Hebbian learning can explain rhythmic entrainment to statistical regularities

Ansgar D. Endress, City, University of London
Ana Fló, Cognitive Neuroimaging Unit, CNRS ERL 9003, INSERM U992, CEA, Université Paris-Saclay

Abstract

THIS IS THE ABSTRACT FROM SOME OLD PAPER. IGNORE In many domains, organisms need to split continuous signals into sequences of recurring units. For example, during language acquisition, humans need to split fluent speech into its underlying words. One prominent candidate mechanism involves computation of co-occurrence statistics such Transitional Probabilities (TPs). TPs indicate how predictive items are of each other. For example, items such as syllables are more predictive of each other when they are part of the same unit (i.e., word) than when they come from different units. TP computations are surprisingly flexible and sophisticated. Humans are sensitive to (1) forward and backward TPs, (2) TPs between adjacent items and longer-distance TPs and (3) recognize TPs in both known and novel units. Here, we show that a simple and biologically plausible model explains these data. We present a network where excitatory interactions are tuned by Hebbian learning and where inhibitory interactions control the overall level of activation. We show that (1) if forgetting is weak, activations are so long-lasting that indiscriminate associations occur among *all* items; (2) if forgetting is strong, activations are so short-lived that they do not persist after the offset of stimuli, and no associations are formed; and (3) for intermediate forgetting rates, this simple network accounts for all of the hallmarks mentioned above. Ironically, forgetting seems to be a key ingredient that enables these sophisticated learning abilities.

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 $P(\sigma_2|\sigma_1)$. In fact, infants, newborns and non-human animals are all sensitive to TPs [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 are first familiarized with a statistically structured speech stream (or a sequence in another modality). The speech stream is a random concatenation of triplets of non-sense syllables (hereafter "words"). For example, if *ABC*, *DEF*, *GHJ* and *KLM* are "words" (where each letter represents a syllable), syllables

within words are more predictive of one another than syllable across word-boundaries. After all, the *C* at the end of *ABC* can be followed by the word-initial syllables of any of the other words. 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 such views often have alternative explanations that do not involve declarative LTM representations (see [Endress et al., 2020, Endress and de Seyssel, under review] for critical reviews). For example, while some high-TP items are sometimes easier to memorize [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 facilitate subsequent associations. Likewise, while sub-items of high-TP items are sometimes harder to recognize than entire items, such results can be complained 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 also be dissociable from 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 to 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, 2009]. 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 them. Further, when learners are asked to repeat back the items they have encountered during a familiarization stream, they are unable to do so [Endress and de Seyssel, under review].

However, there is a simple alternative explanation to such results: a sensitivity to TPs might reflect Hebbian learning [Endress, 2010, Endress and Johnson, 2021]. After all, the representation of syllables (or other elements in a stream) presumably does not cease to be active as soon as the syllable ended. As a result, multiple syllables can be active together, 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, if statistical learning really reflects Hebbian, associative learning, it is difficult to see how one can explain the neurophysiological correlates of statistical learning. We will discuss this literature in the next section.

1.2 Electrophysiological correlates of statistical learning

*** TO DO ANA ***

[Batterink and Paller, 2017, Buiatti et al., 2009, Fló et al., 2022, Kabdebon et al., 2015, Moser et al., 2021]

- DISCUSS N400 AND SIMILAR LITERATIRES, THEN ARGUE THAT IT IS ALSO CONSISTENT WITH A NON-MEMORY EXPLANATION IF THE N400 INDEXES SURPRISING SYLLABLES:

IT WOULD NOT INDEX WORD ONSETS, BUT RATHER THE LACK OF PREDICTABILITY AFTER PREDICTABLE SYLLABLES
- DISCUSS ENTRAINMENT LITERATIRE
- LINK THIS DISCUSSION BACK TO THE ISSUE THAT STATISTICAL LEARNING MIGHT BE MORE USEFUL FOR PREDICTIVE PROCESSING. (CAN DO THIS MYSELF)

*** END TO DO ***

2 The current study

Here, we show that such elecrophysiological 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. The network is 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 $ABCD\ldots$, 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 will 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] confirmed these predictions and showed that this 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 rhythmic entrainment results above seem to suggest that learners do more than merely computing associations among syllables. Here, we argue that a simple Hebbian network can account for the periodic activity found in electrophysiological recordings as well. 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, 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, 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.

We tested this idea in [Endress and Johnson, 2021] model. We expose the network to a continuous sequence inspired by [Saffran et al., 1996a] Experiment 2. The sequence consists of 4 distinct words of 3 syllables each. The familiarization sequence is a random concatenations of these words, with each word occurring 100 times. During the test phase, we record 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] We 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

[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 we seek to uncover periodic activity during familiarization.

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), we 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], we 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]

[2]While we use the label "decay", we do not claim that "decay" reflects a psychological processes. Our 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 we observe discrimination between words and part-words (and thus learning), we first repeat some of [Endress and Johnson, 2021] results. We calculate normalized difference scores of activations for words and part-words, $d = \frac{\text{Word}-\text{Part-Word}}{\text{Word}+\text{Part-Word}}$, and evaluate these difference scores in two ways. First, we compare them to the chance level of zero using Wilcoxon tests. Second, we count the number of simulations (representing different participants) preferring words, and to evaluate 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 1 and Table **??**. 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, we will thus use forgetting rates between 0.4 and 0.9 to model the electrophysiological results.

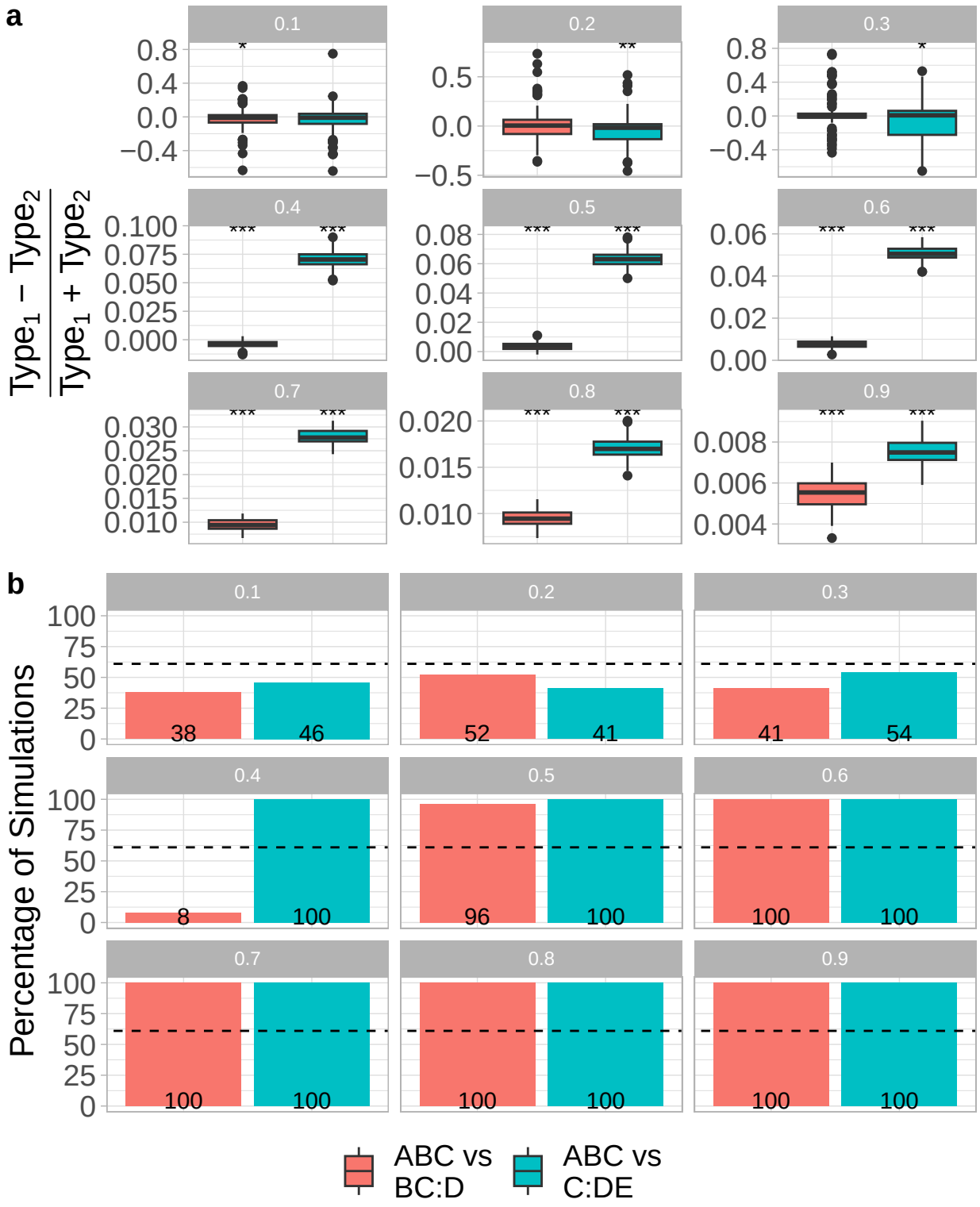

Figure 1: (a) Difference scores between words and part-words for different forgetting rates (between .1 and .9). The scores are calculated based the global activation as a measure of the network's familiarity with the items. Significance is assessed based on Wilcoxon tests against the chance level of zero. (b). Percentage of simulations with a preference for words for different forgetting rates (between .1 and .9). The simulations are assessed based on the global activation in the network. The dashed line shows the minimum percentage of simulations that is significant based on a binomial test.

Table 1: 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	-0.1767709	0.0006546	0	0.0707578	0.0009093	0	-0.1060131	0.0013797	0
0.5	-0.1853090	0.0003695	0	0.1695923	0.0002698	0	-0.0157167	0.0001679	0
0.6	-0.0889537	0.0002555	0	0.1439090	0.0002696	0	0.0549552	0.0000803	0
0.7	0.0120413	0.0000504	0	0.0668421	0.0001194	0	0.0788834	0.0000807	0
0.8	0.0237180	0.0000243	0	0.0322516	0.0000526	0	0.0559696	0.0000467	0
0.9	0.0198095	0.0000204	0	0.0075744	0.0000249	0	0.0273839	0.0000249	0

3.2 Electrophysiological results

3.2.1 Activation differences within words

We next asked whether a basic Hebbian learning model can explain periodic activity found in electrophysiological recordings (XXX), focusing on the forgetting rates in which the network preferred words to part-words. In a first analysis, we simply recorded the total network activation after each syllable in a word has 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 2 and Table 1, activation was highest after word-final syllables (though not for very low forgetting rates for which we 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, however, while previous electrophysiological responses to statistical structured streams were interpreted in terms of a response to word onsets (XXX), our results suggest an alternative interpretation of such results. Rather than signalling the beginnings and ends of words, an activation maximum after the third syllable of each word signals the predictability of the third syllable, while a sudden drop in activation after the first syllable indicates the lack of predictability.

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, we recorded the number of active neurons after a burn-in phase of 600 items. As shown in Table **??** and Figure 6, 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

We next analyzed the frequency response of the network to the speech streams. Specifically, we 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. We 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. This 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

Our 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, we calculated the phase of the network activation with respect to waveforms

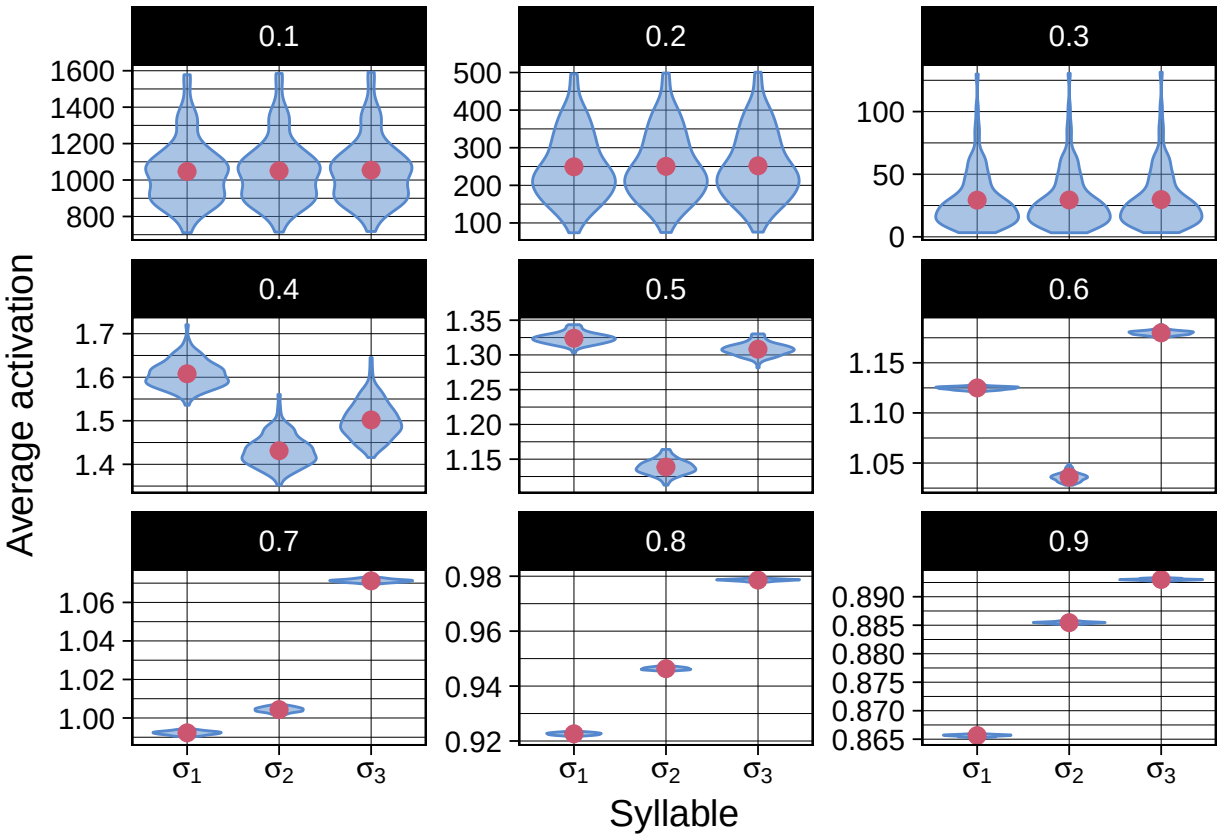

Figure 2: Average total network activation for different syllables in 100 simulations in Endress and Johnson's network during the familiarization with a stream following Saffran et al. (1996). The facets show different forgetting rates. The results reflect the network behavior after the first 20 presentations of each word.

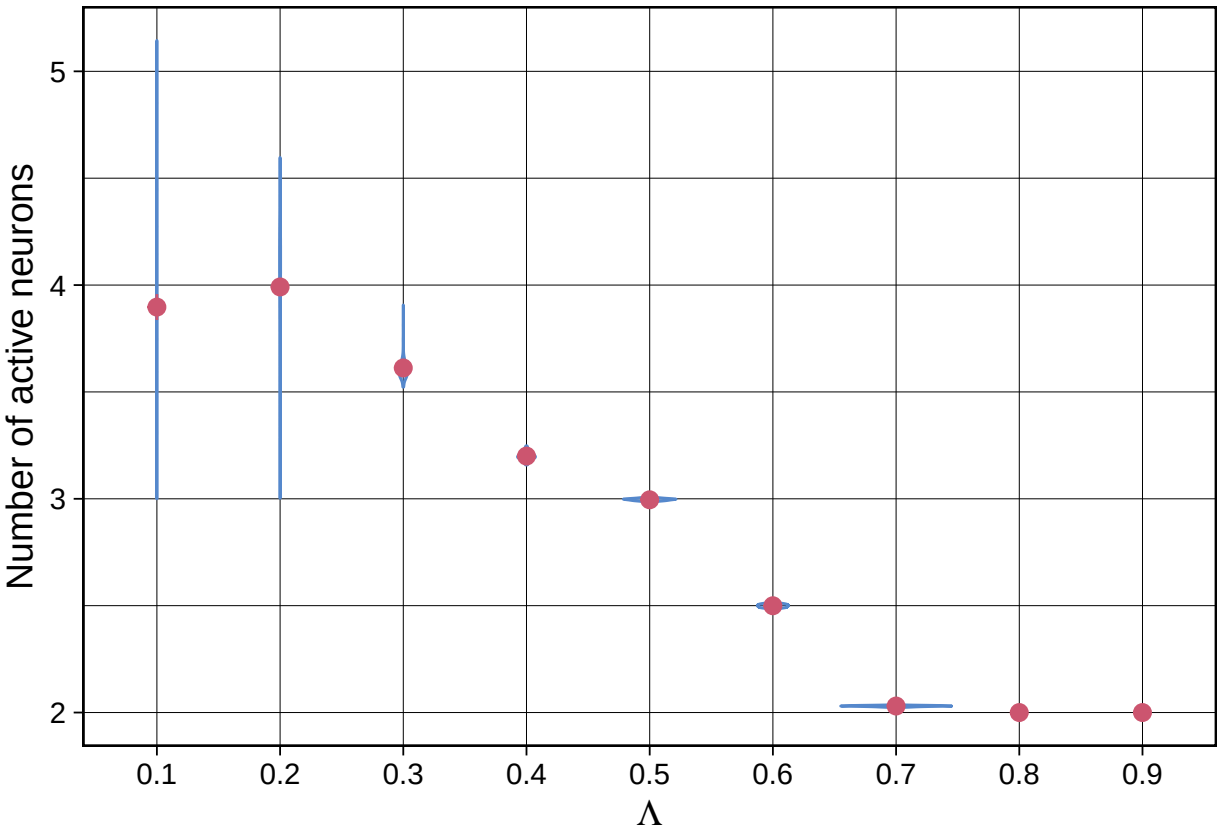

Figure 3: Average number of simultaneously active neurons as a function of the forgetting rate.

Table 2: Difference scores between syllable activations in different positions. P values reflect a wilcoxon

	Phase in degrees relative to							
	σ_1		σ_2		σ_3		Saw tooth	
Λ	M	SE	M	SE	M	SE	M	SE
0.6	82.00705	0.0511961	-157.99295	0.0511961	-37.992953	0.0511961	-67.992952	0.0511961
0.7	128.41318	0.0445116	-111.58682	0.0445116	8.413184	0.0445116	-21.586815	0.0445116
0.8	145.04820	0.0362621	-94.95180	0.0362621	25.048204	0.0362621	-4.951796	0.0362621
0.9	164.56772	0.0473923	-75.43228	0.0473923	44.567719	0.0473923	14.567719	0.0473923

with maximua on word-initial, word-medial and word-final syllables, respectively. Specifically, we 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 sawtooth function with its maximum on the third syllable. As shown in Figure 4(b) and Table~2, the activation had a small relative phase with respect 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 with respect 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 syllabe.

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. Specifically, they presented newborns with items starting with two syllables that occurred word-initially (AB...), and with items starting with a word-medial syllable (BC...) and observed early ERP differences between these items.

To reproduce these results, we measured the activation of the network in response to isolated, bisyllabic AB and BC test items, respectively. As shown in Figure 5 and Table~3, the network activation was always greater in response to BC items than to AB items except for the smallest 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.

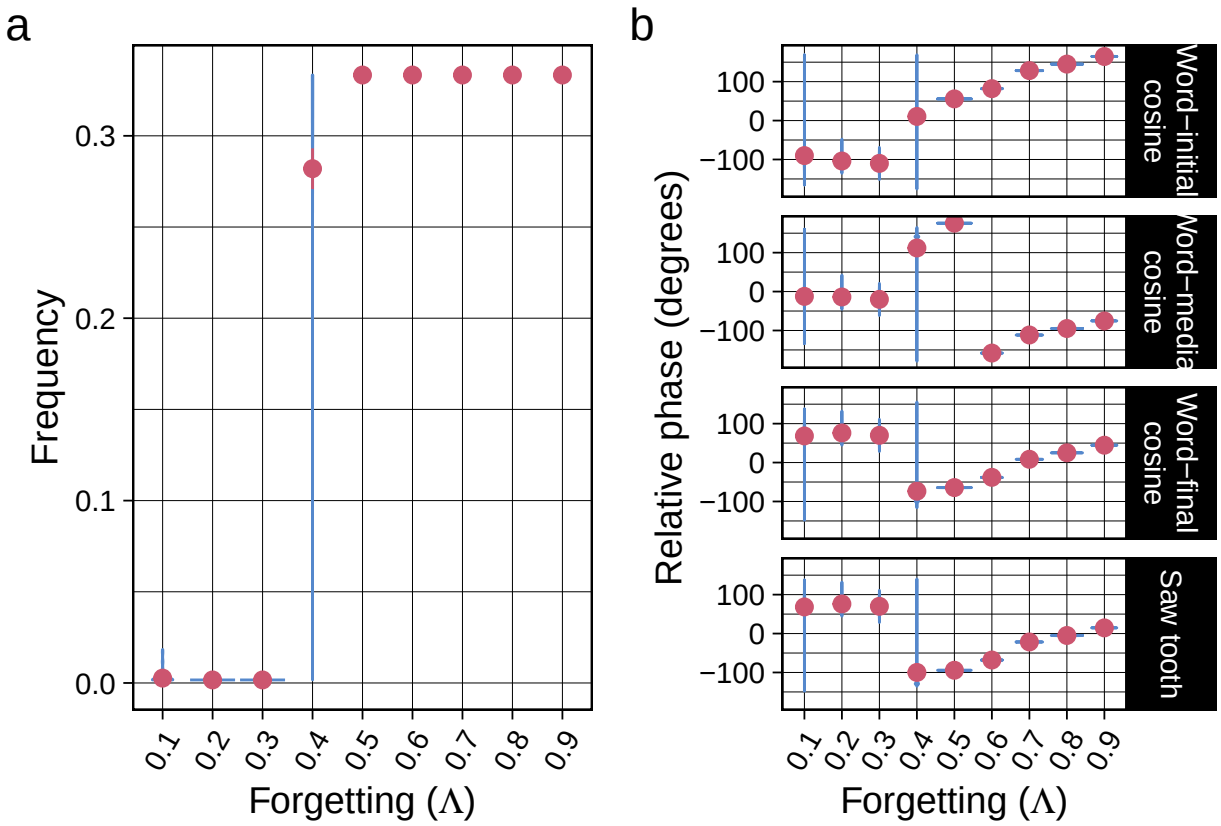

Figure 4: Spectral analysis of the total network activation in 100 simulations in Endress and Johnson's network during the familiarization with a stream following Saffran et al. (1996). The results reflect the network behavior after the first 20 presentations of each word. (a) Maximal frequency as a 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-second syllables and word-final syllables and 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.

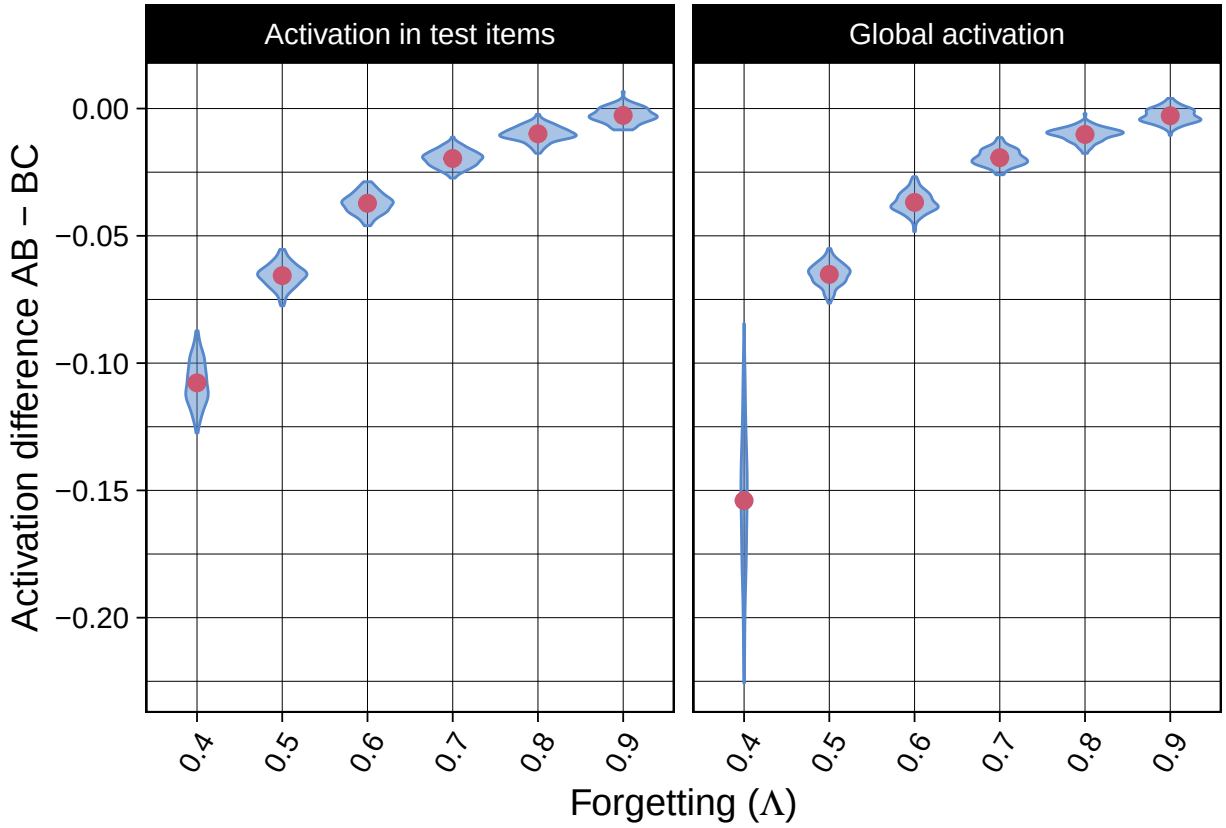

Figure 5: Average difference in the total network activation for the first two syllables of a word (AB) and the first to syllables of a part-word (BC) in 100 simulations in Endress and Johnson's network after familiarization with a stream following Saffran et al. (1996). The results reflect the network behavior after the first 20 presentations of each word. Positive values indicate greater activation for the AB items than the BC items.

Table 3: Activation difference between items composed of the first two items of a word and the last two items 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	0.2826406	0.7313773	0.4280494
0.2	0.1777287	0.4096324	0.7322656
0.3	-0.0686097	0.0461245	0.0002059
0.4	-0.1540030	0.0029380	0.0000000
0.5	-0.0651579	0.0004296	0.0000000
0.6	-0.0368588	0.0004031	0.0000000
0.7	-0.0193043	0.0003164	0.0000000
0.8	-0.0101854	0.0002769	0.0000000
0.9	-0.0028198	0.0002854	0.0000000

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 (XXX); 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 (XXX).

Here, we show that such results can be explained by a simple Hebbian learning model. When exposed to statistically structure 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 activations 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 thus index the statistical predictiveness of syllables rather than the acquisition of words. For example, and as mentioned above, N400 effects observed in statistical learning tasks (XXX) might not index the onset of words, but rather the lack of predictibility of word-initial syllables (or the increased predictibility 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 (XXX).

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, 2009, 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].

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. (Here we use $F(x) = \frac{x}{1+x}$. 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 we set to zero in our 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), we use the duration of individual units (e.g., syllables, visual symbols etc.) as the time unit in our 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, we simply focus on the result that a fairly generic network architecture accounts for the hallmarks of statistical learning that, so far, have eluded explanation.

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 4. An *R* implementation is available at XXX.

Table 4: Parameters used in the simulations

Name	Value
A	0.7
B	0.4
current_l	0.9
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 ?? 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 5: Number of simultaneously active neurons as a function of the forgetting rate.

Λ	M	SE
0.1	3.897	0.061
0.2	3.991	0.029
0.3	3.612	0.006
0.4	3.200	0.002
0.5	2.995	0.001
0.6	2.500	0.001
0.7	2.030	0.000
0.8	2.000	0.000
0.9	2.000	0.000

References

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; a journal devoted to the study of the nervous system and behavior*, 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.

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.

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 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.

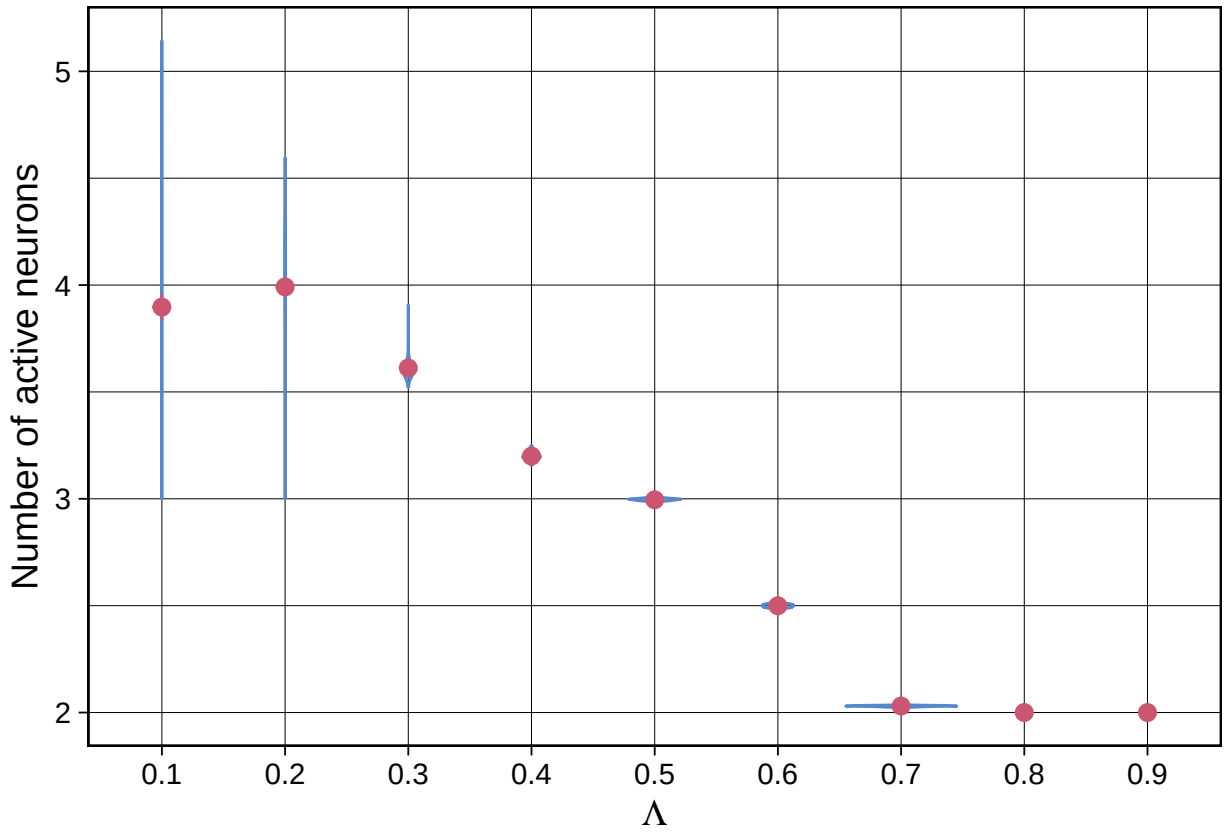

Figure 6: Average number of simultaneously active neurons as a function of the forgetting rate.

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. 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, 2009. 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.

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.

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.

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.

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.

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.

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.

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.

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.

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.