarity with the words.[1] I simulated 100 participants by repeating the familiarization and test cycle 100 times.

The test items follow by [Saffran et al., 1996a] and [Saffran et al., 1996b], among many others. After exposure to the familiarization sequence, activation is recorded in response to words such as ABC and "part-words." As mentioned above, part-words comprise either the last two syllables from one word and the first syllable from the next word (e.g., *BC:D*, where the colon indicates the former word boundary that is not present in the stimuli) or the last syllable from one word and the first two syllables from the next word (e.g., *C:DE*). Part-words are thus attested in the familiarization sequence, but straddle a word boundary. As a result, they have weaker TPs than words. Accordingly, the network should be more familiar with words than with part-words. To assess whether the network can also account for results presented by [Fló et al., 2022] (see below), I also record activation after presenting the first two syllables of a word (e.g., AB) or the last two syllables (e.g., BC).

During the simulations, the network parameters for self-excitation and mutual inhibition are kept constant (α and β in Supplementary Material XXX). However, in line with [Endress and Johnson, 2021], I used different forgetting rates (λ_{act} in Supplementary Material XXX) between 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9. With exponential forgetting, a forgetting rate of 1 means that the activation completely disappears on the next time step (in the absence of excitatory input), a forgetting rate of zero means no forgetting at all, while a forgetting rate of .5 implies the activation is halved on the next time step (again, in the absence of excitatory input).[2]

[1] [Endress and Johnson, 2021] also reported simulations where they recorded the activation in the items comprising the current test-item rather than the global network activation. While the results were very similar to those using the total network activation, measuring activation in test items would not be meaningful in the current simulations as I seek to uncover periodic activity during familiarization.

[2] While I use the label "decay", I do not claim that "decay" reflects a psychological processes. The current implementation uses decay as a mechanism to limit activations in time, but the same effect could likely be obtained through inhibitory interactions or

other mechanisms.

3 Results

3.1 Preference for words over part-words

To establish the forgetting rates at which discrimination between words and part-words (and thus learning) can be observed, I first repeat some of [Endress and Johnson, 2021] results. I calculated normalized difference scores of activations for words and part-words, $d = \frac{\text{Word}-\text{Part-Word}}{\text{Word}+\text{Part-Word}}$, and evaluated these difference scores in two ways. First, I comparde them to the chance level of zero using Wilcoxon tests. Second, I counted the number of simulations (representing different participants) preferring words, and evaluated this count using a binomial test. With 100 simulations per parameter set, performance is significantly different from the chance level of 50% if at least 61 % of the simulations show a preference for the target items.

The results are shown in Figure 2 and Table 2. Except for low forgetting rates of up to .4, the network prefers words over part-words, with somewhat better performance for words against *C:DE* part-words, as has been observed in human participants by [Fiser and Aslin, 2002a]. In the following, I will thus use forgetting rates between 0.4 and 0.9 to model the electrophysiological results.

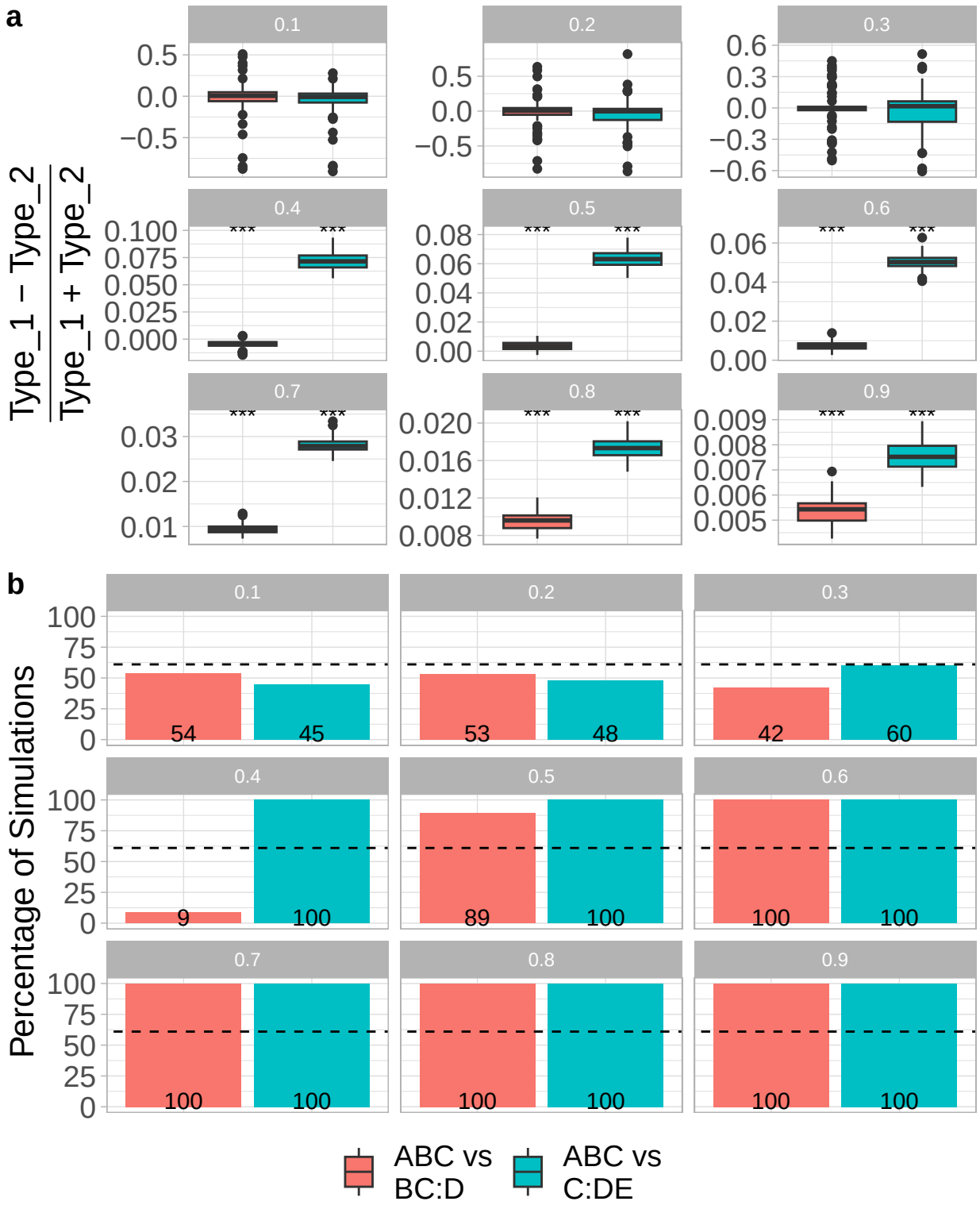

Figure 2: Results based on the global activation as a measure of the network's familiarity with the items for forgetting rates (between 0.1 and 0.9). (a) Difference scores between words and part-words. Significance is assessed based on Wilcoxon tests against the chance level of zero. (b) Percentage of simulations with a preference for words over part-words. The dashed line shows the minimum percentage of simulations that is significant based on a binomial test.

3.2 Electrophysiological results

3.2.1 Activation differences within words

I next asked whether a basic Hebbian learning model can explain periodic activity found in electrophysiological recordings [Buiatti et al., 2009, Batterink and Paller, 2017, Fló et al., 2022, Kabdebon et al., 2015, Moreau et al., 2022, Moser et al., 2021], focusing on the forgetting rates for which the network preferred words to part-words. In a first analysis, I simply recorded the total network activation after each syllable in a word had been presented. These activations were averaged for each syllable position (word-initial, word-medial and word-final) and for each participant after removing the first 200 words from the familiarization stream (during which the network was meant to learn).

As shown in Figure 3 and Table 3, activation was highest after word-final syllables (though not for very low forgetting rates for which I did not observe learning in the first place). As a result, a simple Hebbian learning model can account for rhythmic activity in electrophysiological recordings with a period equivalent to the word duration. Critically, and as mentioned above, while previous electrophysiological responses to statistical structured streams were interpreted in terms of a response to word onsets [Abla et al., 2008, Cunillera et al., 2006, Kudo et al., 2011, Sanders et al., 2002, Teinonen et al., 2009], the current results suggest an alternative interpretation of such effects. Rather than signalling the beginnings and ends of words, an activation maximum after the third syllable of each word might reflect the predictability of the third syllable, while a sudden drop in activation after the first syllable might indicate the lack of predictability. Importantly, such activation maxima can arise even if no word is stored in memory.

The reason for which lower forgetting rates do not necessarily lead to rhythmic activity is the interplay between decay and inhibition. To assess this possibility, I recorded the number of active neurons after a burn-in phase of 600 items. As shown in Table 4 and Figure 8, more neurons remain active at any point in time when the decay rate is lower, and might thus inhibit other neurons. When decay limits the effect of residual inhibitory input from other neurons, the pattern of connections between neurons then enables the network to exhibit periodic activity.

3.2.2 Spectral density

I next analyzed the frequency response of the network to the speech streams. Specifically, I estimated the spectral density of the time series corresponding to the total network activation after each time step (again after a burnin of 200 words), separately for each decay rate and simulation. I then extracted the frequency with the maximal density. As shown in Figure 4(a), the modal frequency for decay rates of least .4 was 1/3, corresponding to a period of three syllables. These results thus suggest again that a simple Hebbian learning mechanism can entrain to statistical rhythms in the absence of memory for words.

3.2.3 Phase analysis

The analyses of the network activations suggest that activations are strongest for word-final syllables, and that the network entrains to a periodicity of three syllables. However, the traditional interpretation of electrophysiological responses to statistical learning is that neural responses index word-initial syllables. To address this issue more directly, I calculated the phase of the network activation relative to wave forms with maxima on word-initial, word-medial and word-final syllables, respectively. Specifically, I calculated the cross-spectrum phase at the winning frequency between the total network activation and (1) three cosine reference waves with their maxima on the first, second or third syllable of a word as well as (2) a saw-tooth function with its maximum on the third syllable. As shown in Figure 4(b) and Table 5, the activation had a small relative phase relative to the cosine with the maximum on the third syllable or the saw tooth function. In contrast the phase relative to the cosine with the word-initial maximum was around 120 degrees, while that relative to the cosine with the maximum on the second syllable was around -120 degrees. These spectral analyses thus confirm that, at least for larger decay rates, the activation increases towards the end of a word, and that the network activation is roughly in phase with a function with a maximum on the third syllable.

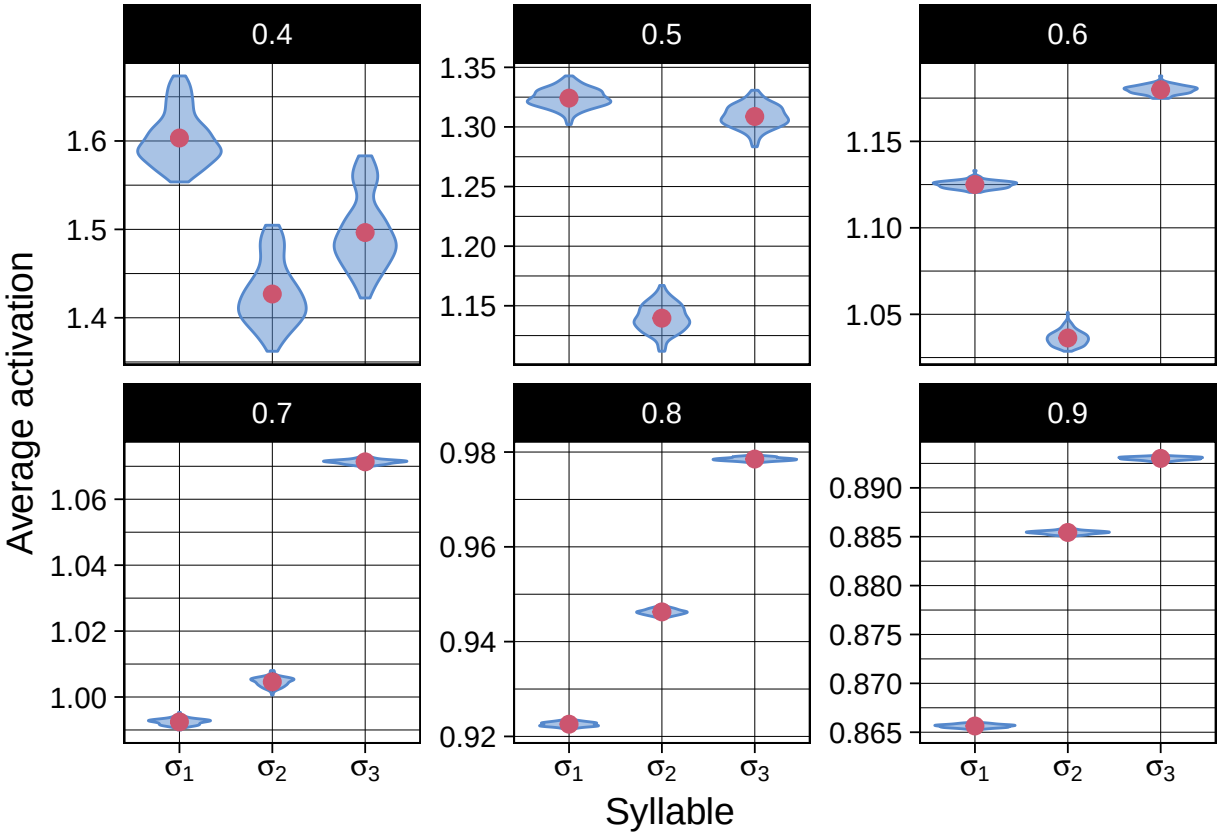

Figure 3: Average total network activation for different syllable positions during the familiarization with a stream following citeSaffran-Science. The facets show different forgetting rates. The results reflect the network behavior after the first 50 presentations of each word.

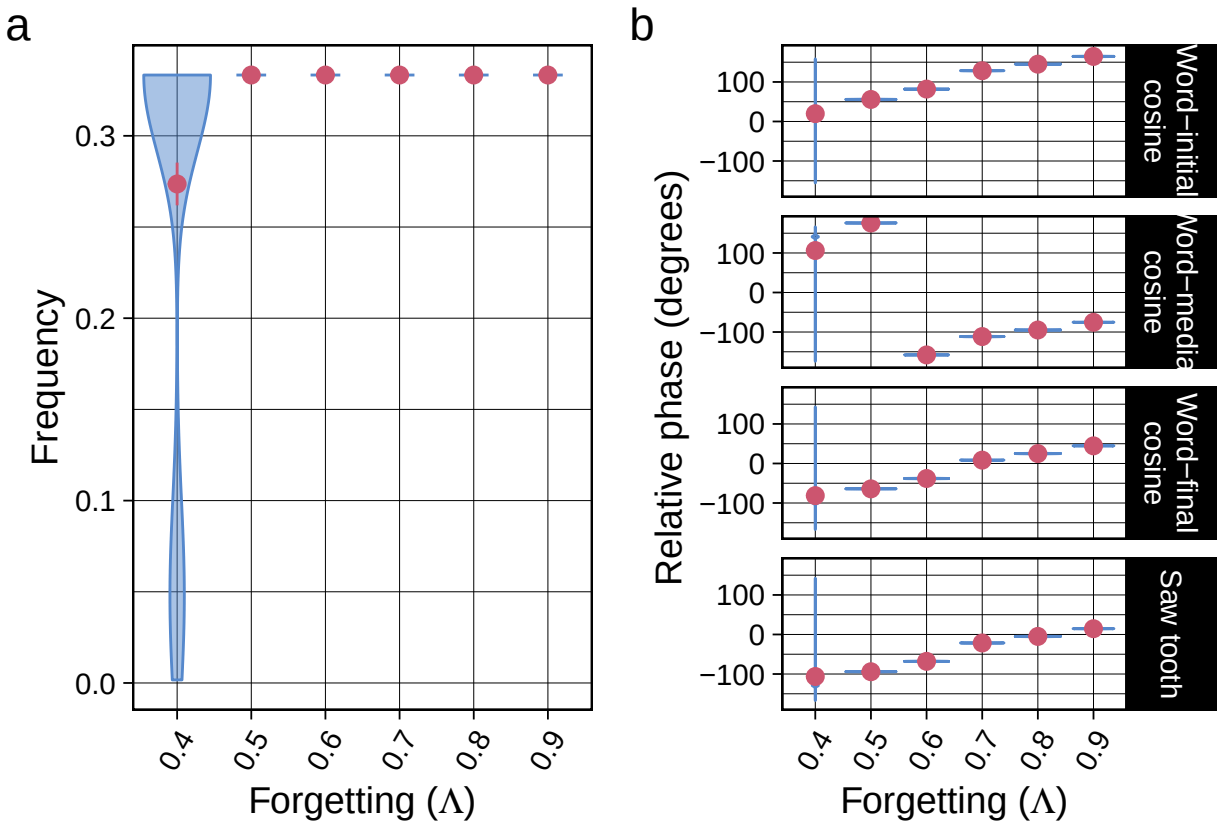

Figure 4: Spectral analysis of the total network activation during the familiarization with a stream following citeSaffran-Science. The results reflect the network behavior after the first 50 presentations of each word. (a) Maximal frequency as a function of the forgeting rates. For forgetting rates where learning takes place, the dominant frequency is 1/3, and thus corresponds to the word length. (b) Relative phase (in degrees) at the maximal frequency of the total network activation relative to (from top to bottom) a cosine function with its maximum at word-intial syllables, word-medial syllables and word-final syllables as well as a saw tooth function with the maximum on the third syllable. For forgetting rates where learning takes place, the total activation is in phase with a cosine with its maximum on the word-final syllable as well as with the corresponding saw tooth function.

3.2.4 Memory for word onsets vs. offsets [Fló et al., 2022]

The results so far suggest that a simple Hebbian network can reproduce rhythmic activity in the absence of memory for words. However, [Fló et al., 2022] suggested that neonates retain at least the first syllable of statistical defined words, if not the entire words. Specifically, they presented newborns with items starting with two syllables that occurred word-initially ($AB\ldots$), and with items starting with a word-medial syllable ($BC\ldots$) and observed early ERP differences between these items.

To reproduce these results, I measured the activation of the network in response to isolated, bisyllabic AB and BC test items, respectively. As shown in Figure 5 and Table 6, the network activation was always greater in response to BC items than to AB items except for the largest decay rates. The reasons is presumably that a B syllable is strongly associated with both A and C syllables, which are associated with each other in turn. In contrast, A syllables are only strongly associated with B syllables and more weakly with C syllables. Upon presentation of the second syllable, second order activation should thus be greater for BC items than for AB items. Be that as it might, these analyses show that a memory-less system can reproduce differential responses to AB and BC items.

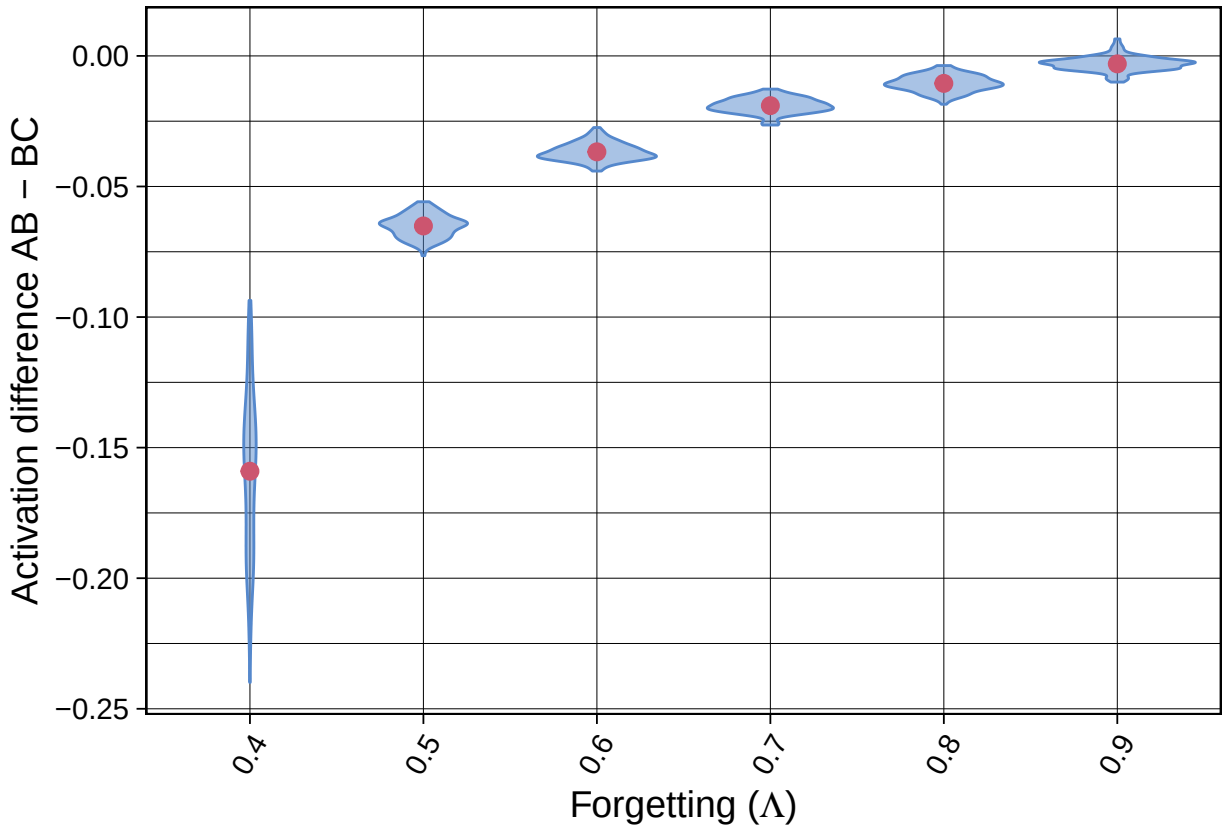

Figure 5: Average difference in the total network activation for the first two syllables of a word (AB) and the first to syllables of a part-word (BC) after familiarization with a stream following citeSaffran-Science. The results reflect the network behavior after the first 50 presentations of each word. Positive values indicate greater activation for the AB items than the BC items.

3.2.5 Positional coding (NEW IN REVISION)

Other investigators used electrophysiological recordings to reveal more abstract information extracted from statistically structured sequences. For example, [Henin et al., 2021] found that, in intracranial recordings, the representations of items sharing a sequential position within words (i.e., word-initial, word-medial or word-final) were more similar than the representations of items from different positions, and concluded that this representational similarity might support positional codes.

However, behavioral evidence suggests that such positional codes are not available after familiarizations with continuous sequences (e.g. [Marchetto and Bonatti, 2013, Endress and Bonatti, 2007, Endress and Mehler, 2009a, Endress and Bonatti, 2016, Peña et al., 2002]), raising the question of whether this similarity is behaviorally relevant for learning.

Here, I thus suggest an alternative explanation: The positional similarity might be a side effect of associative processing. Specifically, items in the same sequential position share their context. For example, all word-initial syllables are surrounded by the same set of word-final syllables and vice-versa. If so, and if the context and the focal syllables activate each other, one would expect a certain degree of representational overlap of syllables sharing a sequential position, without this necessarily having behavioral consequences.

To evaluate this idea, I extracted the activation in all neurons in all time steps (after burn in). I then calculated, for each forgetting rate, simulated participant and syllable, the average activation in all neurons. I then extracted the "representation" of all syllables, that is, the vector of activations observed after a syllable has been presented. I calculated the cosine similarity across the representations of all syllable pairs (i.e., the normalized dot product), and calculated an average similarity for each forgetting rate, simulated participant and match type (positional match vs. non-match). Finally, I calculated, for each forgetting rate and simulated participant, the relative similarity difference for non-matching vs. matching pairs, $\frac{\text{non-match}-\text{match}}{\text{non-match}+\text{match}}$ and evaluated these difference scores with a Wilcoxon test against the chance level of zero.

As shown in Figure 6 and Table 7, the similarity score for non-match pairs was significantly higher than for match pairs for all forgetting rates. The reason for the inversion of the sign of the difference with respect to [Henin et al., 2021] is presumably the time course in the current model vs. in actual biological tissue. In the current simulations, a syllable duration is a discrete time-step. The activations reported here are thus snapshots of more continuously evolving activations. As a result, the representations of, say, word-initial and word-final syllables overlap, given that these syllables excite each other in the same time step. In contrast, given the localist coding scheme used here, there is no overlap in the representations of syllables occupying the same sequential position.

In contrast, with realistic activation time courses, the time-resolved similarity measures used by [Henin et al., 2021] can capture the actual time course of the associative activation. For example, upon presentation of a word-initial syllable, such measures can capture any lingering activation of the representations of the previous word-final syllable as well as its reactivation through excitation from the word-initial syllable. I surmise that these time series make the same same-position representations more similar. Be that as it might, the current model can differentiate between different sequential positions.

Given that, behaviorally, the positional codes do not seem to be available to actual learners, the current results also suggest that caution is required when interpreting information that can be decoded from the brain without being behaviorally relevant. To take a non-psychological example, audio recordings often contain noise from the electric grid from which spatial and temporal localization information can be decoded (e.g., for forensic purposes [Grigoras, 2005]). However, while this information is clearly present in the recordings, it is not relevant for the primary means by which audio information is consumed (i.e., by listening to it). Mutatis mutandis, some information might be present in neural activity as a side effect of the mechanics of neural processing, but whether this information is behaviorally relevant is an independent and empirical question.

3.2.6 Effects of word length (NEW IN REVISION)

I next asked whether the network can entrain to statistical regularities when the familiarization streams are composed of words of arbitrary length. Intuitively, given that the periodicity reported here arises due to the increasing cumulative excitatory input towards the ends of words, one would expect the network to be unable

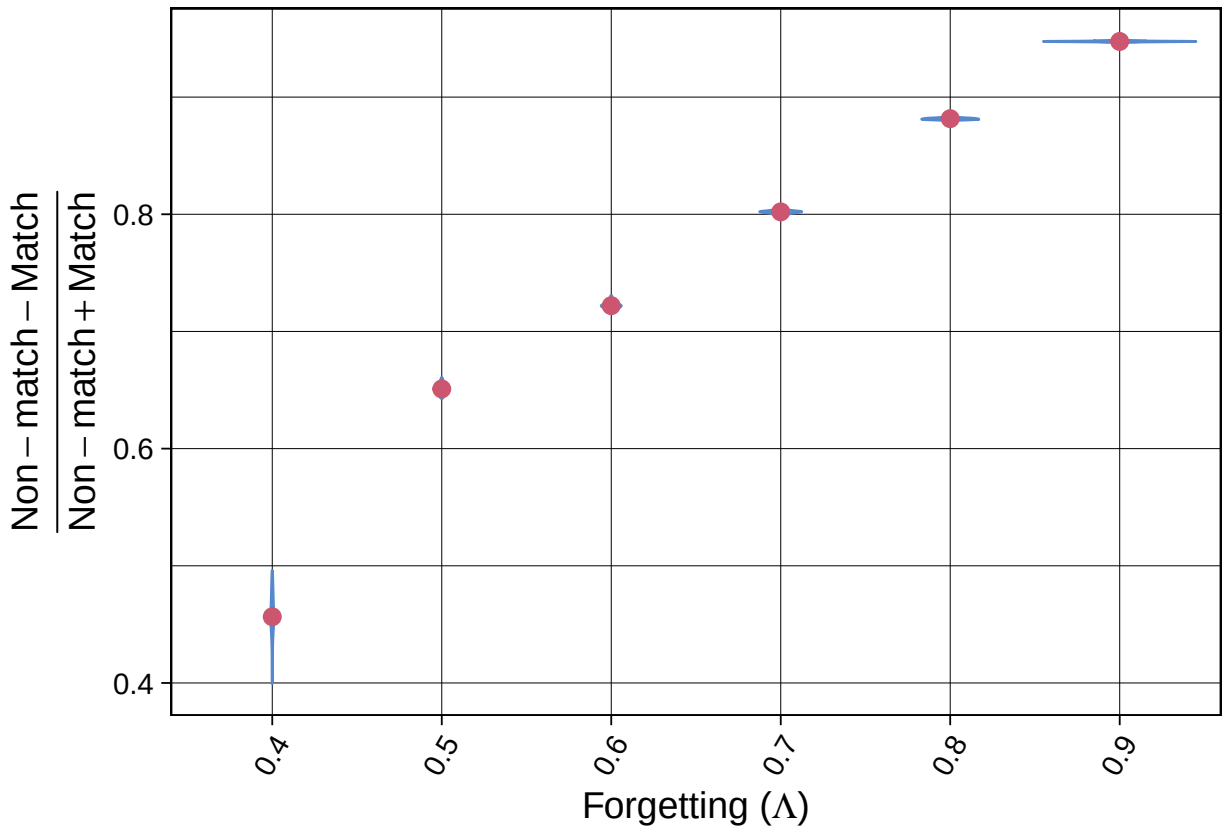

Figure 6: Simulations of positional codes citeHenin2021. Representations of syllables *not* sharing a sequential position (i.e., word-initial, word-medial, word-final) are more similar to each other than representations of syllables sharing a sequential position. The similarity was evaluated using the cosine similarity across the vectors of activation elicited by the syllables.

to track statistical periodicity for excessively long words. After all, for sufficiently long words, the activation from to earlier syllables will have disappeared once the input reaches the end of a word.

There is some evidence supporting this idea. For example, [Benjamin et al., 2023] did not find neural entrainment to 4-syllable words in newborns (though a failure to detect entrainment might have other reasons than the word length.) Computationally, however, it is also conceivable that networks can deal with longer words, if spreading activation due to higher order associations are sufficient to increase the network activation towards the end of a word.

To examine this issue, I repeated the simulations above, but with word lengths between 3 and 18 syllables, again for the same forgetting rates as in the simulations above and 100 simulated participants. I estimated the spectral density of the time series corresponding to the total network activation after each time step (again after a burnin of 200 words), separately for each decay rate and simulation. I then extracted the frequency with the maximal density, and averaged these frequencies across participants.

As shown in Figures **??** and **??**, the network successfully tracked the periodicity up to a word length of up to and including 8 syllables. For 8 syllable-words, the winning frequency was either $\frac{1}{8}$ or $\frac{1}{4}$, depending on the forgetting rate. In other words, the network extracted a periodicity whose period was a fraction of the word length. For longer words, the winning frequency was generally a fraction of the word length, with a multiplier of 2 or 3.

As a result, there seems to be a limit to how long words can be so that the network can entrain to a statistically induced rhythm. Here, the limit seems to be a word length of 8 syllables, but the specific limit likely depends on the interplay between the forgetting, excitation and inhibition parameters.[3]

[3]See SI XXX for other quantitive results where the network makes incorrect predictions.

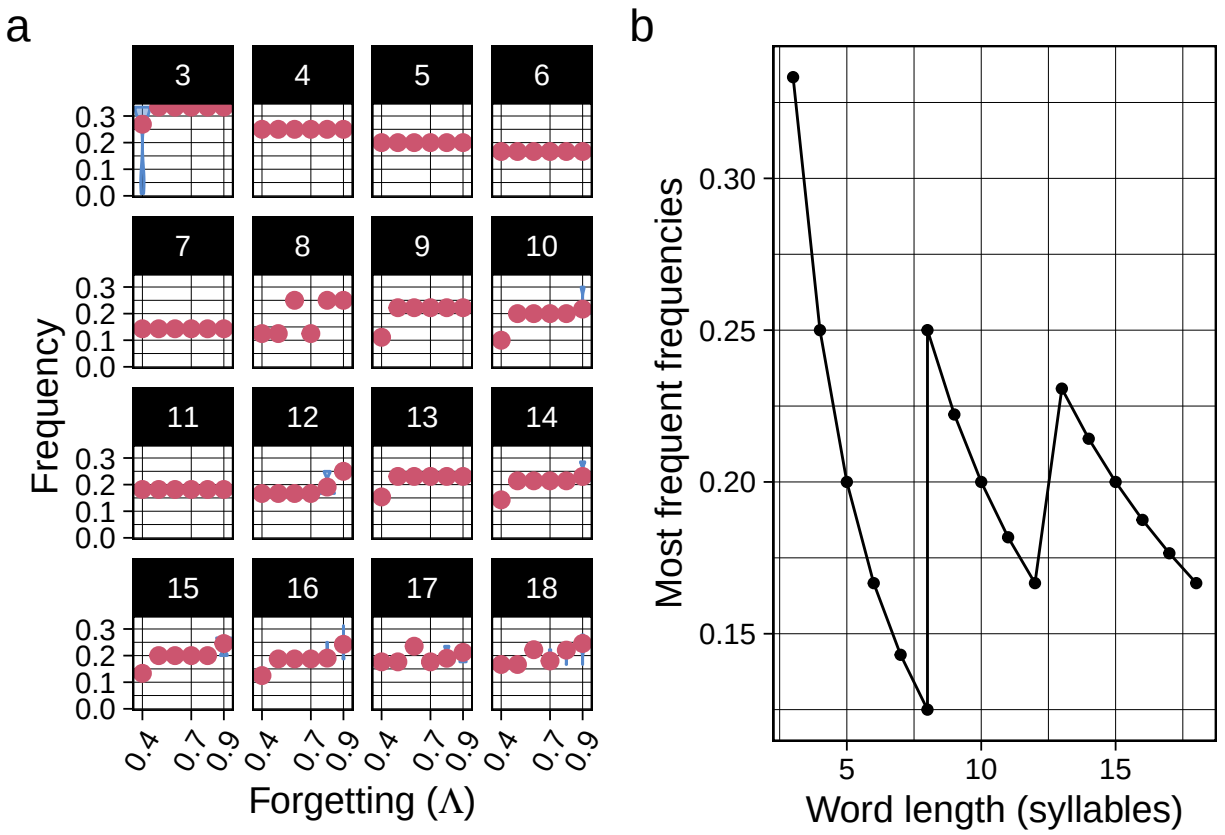

Figure 7: Entrainment as a function of word-length. (a) Dominant frequency as a function of the forgetting rate (x-axix, Λ) and the word-length (3-18 syllables; facets). (b) Most frequent dominant frequency across forgetting rates as a function of the word-length. For words of up to and including 8 syllables, the network entrains to a frequency equivalent to the word-length. For words of 8 or more syllable, the network entrains to a multiple of that frequency.

4 Discussion

To acquire the words of their native language, learners need to extract them from fluent speech, and might use co-occurrence statistics such as TPs to do so. If so, high-TP items should be stored in memory for later use as words. Strong evidence in favor of this possibility comes from electrophysiology, where rhythmic activity has been observed in response to statistically structured sequences. In the time domain, different authors have observed amplitude peaks around the boundaries of statistically defined words [Abla et al., 2008, Cunillera et al., 2006, Kudo et al., 2011, Sanders et al., 2002, Teinonen et al., 2009]; in the frequency domain, a frequency response with a period of the word duration emerges as participants learn the statistical structure of the speech stream [Buiatti et al., 2009, Batterink and Paller, 2017, Fló et al., 2022, Kabdebon et al., 2015, Moreau et al., 2022, Moser et al., 2021].

Here, I show that such results can be explained by a simple Hebbian learning model. When exposed to statistically structured sequences, the network activation increased towards the end of words due to increased excitatory input from second order associations. As a result, the network exhibits rhythmic activity with a period of a word duration. Critically, given that the network could reproduce these results in the absence of memory representations for words, earlier electrophysiological results might also index the statistical predictability of syllables rather than the acquisition of coherent units such as words. For example, and as mentioned above, N400 effects observed in statistical learning tasks [Abla et al., 2008, Cunillera et al., 2006, Kudo et al., 2011, Sanders et al., 2002, Teinonen et al., 2009] might not index the onset of words, but rather the lack of predictability of word-initial syllables (or the increased predictability of word-final syllables). This would also be more consistent with the initial description of the N400 component as an ERP component that indexes *unpredictable* events [Kutas and Federmeier, 2000].

As mention in the introduction, the view that statistical learning does not necessarily lead to storage in declarative memory memory is consistent with long-established dissociations between declarative memory and implicit learning [Cohen and Squire, 1980; Finn et al., 2016; Graf and Mandler, 1984, Knowlton et al. [1996]; Poldrack et al., 2001; Squire, 1992]. It is also consistent with a variety of behavioral results (see [Endress et al., 2020, Endress and de Seyssel [under review]] for critical reviews), including behavioral preferences for unattested high-TP items [Endress and Wood, 2011, Endress and Langus, 2017, Endress and Mehler, 2009b, Jones and Pashler, 2007, Turk-Browne and Scholl, 2009]), and the inability of adult learners to repeat back words from familiarization streams with as few as four words[Endress and de Seyssel, under review].

To identify words in fluent speech, learners might thus need to rely on other cues, including using known words as cues to word boundaries for other words [Bortfeld et al., 2005, Brent and Siskind, 2001, Mersad and Nazzi, 2012], paying attention to beginnings and ends of utterances [Monaghan and Christiansen, 2010, Seidl and Johnson, 2008, Shukla et al., 2007], phonotactic regularities [McQueen, 1998] and universal aspects of prosody [Brentari et al., 2011, Christophe et al., 2001, Endress and Hauser, 2010, Pilon, 1981]. Computational results suggest that such cues are promising, given that a computational model attending to utterance edges showed excellent segmentation and word-learning abilities [Monaghan and Christiansen, 2010].

In contrast, statistical learning might well be important for predicting events across time [Endress and de Seyssel, under review, Morgan et al., 2019, Sherman and Turk-Browne, 2020, Turk-Browne et al., 2010, Verosky and Morgan, 2021] and space [Theeuwes et al., 2022], an ability that is clearly critical for mature language processing [Levy, 2008, Trueswell et al., 1999] (as well as many other processes [Clark, 2013, Friston, 2010, Keller and Mrsic-Flogel, 2018]). This suggests the possibility that predictive processing might also be crucial for word learning, but it is an important topic for further research to find out how predictive processing interacts with language acquisition.

5 Supplementary Information

5.1 Supplementary Information 1: Model definition

The activation of the i^{th} unit is given by

$$\dot{x_i} = -\lambda_a x_i + \alpha \sum_{j \neq i} w_{ij} F(x_j) - \beta \sum_{j \neq i} F(x_j) + \text{noise}$$

where $F(x)$ is some activation function. The current simulations use the function $F(x) = \frac{x}{1+x}$. The first term represents exponential forgetting with a time constant of λ_a, the second term activation from other units, and the third term inhibition among items to keep the overall activation in a reasonable range.

The weights w_{ij} are updated using a Hebbian learning rule

$$\dot{w}_{ij} = -\lambda_w w_{ij} + \rho F(x_i) F(x_j)$$

λ_w is the time constant of forgetting (which I set to zero in the current simulations) while ρ is the learning rate.

A discrete version of the activation equation is given by

$$x_i(t+1) = x_i(t) - \lambda_a x_i(t) + \alpha \sum_{j \neq i} w_{ij} F(x_j) - \beta \sum_{j \neq i} F(x_j) + \text{noise}$$

While the time step is arbitrary in the absence of external input (see [Endress and Szabó, 2020] for a proof), I use the duration of individual units (e.g., syllables, visual symbols etc.) as the time unit in the discretization as associative learning is generally invariant under temporal scaling of the experiment [Gallistel and Gibbon, 2000, Gallistel et al., 2001]. Further, while only excitatory connections are tuned by learning in our model, the same effect could be obtained by tuning inhibition, for example through tunable disinhibitory interneurons [Letzkus et al., 2011]. Here, I simply focus on the result that a fairly generic network architecture accounts for rhythmic network activity in response to statistically structured sequences.

The discrete updating rule for the weights is

$$w_{ij}(t+1) = w_{ij}(t) - \lambda_w w_{ij}(t) + \rho F(x_i) F(x_j)$$

Simulation parameters are listed in Table 1. An *R* implementation is available at XXX.

Table 1: Parameters used in the simulations XXX BEAUTIFY AS IN ORIGINAL PAPER

Name	Value
A	0.7
B	0.4
L_ACT	0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9
L_ACT_DEFAULT	0.5
L_ACT_SAMPLES	0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9
L_W	0
N_ITEMS_BEFORE_ACTIVATION_INSPECTION	600
N_NEURONS	19
N_REP_PER_WORD	100
N_REP_PER_WORD_BURNIN	50
N_SIM	100
N_SYLL_PER_WORD	3
N_WORDS	4
NOISE_SD_ACT	0.001
NOISE_SD_W	0
R	0.05

5.2 Supplementary Information 2: Detailed results

5.2.1 Activation differences between words and part-words

Table 2 provides detailed results for the simulations in terms of descriptive statistics and statistical tests for the simulation testing the recognition of words and part-words.

Table 2: Detailed results for the different forgetting rates and the word vs. part-word comparison (*ABC* vs. *BC:D* and *ABC* vs. *C:DE*), using the global activation as a measure of the network's familiarity with the items. $p_{Wilcoxon}$ represents the *p* value of a Wilcoxon test on the difference scores against the chance level of zero. $P_{Simulations}$ represents the proportion of simulations showing positive difference scores.

λ_a	Statistic	ABC vs BC:D	ABC vs C:DE
100×10^{-3}	M	-12.5×10^{-3}	-50.0×10^{-3}
100×10^{-3}	SE	-1.26×10^{-3}	-5.02×10^{-3}
100×10^{-3}	$p_{Wilcoxon}$	985×10^{-3}	68.1×10^{-3}
100×10^{-3}	$P_{Simulations}$	540×10^{-3}	450×10^{-3}
200×10^{-3}	M	-7.60×10^{-3}	-45.2×10^{-3}
200×10^{-3}	SE	-764×10^{-6}	-4.54×10^{-3}
200×10^{-3}	$p_{Wilcoxon}$	649×10^{-3}	94.4×10^{-3}
200×10^{-3}	$P_{Simulations}$	530×10^{-3}	480×10^{-3}
300×10^{-3}	M	1.72×10^{-3}	-39.2×10^{-3}
300×10^{-3}	SE	173×10^{-6}	-3.94×10^{-3}
300×10^{-3}	$p_{Wilcoxon}$	114×10^{-3}	627×10^{-3}
300×10^{-3}	$P_{Simulations}$	420×10^{-3}	600×10^{-3}
400×10^{-3}	M	-4.52×10^{-3}	71.5×10^{-3}
400×10^{-3}	SE	-454×10^{-6}	7.19×10^{-3}
400×10^{-3}	$p_{Wilcoxon}$	76.7×10^{-18}	3.96×10^{-18}
400×10^{-3}	$P_{Simulations}$	90.0×10^{-3}	1.00
500×10^{-3}	M	3.64×10^{-3}	63.3×10^{-3}
500×10^{-3}	SE	366×10^{-6}	6.36×10^{-3}
500×10^{-3}	$p_{Wilcoxon}$	468×10^{-18}	3.96×10^{-18}
500×10^{-3}	$P_{Simulations}$	890×10^{-3}	1.00
600×10^{-3}	M	7.42×10^{-3}	50.5×10^{-3}
600×10^{-3}	SE	746×10^{-6}	5.07×10^{-3}
600×10^{-3}	$p_{Wilcoxon}$	3.96×10^{-18}	3.96×10^{-18}
600×10^{-3}	$P_{Simulations}$	1.00	1.00
700×10^{-3}	M	9.46×10^{-3}	28.0×10^{-3}
700×10^{-3}	SE	951×10^{-6}	2.82×10^{-3}
700×10^{-3}	$p_{Wilcoxon}$	3.96×10^{-18}	3.96×10^{-18}
700×10^{-3}	$P_{Simulations}$	1.00	1.00
800×10^{-3}	M	9.49×10^{-3}	17.3×10^{-3}
800×10^{-3}	SE	954×10^{-6}	1.74×10^{-3}
800×10^{-3}	$p_{Wilcoxon}$	3.96×10^{-18}	3.96×10^{-18}
800×10^{-3}	$P_{Simulations}$	1.00	1.00
900×10^{-3}	M	5.39×10^{-3}	7.56×10^{-3}
900×10^{-3}	SE	542×10^{-6}	760×10^{-6}
900×10^{-3}	$p_{Wilcoxon}$	3.96×10^{-18}	3.96×10^{-18}
900×10^{-3}	$P_{Simulations}$	1.00	1.00

Table 2: Detailed results for the different forgetting rates and the word vs. part-word comparison (*ABC* vs. *BC:D* and *ABC* vs. *C:DE*), using the global activation as a measure of the network's familiarity with the items. $p_{Wilcoxon}$ represents the *p* value of a Wilcoxon test on the difference scores against the chance level of zero. $P_{Simulations}$ represents the proportion of simulations showing positive difference scores. *(continued)*

λ_a	Statistic	ABC vs BC:D	ABC vs C:DE

Table 3: Difference scores between syllable activations in different positions. P values reflect a Wilcoxon test against the chance level of zero.

	$\sigma_2 - \sigma_1$			$\sigma_3 - \sigma_2$			$\sigma_3 - \sigma_1$		
Λ	M	SE	p	M	SE	p	M	SE	p
0.4	-0.1764595	0.0007211	0	0.0695458	0.0008207	0	-0.1069137	0.0013255	0
0.5	-0.1845433	0.0004358	0	0.1691245	0.0002991	0	-0.0154188	0.0001861	0
0.6	-0.0887139	0.0002330	0	0.1435532	0.0002557	0	0.0548392	0.0000765	0
0.7	0.0121462	0.0000536	0	0.0667430	0.0001267	0	0.0788892	0.0000831	0
0.8	0.0236909	0.0000262	0	0.0322496	0.0000619	0	0.0559405	0.0000507	0
0.9	0.0197990	0.0000217	0	0.0075760	0.0000247	0	0.0273750	0.0000228	0

5.2.2 Electrophysiological results

5.2.2.1 Activation differences within words. Table 3 shows activation differences between pairs of neurons in different positions within words.

Table 4: Number of simultaneously active neurons as a function of the forgetting rate.

Λ	M	SE
0.1	3.790	0.058
0.2	4.006	0.028
0.3	3.611	0.006
0.4	3.198	0.002
0.5	2.996	0.001
0.6	2.499	0.001
0.7	2.030	0.000
0.8	2.000	0.000
0.9	2.000	0.000

5.2.2.2 Number of active neurons Table 4 and Figure 8 show the number of simultaneously active neurons as a function of the forgetting rate.

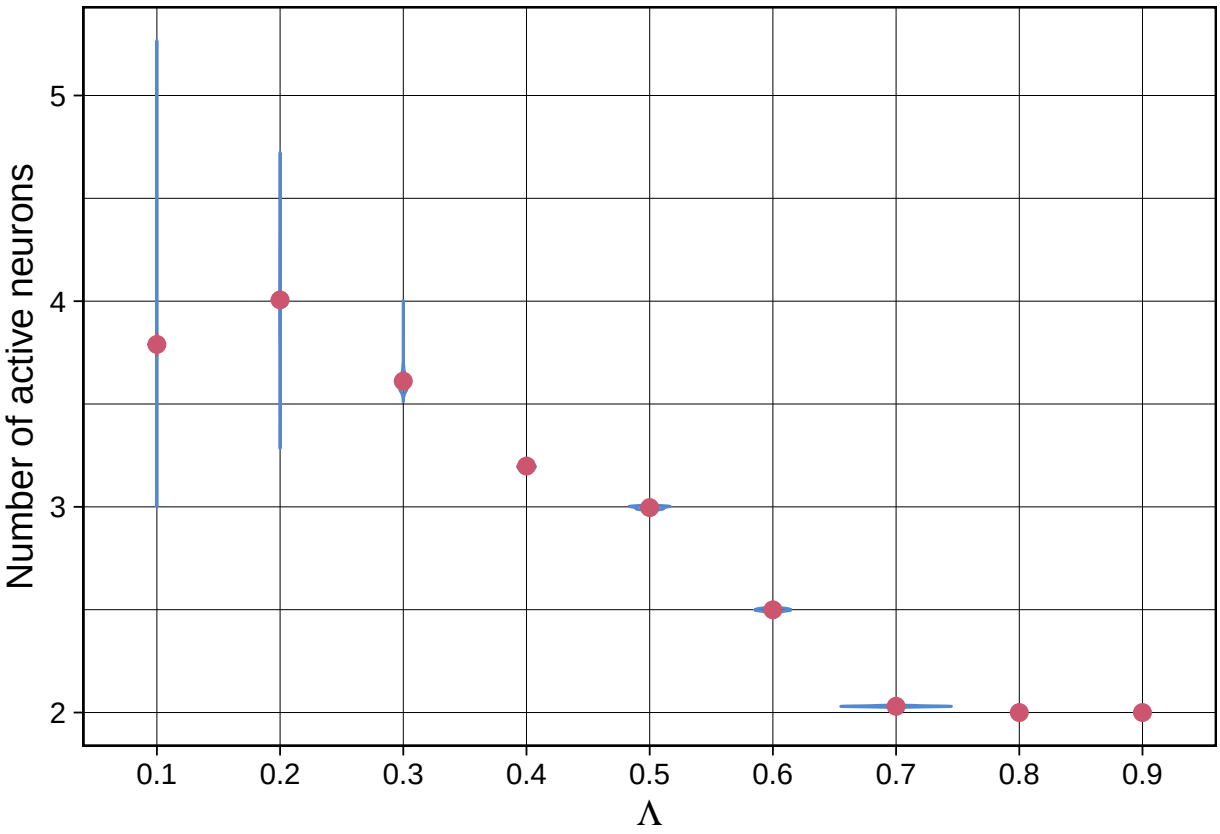

Figure 8: Average number of simultaneously active neurons as a function of the forgetting rate.

Table 5: Relative phases of the network activation relative to different syllable positions in degrees.

	Phase in degrees relative to							
	σ_1		σ_2		σ_3		Saw tooth	
Λ	M	SE	M	SE	M	SE	M	SE
0.4	19.79705	4.8859704	106.13301	8.3828050	-81.304162	5.2661921	-106.35356	6.0032660
0.5	55.77019	0.0411637	175.77019	0.0411637	-64.229806	0.0411637	-94.22981	0.0411637
0.6	82.00547	0.0427665	-157.99453	0.0427665	-37.994529	0.0427665	-67.99453	0.0427665
0.7	128.47145	0.0476116	-111.52855	0.0476116	8.471447	0.0476116	-21.52855	0.0476116
0.8	145.05548	0.0439595	-94.94452	0.0439595	25.055480	0.0439595	-4.94452	0.0439595
0.9	164.59035	0.0490270	-75.40966	0.0490270	44.590344	0.0490270	14.59034	0.0490270

5.2.3 Spectral density

5.2.4 Phase analysis

Table 5 shows the descriptives of the relative phase of the network activation relative to the different syllable positions.

Table 6: Activation difference between items composed of the first two syllables of a word and the last two syllables of a word, when these bigrams were presented in isolation. Positive values indicate greater activation for the AB items than the BC items. The p value reflects a two-sided Wilcoxon signed rank test against the chance level of zero

	Activation difference between AB and BC items		
Λ	M	SE	p
0.1	-0.0250175	0.6970091	0.4671000
0.2	0.2191508	0.3494696	0.5976501
0.3	-0.1952755	0.0335688	0.0000000
0.4	-0.1590318	0.0031236	0.0000000
0.5	-0.0651194	0.0004283	0.0000000
0.6	-0.0367384	0.0003334	0.0000000
0.7	-0.0190492	0.0002921	0.0000000
0.8	-0.0105219	0.0003090	0.0000000
0.9	-0.0030480	0.0002854	0.0000000

5.2.5 Memory for word-onsets vs. offsets

Table 6 shows the descriptives of activation differences between syllable bigrams at word onsets and word offsets, respectively.

5.2.6 Positional similarity (NEW IN REVISION)

Table 7 shows the descriptives of the relative difference in cosine similarity for pairs of representations mismatching and matching in the their sequential positions, respectively

5.2.7 Heard vs. unheard items [Fló et al., 2022] (NEW IN REVISION)

While the current model provides a qualitative explanation of a number of statistical learning results, it is unlikely to make quantitative predictions. A case in point are some test items used by [Fló et al., 2022]. Specifically, they sought ERP responses to TP violations by pitting hear items against non-heard items.

The heard items were words (of the form $A_iB_iC_i$) and part-words (of the form $B_iC_iA_k$). The unheard items were edge-words (of the form $A_iB_iC_k$) and non-words (of the form $B_iC_iA_i$). While [Fló et al., 2022] did not observe significant ERP differences between these item types, modeling this null effect likely requires quantitative predictions of different cues that might drive such differences (apart from issues associated with modeling null effects).

Specifically, while the forward and backward TPs are stronger in heard items than in un-heard items, this description of the test items depends on calculating TPs at a constant lag. For example, in edge-words ($A_iB_iC_k$), the C_k syllable is associated with the other syllables, given that participants encountered part-words such as $C_kA_iB_i$. Likewise, in non-words ($B_iC_iA_i$), all syllables are strongly associated with one another, given that the item is a scrambled version of a word $A_iB_iC_i$. Depending on how learners integrate associations across directions (forward or backward) and lags, they might or might not differentiate [Fló et al., 2022] heard items from their unheard items.

In contrast to [Fló et al., 2022] results, Figure 9 and Table 8 show that, with the current parameter set, the network prefers heard items over unheard items for forgetting rates of at least 0.4. However, these results likely depend on the interplay between forgetting, excitation, interference as well as the structure of the speech stream.

Table 7: Simulations of positional codes citeHenin2021. Representations of syllables *not* sharing a sequential position (i.e., word-initial, word-medial, word-final) are more similar to each other than representations of syllables sharing a sequential position. The similarity was evaluated using the cosine similarity across the vectors of activation elicited by the syllables. The p value reflects a Wilcoxon test against the chance level of zero.

	Relative cosine similarity difference $\frac{Non-match - Match}{Non-match + Match}$		
Λ	M	SE	p
0.1	0.0000423	0.0000056	0
0.2	0.0000808	0.0000100	0
0.3	0.0042055	0.0017344	0
0.4	0.4565052	0.0019912	0
0.5	0.6509563	0.0003644	0
0.6	0.7219907	0.0002754	0
0.7	0.8020529	0.0001526	0
0.8	0.8815974	0.0001024	0
0.9	0.9474764	0.0000495	0

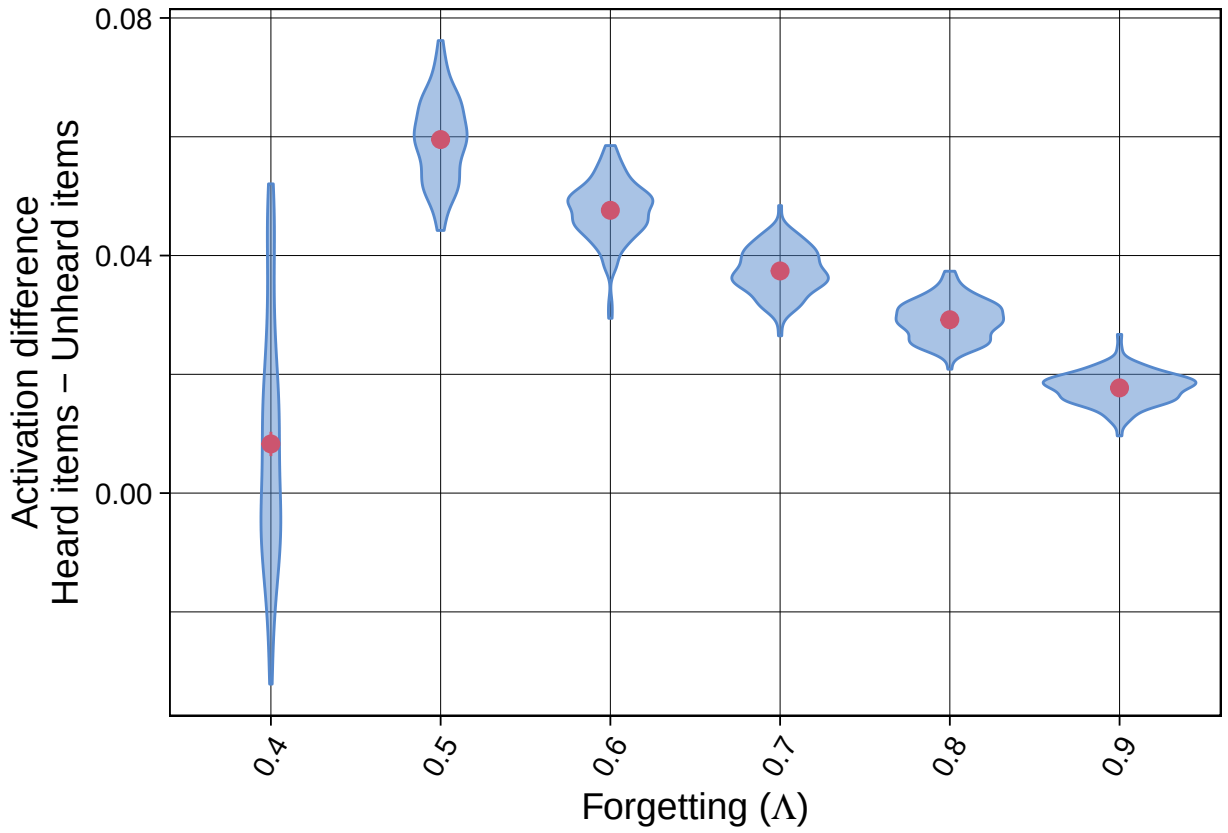

Figure 9: Average difference in the total network activation for the heard item ($A_iB_iC_i$ and $B_iC_iA_k$) vs. unheard item contrast ($A_iB_iC_k$; $B_iC_iA_i$) used by citeFlo2022. The results reflect the network behavior after the first 50 presentations of each word. Positive values indicate greater activation for the heard items than the unheard items.

Table 8: Average difference in the total network activation for the heard item ($A_iB_iC_i$ and $B_iC_iA_k$) vs. unheard item contrast ($A_iB_iC_k$; $B_iC_iA_i$) used by citeFlo2022. The results reflect the network behavior after the first 50 presentations of each word. Positive values indicate greater activation for the heard items than the unheard items. The p value reflects a two-sided Wilcoxon signed rank test against the chance level of zero

	Activation difference between heard items and unheard items		
Λ	M	SE	p
0.1	-1.853	2.772	0.641
0.2	0.224	1.676	0.619
0.3	-0.196	0.267	0.315
0.4	0.008	0.002	0.001
0.5	0.060	0.001	0.000
0.6	0.048	0.001	0.000
0.7	0.037	0.000	0.000
0.8	0.029	0.000	0.000
0.9	0.018	0.000	0.000

References

Dilshat Abla, Kentaro Katahira, and Kazuo Okanoya. On-line assessment of statistical learning by event-related potentials. *J Cogn Neurosci*, 20(6):952–964, 2008. doi: 10.1162/jocn.2008.20058.

Richard N Aslin, Jenny R Saffran, and Elissa L Newport. Computation of conditional probability statistics by 8-month-old infants. *Psychol Sci*, 9:321–324, 1998.

Laura J. Batterink and Ken A. Paller. Online neural monitoring of statistical learning. *Cortex*, 90:31–45, May 2017. ISSN 1973-8102. doi: 10.1016/j.cortex.2017.02.004.

Paul M Bays, Victoria Singh-Curry, Nikos Gorgoraptis, Jon Driver, and Masud Husain. Integration of goal- and stimulus-related visual signals revealed by damage to human parietal cortex. *J Neurosci*, 30:5968–5978, 2010. doi: 10.1523/JNEUROSCI.0997-10.2010.

Lucas Benjamin, Ana Fló, Marie Palu, Shruti Naik, Lucia Melloni, and Ghislaine Dehaene-Lambertz. Tracking transitional probabilities and segmenting auditory sequences are dissociable processes in adults and neonates. *Developmental science*, 26:e13300, March 2023. ISSN 1467-7687. doi: 10.1111/desc.13300.

Heather Bortfeld, James L Morgan, Roberta Michnick Golinkoff, and Karen Rathbun. Mommy and me: Familiar names help launch babies into speech-stream segmentation. *Psychol Sci*, 16(4):298–304, 2005. doi: 10.1111/j.0956-7976.2005.01531.x.

MR Brent and JM Siskind. The role of exposure to isolated words in early vocabulary development. *Cognition*, 81(2):B33–44, 2001.

Diane Brentari, Carolina González, Amanda Seidl, and Ronnie Wilbur. Sensitivity to visual prosodic cues in signers and nonsigners. *Lang Speech*, 54(1):49–72, 2011.

Marco Buiatti, Marcela Peña, and Ghislaine Dehaene-Lambertz. Investigating the neural correlates of continuous speech computation with frequency-tagged neuroelectric responses. *Neuroimage*, 44(2):509–519, 2009. doi: 10.1016/j.neuroimage.2008.09.015.

Jiani Chen and Carel Ten Cate. Zebra finches can use positional and transitional cues to distinguish vocal element strings. *Behav Processes*, 117:29–34, 2015. doi: 10.1016/j.beproc.2014.09.004.

Morten H. Christiansen. Implicit statistical learning: A tale of two literatures. *Topics in Cognitive Science*, 11(3):468–481, 2018. doi: 10.1111/tops.12332.

Anne Christophe, Jacques Mehler, and Nuria Sebastian-Galles. Perception of prosodic boundary correlates by newborn infants. *Infancy*, 2(3):385–394, 2001.

Andy Clark. Whatever next? predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, 36(3):181–204, 2013. doi: 10.1017/s0140525x12000477.

N. Cohen and L. Squire. Preserved learning and retention of pattern-analyzing skill in amnesia: dissociation of knowing how and knowing that. *Science*, 210(4466):207–210, 1980. doi: 10.1126/science.7414331.

Sarah C Creel, Elissa L Newport, and Richard N Aslin. Distant melodies: Statistical learning of nonadjacent dependencies in tone sequences. *J Exp Psychol Learn Mem Cogn*, 30(5):1119–30, 2004. doi: 10.1037/0278-7393.30.5.1119.

Toni Cunillera, Juan M. Toro, Nuria Sebastián-Gallés, and Antoni Rodríguez-Fornells. The effects of stress and statistical cues on continuous speech segmentation: an event-related brain potential study. *Brain research*, 1123:168–178, December 2006. ISSN 0006-8993. doi: 10.1016/j.brainres.2006.09.046.

Ansgar D. Endress. Learning melodies from non-adjacent tones. *Acta Psychologica*, 135(2):182–190, 2010. doi: 10.1016/j.actpsy.2010.06.005.

Ansgar D. Endress and L. L. Bonatti. Words, rules, and mechanisms of language acquisition. *Wiley Interdisciplinary Reviews: Cognitive Science*, 7(1):19–35, 2016. doi: 10.1002/wcs.1376.

Ansgar D. Endress and Luca L. Bonatti. Rapid learning of syllable classes from a perceptually continuous speech stream. *Cognition*, 105(2):247–299, 2007. doi: 10.1016/j.cognition.2006.09.010.

Ansgar D. Endress and Maureen de Seyssel. The specificity of sequential statistical learning: Statistical learning accumulates predictive information from unstructured input but is dissociable from (declarative) memory. *JEPG:G*, under review.

Ansgar D. Endress and Marc D. Hauser. Word segmentation with universal prosodic cues. *Cognit Psychol*, 61 (2):177–199, 2010. doi: 10.1016/j.cogpsych.2010.05.001.

Ansgar D Endress and S P Johnson. When forgetting fosters learning: A neural network model for statistical learning. *Cognition*, 104621, 2021. doi: 10.1016/j.cognition.2021.104621.

Ansgar D. Endress and A Langus. Transitional probabilities count more than frequency, but might not be used for memorization. *Cognitive Psychology*, 92:37–64, 2017. doi: 10.1016/j.cogpsych.2016.11.004.

Ansgar D. Endress and Jacques Mehler. Primitive computations in speech processing. *Quarterly Journal of Experimental Psychology*, 62(11):2187–2209, 2009a. doi: 10.1080/17470210902783646.

Ansgar D. Endress and Jacques Mehler. The surprising power of statistical learning: When fragment knowledge leads to false memories of unheard words. *Journal of Memory and Language*, 60(3):351–367, 2009b. doi: 10.1016/j.jml.2008.10.003.

Ansgar D. Endress and S. Szabó. Sequential presentation protects memory from catastrophic interference. *Cognit Sci*, 44(5), 2020. doi: 10.1111/cogs.12828.

Ansgar D. Endress and Justin N Wood. From movements to actions: Two mechanisms for learning action sequences. *Cognit Psychol*, 63(3):141–171, 2011. doi: 10.1016/j.cogpsych.2011.07.001.

Ansgar D. Endress, Lauren K. Slone, and Scott P. Johnson. Statistical learning and memory. *Cognition*, 204: 104346, 2020. ISSN 1873-7838. doi: 10.1016/j.cognition.2020.104346.

Lucy C. Erickson, Erik D. Thiessen, and Katharine Graf Estes. Statistically coherent labels facilitate categorization in 8-month-olds. *Journal of Memory and Language*, 72:49–58, 2014. doi: 10.1016/j.jml.2014 .01.002.

Amy S. Finn, Priya B. Kalra, Calvin Goetz, Julia A. Leonard, Margaret A. Sheridan, and John D.E. Gabrieli. Developmental dissociation between the maturation of procedural memory and declarative memory. *Journal of Experimental Child Psychology*, 142:212–220, 2016. doi: 10.1016/j.jecp.2015.09.027.

József Fiser and Richard N Aslin. Statistical learning of new visual feature combinations by infants. *Proc Natl Acad Sci U S A*, 99(24):15822–6, 2002a. doi: 10.1073/pnas.232472899.

József Fiser and Richard N Aslin. Statistical learning of higher-order temporal structure from visual shape sequences. *J Exp Psychol Learn Mem Cogn*, 28(3):458–67, 2002b.

József Fiser and Richard N Aslin. Encoding multielement scenes: statistical learning of visual feature hierarchies. *J Exp Psychol Gen*, 134(4):521–37, 2005. doi: 10.1037/0096-3445.134.4.521.

Ana Fló, Lucas Benjamin, Marie Palu, and Ghislaine Dehaene-Lambertz. Sleeping neonates track transitional probabilities in speech but only retain the first syllable of words. *Scientific reports*, 12:4391, March 2022. ISSN 2045-2322. doi: 25865749.

Karl Friston. The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, 11(2):127–138, 2010. doi: 10.1038/nrn2787.

C. R. Gallistel and J Gibbon. Time, rate, and conditioning. *Psychol Rev*, 107(2):289–344, 2000.

C. R. Gallistel, T A Mark, A P King, and P E Latham. The rat approximates an ideal detector of changes in rates of reward: implications for the law of effect. *Journal of experimental psychology. Animal behavior processes*, 27:354–372, 2001. ISSN 0097-7403.

J Gillette, Henry Gleitman, Lila R Gleitman, and A Lederer. Human simulations of vocabulary learning. *Cognition*, 73(2):135–76, 1999.

Arit Glicksohn and Asher Cohen. The role of gestalt grouping principles in visual statistical learning. *Atten Percept Psychophys*, 73(3):708–713, 2011. doi: 10.3758/s13414-010-0084-4.

Jacqueline Gottlieb. From thought to action: the parietal cortex as a bridge between perception, action, and cognition. *Neuron*, 53:9–16, 2007. ISSN 0896-6273.

Peter Graf and George Mandler. Activation makes words more accessible, but not necessarily more retrievable. *Journal of Verbal Learning and Verbal Behavior*, 23(5):553–568, 1984. doi: 10.1016/s0022-5371(84)90346-3.

Katharine Graf-Estes, Julia L Evans, Martha W Alibali, and Jenny R Saffran. Can infants map meaning to newly segmented words? Statistical segmentation and word learning. *Psychol Sci*, 18(3):254–60, 2007. doi: 10.1111/j.1467-9280.2007.01885.x.

Catalin Grigoras. Digital audio recording analysis: the electric network frequency (enf) criterion. *International Journal of Speech, Language and the Law*, 12(1):63–76, Feb. 2005. doi: 10.1558/sll.2005.12.1.63. URL https://journal.equinoxpub.com/IJSLL/article/view/9977.

Marc D Hauser, Elissa L Newport, and Richard N Aslin. Segmentation of the speech stream in a non-human primate: Statistical learning in cotton-top tamarins. *Cognition*, 78(3):B53–64, 2001.

Jessica F. Hay, Bruna Pelucchi, Katharine Graf Estes, and Jenny R. Saffran. Linking sounds to meanings: infant statistical learning in a natural language. *Cogn Psychol*, 63(2):93–106, 2011. doi: 10.1016/j.cogpsych .2011.06.002.

Simon Henin, Nicholas B. Turk-Browne, Daniel Friedman, Anli Liu, Patricia Dugan, Adeen Flinker, Werner Doyle, Orrin Devinsky, and Lucia Melloni. Learning hierarchical sequence representations across human cortex and hippocampus. *Science advances*, 7, February 2021. ISSN 2375-2548. doi: 10.1126/sciadv.abc4530.

Erin S. Isbilen, Stewart M. McCauley, Evan Kidd, and Morten H. Christiansen. Statistically induced chunking recall: A memory-based approach to statistical learning. *Cognitive science*, 44:e12848, 2020. ISSN 1551-6709. doi: 10.1111/cogs.12848.

Elizabeth K Johnson and Peter W. Jusczyk. Word segmentation by 8-month-olds: When speech cues count more than statistics. *J Mem Lang*, 44(4):548–567, 2001.

Elizabeth K Johnson and Amanda H Seidl. At 11 months, prosody still outranks statistics. *Dev Sci*, 12(1): 131–41, 2009. doi: 10.1111/j.1467-7687.2008.00740.x.

Elizabeth K Johnson and Michael D Tyler. Testing the limits of statistical learning for word segmentation. *Dev Sci*, 13(2):339–45, 2010. doi: 10.1111/j.1467-7687.2009.00886.x.

Jason Jones and Harold Pashler. Is the mind inherently forward looking? comparing prediction and retrodiction. *Psychonomic Bulletin & Review*, 14:295–300, 2007. ISSN 1069-9384. doi: 10.3758/bf03194067.

C. Kabdebon, M. Pena, M. Buiatti, and G. Dehaene-Lambertz. Electrophysiological evidence of statistical learning of long-distance dependencies in 8-month-old preterm and full-term infants. *Brain and language*, 148:25–36, September 2015. ISSN 1090-2155. doi: 10.1016/j.bandl.2015.03.005.

Ferhat Karaman and Jessica F. Hay. The longevity of statistical learning: When infant memory decays, isolated words come to the rescue. *J. Exp. Psychol. Learn. Mem. Cogn.*, 44(2):221–232, 2018. doi: 10.1037/xlm0000448.

Georg B. Keller and Thomas D. Mrsic-Flogel. Predictive processing: A canonical cortical computation. *Neuron*, 100(2):424–435, 2018. doi: 10.1016/j.neuron.2018.10.003.

Natasha Z Kirkham, Jonathan A Slemmer, and Scott P Johnson. Visual statistical learning in infancy: evidence for a domain general learning mechanism. *Cognition*, 83(2):B35–B42, 2002. doi: 10.1016/s0010-0277(02)00004-5.

B J Knowlton, J A Mangels, and L R Squire. A neostriatal habit learning system in humans. *Science*, 273: 1399–1402, 1996. ISSN 0036-8075.

Noriko Kudo, Yulri Nonaka, Noriko Mizuno, Katsumi Mizuno, and Kazuo Okanoya. On-line statistical segmentation of a non-speech auditory stream in neonates as demonstrated by event-related brain potentials. *Dev Sci*, 14(5):1100–1106, 2011. doi: 10.1111/j.1467-7687.2011.01056.x.

Kutas and Federmeier. Electrophysiology reveals semantic memory use in language comprehension. *Trends Cogn Sci*, 4(12):463–470, 2000.

Johannes J Letzkus, Steffen B E Wolff, Elisabeth M M Meyer, Philip Tovote, Julien Courtin, Cyril Herry, and Andreas Lüthi. A disinhibitory microcircuit for associative fear learning in the auditory cortex. *Nature*, 480:331–335, 2011. ISSN 1476-4687. doi: 10.1038/nature10674.

Roger Levy. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177, 2008. doi: 10.1016/j.cognition.2007.05.006.

Erika Marchetto and Luca L. Bonatti. Words and possible words in early language acquisition. *Cognit. Psychol.*, 67(3):130 – 150, 2013. ISSN 0010-0285. doi: 10.1016/j.cogpsych.2013.08.001.

James M. McQueen. Segmentation of continuous speech using phonotactics. *J Mem Lang*, 39(1):21–46, 1998.

Tamara Nicol Medina, Jesse Snedeker, John C. Trueswell, and Lila R. Gleitman. How words can and cannot be learned by observation. *Proc Natl Acad Sci U S A*, 108(22):9014–9019, 2011. doi: 10.1073/pnas.1105040108.

Karima Mersad and Thierry Nazzi. When mommy comes to the rescue of statistics: Infants combine top-down and bottom-up cues to segment speech. *Language Learning and Development*, 8(3):303–315, 2012. doi: 10.1080/15475441.2011.609106.

Padraic Monaghan and Morten H. Christiansen. Words in puddles of sound: modelling psycholinguistic effects in speech segmentation. *J Child Lang*, 37(3):545–564, 2010. doi: 10.1017/S0305000909990511.

Christine N. Moreau, Marc F. Joanisse, Jerrica Mulgrew, and Laura J. Batterink. No statistical learning advantage in children over adults: Evidence from behaviour and neural entrainment. *Developmental cognitive neuroscience*, 57:101154, September 2022. ISSN 1878-9307. doi: 10.1016/j.dcn.2022.101154.

Emily Morgan, Allison Fogel, Anjali Nair, and Aniruddh D Patel. Statistical learning and gestalt-like principles predict melodic expectations. *Cognition*, 189:23–34, 2019. ISSN 1873-7838. doi: 10.1016/j.cognition.2018.12.015.

Julia Moser, Laura Batterink, Yiwen Li Hegner, Franziska Schleger, Christoph Braun, Ken A. Paller, and Hubert Preissl. Dynamics of nonlinguistic statistical learning: From neural entrainment to the emergence

of explicit knowledge. *NeuroImage*, 240:118378, October 2021. ISSN 1095-9572. doi: 10.1016/j.neuroimage .2021.118378.

Marcela Peña, Luca L Bonatti, Marina Nespor, and Jacques Mehler. Signal-driven computations in speech processing. *Science*, 298(5593):604–7, 2002. doi: 10.1126/science.1072901.

Pierre Perruchet. What mechanisms underlie implicit statistical learning? transitional probabilities versus chunks in language learning. *Topics in cognitive science*, 11:520–535, July 2019. ISSN 1756-8765. doi: 10.1111/tops.12403.

Pierre Perruchet and Sebastien Pacton. Implicit learning and statistical learning: one phenomenon, two approaches. *Trends in cognitive sciences*, 10:233–238, 2006. ISSN 1364-6613. doi: 10.1016/j.tics.2006.03.006.

Robert Pilon. Segmentation of speech in a foreign language. *J. Psycholinguist. Res.*, 10(2):113 – 122, 1981. ISSN 0090-6905.

R A Poldrack, J Clark, E J Paré-Blagoev, D Shohamy, J Creso Moyano, C Myers, and M A Gluck. Interactive memory systems in the human brain. *Nature*, 414:546–550, 2001. ISSN 0028-0836. doi: 10.1038/35107080.

Edwin M. Robertson. Memory leaks: information shared across memory systems. *Trends in cognitive sciences*, 26:544–554, 2022. doi: 10.1016/j.tics.2022.03.010.

Chantal Roggeman, Wim Fias, and Tom Verguts. Salience maps in parietal cortex: imaging and computational modeling. *NeuroImage*, 52:1005–1014, 2010. ISSN 1095-9572. doi: 10.1016/j.neuroimage.2010.01.060.

Jenny R. Saffran and G. J. Griepentrog. Absolute pitch in infant auditory learning: evidence for developmental reorganization. *Dev Psychol*, 37(1):74–85, 2001.

Jenny R Saffran, Richard N Aslin, and Elissa L Newport. Statistical learning by 8-month-old infants. *Science*, 274(5294):1926–8, 1996a.

Jenny R. Saffran, Elissa L Newport, and Richard N Aslin. Word segmentation: The role of distributional cues. *J Mem Lang*, 35:606–21, 1996b.

Jenny R Saffran, EK Johnson, Richard N Aslin, and Elissa L Newport. Statistical learning of tone sequences by human infants and adults. *Cognition*, 70(1):27–52, 1999.

Lisa D. Sanders, Elissa L. Newport, and Helen J. Neville. Segmenting nonsense: an event-related potential index of perceived onsets in continuous speech. *Nature neuroscience*, 5:700–703, July 2002. ISSN 1097-6256. doi: 10.1038/nn873.

Amanda Seidl and Elizabeth K Johnson. Boundary alignment enables 11-month-olds to segment vowel initial words from speech. *J Child Lang*, 35(1):1–24, 2008.

Rakesh Sengupta, Bapi Raju Surampudi, and David Melcher. A visual sense of number emerges from the dynamics of a recurrent on-center off-surround neural network. *Brain Res*, 1582:114–124, 2014. doi: 10.1016/j.brainres.2014.03.014.

Brynn E. Sherman and Nicholas B. Turk-Browne. Statistical prediction of the future impairs episodic encoding of the present. *Proceedings of the National Academy of Sciences of the United States of America*, 117: 22760–22770, 2020. ISSN 1091-6490. doi: 10.1073/pnas.2013291117.

Amber Shoaib, Tianlin Wang, Jessica F. Hay, and Jill Lany. Do infants learn words from statistics? evidence from English-learning infants hearing Italian. *Cognitive Science*, 42(8):3083–3099, 2018. doi: 10.1111/cogs.12673.

Mohinish Shukla, Marina Nespor, and Jacques Mehler. An interaction between prosody and statistics in the segmentation of fluent speech. *Cognit Psychol*, 54(1):1–32, 2007. doi: 10.1016/j.cogpsych.2006.04.002.

Mohinish Shukla, Katherine S White, and Richard N Aslin. Prosody guides the rapid mapping of auditory word forms onto visual objects in 6-mo-old infants. *Proc Natl Acad Sci U S A*, 108(15):6038–6043, 2011. doi: 10.1073/pnas.1017617108.

Juwairia Sohail and Elizabeth K. Johnson. How transitional probabilities and the edge effect contribute to listeners' phonological bootstrapping success. *Language Learning and Development*, pages 1–11, 2016. doi: 10.1080/15475441.2015.1073153.

Larry R. Squire. Memory and the hippocampus: A synthesis from findings with rats, monkeys, and humans. *Psychological Review*, 99(2):195–231, 1992. doi: 10.1037/0033-295x.99.2.195.

Tuomas Teinonen, Vineta Fellman, Risto Näätänen, Paavo Alku, and Minna Huotilainen. Statistical language learning in neonates revealed by event-related brain potentials. *BMC Neurosci*, 10:21, 2009. doi: 10.1186/1471-2202-10-21.

Jan Theeuwes, Louisa Bogaerts, and Dirk van Moorselaar. What to expect where and when: how statistical learning drives visual selection. *Trends in cognitive sciences*, July 2022. ISSN 1879-307X. doi: 10.1016/j.tics.2022.06.001.

Juan M Toro, Josep B Trobalon, and Núria Sebastián-Gallés. Effects of backward speech and speaker variability in language discrimination by rats. *J Exp Psychol Anim Behav Process*, 31(1):95–100, 2005. doi: 10.1037/0097-7403.31.1.95.

J. C. Trueswell, I. Sekerina, N. M. Hill, and M. L. Logrip. The kindergarten-path effect: studying on-line sentence processing in young children. *Cognition*, 73(2):89–134, 1999.

Nicholas B Turk-Browne and Brian J Scholl. Flexible visual statistical learning: Transfer across space and time. *J Exp Psychol: Hum Perc Perf*, 35(1):195–202, 2009.

Nicholas B Turk-Browne, Justin Jungé, and Brian J Scholl. The automaticity of visual statistical learning. *J Exp Psychol Gen*, 134(4):552–64, 2005. doi: 10.1037/0096-3445.134.4.552.

Nicholas B Turk-Browne, Brian J Scholl, Marcia K Johnson, and Marvin M Chun. Implicit perceptual anticipation triggered by statistical learning. *Journal of neuroscience*, 30:11177–11187, 2010. ISSN 1529-2401. doi: 10.1523/JNEUROSCI.0858-10.2010.

Niels J. Verosky and Emily Morgan. Pitches that wire together fire together: Scale degree associations across time predict melodic expectations. *Cognitive science*, 45:e13037, 2021. ISSN 1551-6709. doi: 10.1111/cogs.13037.