Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

When forgetting fosters learning: A neural network model for statistical learning[☆]

Ansgar D. Endress[a,*], Scott P. Johnson[b]

[a] Department of Psychology, City, University of London, UK
[b] Department of Psychology, UCLA, United States

ARTICLE INFO

Keywords:
Statistical learning
Implicit learning
Transitional probabilities
Neural networks
Chunking

ABSTRACT

Learning often requires splitting continuous signals into recurring units, such as the discrete words constituting fluent speech; these units then need to be encoded in memory. A prominent candidate mechanism involves statistical learning of co-occurrence statistics like transitional probabilities (TPs), reflecting the idea that items from the same unit (e.g., syllables within a word) predict each other better than items from different units. TP computations are surprisingly flexible and sophisticated. Humans are sensitive to forward and backward TPs, compute TPs between adjacent items and longer-distance items, and even recognize TPs in novel units. We explain these hallmarks of statistical learning with a simple model with tunable, Hebbian excitatory connections and inhibitory interactions controlling the overall activation. With weak forgetting, activations are long-lasting, yielding associations among all items; with strong forgetting, no associations ensue as activations do not outlast stimuli; with intermediate forgetting, the network reproduces the hallmarks above. Forgetting thus is a key determinant of these sophisticated learning abilities. Further, in line with earlier dissociations between statistical learning and memory encoding, our model reproduces the hallmarks of statistical learning in the absence of a memory store in which items could be placed.

1. Introduction

Observers often need to segment continuous signals into discrete recurring units, from the recognition of meaningful actions, where observers need to identify meaningful units in the continuous movement of other agents (Newtson, 1973; Zacks & Swallow, 2007) to language acquisition, where learners need to find out where words start and where they end in fluent speech (Aslin, Saffran, & Newport, 1998; Saffran, Aslin, & Newport, 1996; Saffran, Newport, & Aslin, 1996). In the context of language acquisition, this challenge is called the segmentation problem (Aslin et al., 1998; Saffran, Aslin, & Newport, 1996; Saffran, Newport, & Aslin, 1996) and is clearly one of the first challenges infants face, even before they can acquire the meaning of any word.

To extract and understand words from fluent speech, adults can rely on a variety of speech cues, including rhythmic, prosodic and phonotactic cues (e.g., Christophe, Peperkamp, Pallier, Block, & Mehler, 2004; McQueen, 1998; Mehler, Dommergues, Frauenfelder, & Segui, 1981; Norris, McQueen, Cutler, & Butterfield, 1997; Salverda et al., 2007). However, while some of these cues can be perceived at birth and across languages (e.g., Brentari, González, Seidl, & Wilbur, 2011; Christophe, Dupoux, Bertoncini, & Mehler, 1994; Christophe, Mehler, & Sebastian-Galles, 2001; Endress & Hauser, 2010; Pilon, 1981) and are among the most salient cues for word-learning (e.g., Johnson & Jusczyk, 2001; Johnson & Seidl, 2009; Shukla, Nespor, & Mehler, 2007; Shukla, White, & Aslin, 2011), others tend to be language-specific (e.g., Otake, Hatano, Cutler, & Mehler, 1993; Cutler, Mehler, Norris, & Segui, 1986, 1992). As a result, a language-universal mechanism for solving the segmentation problem would be desirable.

A prominent set of potentially language-universal mechanisms that might solve the segmentation problem relies on co-occurrence statistics of various sorts (but see Gervain & Guevara Erra, 2012; Saksida, Langus, & Nespor, 2017). These mechanisms track the predictability of items such as syllables. For example, predicting the next syllable after “the” is much harder than predicting the next syllable after “whis”, because “the” can be followed by any noun while there are few possible continuations after “whis” (e.g., whiskey, whisker, …). More formally, these predictive relationships have been quantified using Transitional Probabilities (TPs), i.e., the conditional probability of a syllable σ_2 following

[☆] This paper is a part of special issue “Special Issue in Honour of Jacques Mehler, Cognition’s founding editor”.

[*] Corresponding author at: Department of Psychology, City, University of London, UK.
E-mail address: ansgar.endress@m4x.org (A.D. Endress).

https://doi.org/10.1016/j.cognition.2021.104621
Received 20 June 2020; Received in revised form 19 December 2020; Accepted 28 January 2021
Available online 17 February 2021
0010-0277/

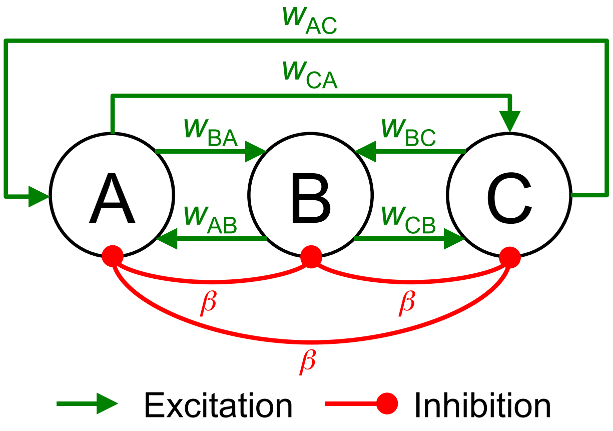

Fig. 1. Schematic representation of the network architecture with three units *A*, *B* and *C* (e.g., representing syllables). All units inhibit each other with a fixed weight of β. They also have tunable excitatory connections. For example, unit *A* sends excitatory input to unit *B* with a weight of w_{BA} and sends excitatory input to unit *C* with a weight of w *CA*. In addition to excitation and inhibition, all units undergo forgetting.

another syllable $\sigma_1, P(\sigma_2|\sigma_1)$.

After the initial discovery that infants and other animals are sensitive to TPs in general (Aslin et al., 1998; Chen & Ten Cate, 2015; Creel, Newport, & Aslin, 2004; Endress, 2010; Endress & Wood, 2011; Fiser & Aslin, 2002a; Hauser, Newport, & Aslin, 2001; Saffran, Aslin, & Newport, 1996; Saffran & Griepentrog, 2001; Saffran, Johnson, Aslin, & Newport, 1999; Saffran, Newport, & Aslin, 1996; Sohail & Johnson, 2016; Toro & Trobalón, 2005; Turk-Browne & Scholl, 2009), further research revealed the astonishing sophistication of these abilities.

For example, adults and infants can track backwards TPs (Endress & Wood, 2011; Pelucchi, Hay, & Saffran, 2009; Perruchet & Desaulty, 2008; Turk-Browne & Scholl, 2009) and discriminate high-TP items from low-TP items when the test-items are played in reverse order with respect to the familiarization (i.e., they readily recognize the item *CBA* after familiarization with *ABC*; Endress & Wood, 2011; Turk-Browne & Scholl, 2009). Learners can also track TPs between non-adjacent items (Endress, 2010; Endress & Wood, 2011; Peña, Bonatti, Nespor, & Mehler, 2002), though in some experiments, additional manipulations were required (Creel et al., 2004; Pacton & Perruchet, 2008). Both abilities are critical for language acquisition, because backwards TPs are in some languages more informative than forward TPs (e.g., Gervain & Guevara Erra, 2012) and because, across languages, non-adjacent dependencies abound (e.g., Newport & Aslin, 2004).

Learners prefer high-TP items to low-TP items even when the items are equated for frequency of occurrence (Aslin et al., 1998), and even when they had heard or seen only the low-TP items but not the high-TP items (Endress & Langus, 2017; Endress & Mehler, 2009; Perruchet & Poulin-Charronnat, 2012).

How can we make sense of these data? While a variety of computational models have been proposed to explain word segmentation (e.g., Batchelder, 2002; Brent & Cartwright, 1996; Christiansen, Allen, & Seidenberg, 1998; Frank, Goldwater, Griffiths, & Tenenbaum, 2010; Orbán, Fiser, Aslin, & Lengyel, 2008; Perruchet & Vinter, 1998), none of the extant models captures the sophistication of statistical learning abilities in their entirety.

For example, network models (such as Simple Recurrent Networks; Elman, 1990) are directional, and thus do not account for backward TPs, while their sensitivity to non-adjacent TPs will likely depend on the network parameters. "Chunking models" that store items in memory (Batchelder, 2002; Perruchet & Vinter, 1998; Thiessen, 2017) and information-theoretic models (or related Bayesian models) that minimize storage space in memory (Brent & Cartwright, 1996; Orbán et al., 2008) will not track (adjacent or non-adjacent) TPs in unattested items, and thus do not account for the entire range of data either.

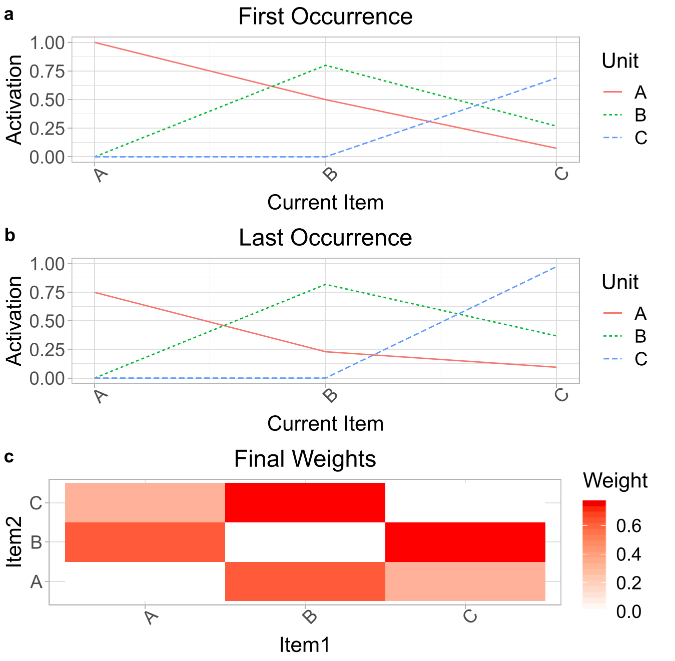

Fig. 2. Illustration of the computational principles of the simulations. We plot the network activation when stimulated by a recurring unit *ABC*. (a) On the first occurrence of the unit, no associations have been formed yet. Hence, when *A* is presented, *A* (but no other items) becomes active, and then decays, though some activation persists even while *C* is presented. Likewise, *B* and *C* become active upon presentation, and then decay. The initial activation is weaker for *B* and *C* than for *A* due to the presence of inhibitory interactions; this is because, for *A*, no other potentially inhibiting representations are active yet, while other activated items (e.g., *A*) have inhibitory input for *B* and *C*. (b) On the last occurrence of a unit, associations between the items have been formed. When the network is externally stimulated with a unit such as *ABC*, the activation of *B* and *C* is greater than that of *A* when the corresponding items are stimulated. This is because *B* and *C* (but not *A*) receive excitatory input from the strongly associated, preceding items. (c) Weights at the end of the familiarization phase. The connection weights between adjacent items are stronger than those between non-adjacent items (i.e., between *A* and *C*).

Here, we suggest that an ability to succeed in the crucial test cases above follows naturally from a correlational learning mechanism such as Hebbian learning. Specifically, we assume that each item (syllable, visual shape, …) is represented by some population of neurons, and that participants are exposed to some sequence *ABCD*…, where each letter stands for an item. If the activation of such a population decays more slowly than the duration of an item, two adjacent items will be active simultaneously, and thus form an association. For example, if the representation of *A* is still active while *B* occurs, these representations will form an association. But if the representation of *A* is still active while *C* occurs, *A* and *C* will form an association as well even though they are not temporarily adjacent (see also Endress, 2010). Importantly, these associations are not directional: just as presenting *A* will activate *B*, presenting *B* will activate *A*.

Here, we provide a computational implementation of this model. The model is a fairly generic network, based on a widely used model of saliency maps in the parietal cortex to which we added a Hebbian learning component (see Fig. 1 for a schematic representation of the network architecture and Supplementary Material A for more details). We use this network architecture as it is fairly generic and widely used, but have no particular claims about attentional involvement in TP computations (but see e.g. Toro, Sinnett, & Soto-Faraco, 2005; Turk-Browne, Jungé, & Scholl, 2005). Further, as this model is rather generic, we do not attempt to fit it to specific experiments. Rather, our critical point is a conceptual one: the sophisticated properties of statistical learning follow naturally from the combination of two simple mechanisms, namely correlational learning and forgetting.

Specifically, the network consists of units that stand for populations of neurons encoding the items (see Fig. 1 for a schematic representation of the network architecture). Excitatory connections between units follow a Hebbian learning rule. To keep the total activation in the network at a reasonable level, we also added mutual interference among the units; the inhibitory interactions do not undergo learning.

Further specifics of the model can be found in Supplementary Material A.

2. Computational principles

We first illustrate the computational principles of the model by running a simulation with a stream consisting of 9 symbols *A*, *B*, …*I* that are arranged into three three-item units *ABC*, *DEF* and *GHI*. Units were concatenated in random order so that each unit occurred 100 times.

Fig. 2 shows the activation in response to the presentation of each item when the unit *ABC* is presented for the first time (a) and for the last time (b) as well as the weights between the underlying items after the last presentation.

Fig. 2a shows that the *A* unit is still active when the *C* item is presented. As a result, we would expect a strong and reciprocal associative link between *A* and *B* and a weaker one between *A* and *C*, which is just what Fig. 2c shows.

Comparing Figs. 2a and b reveals that the activation of *A* is more reduced at its last occurrence. This is due to the inhibitory input from other units: On the first occurrence, no other units are active yet, and activation of *A* can only be reduced through inhibition when other units are active. In contrast, the activations of *B* and *C* do not seem reduced between Figs. 2 a and c. This is because they receive excitatory input from *A* (and *B* in the case of *C*) which compensates the inhibitory input from other units.

While we focus on statistical learning in sequences, there is a considerable literature about statistical learning in simultaneously presented arrays of visual items (see e.g., Fiser & Aslin, 2002b, 2005, among many others). In principle, correlational learning applies to simultaneously presented items as well. For example, consider an experiment where the network is familiarized with two strongly associated pairs of shapes *A*-*B* and *a*-*b* (where each letter stands for a shape). If the network is then tested on the strongly associated pair *A*-*B* and the weakly associated pair *A*-*b*, it will be more familiar with the *A*-*B* pair than with the *A*-*b* pair. In either pair, *A* receives the same amount of inhibition from *B* or *b*. However, only *A* and *B* will have formed an association that provide *A* with excitatory input from *B*, while less excitation is received from *b*.[1]

However, associations among spatially distributed items present other complications, including associations between items and (relative or absolute) spatial positions, Gestalt-based grouping principles and the allocation of spatial attention. As little is known about these factors in statistical learning (but see e.g., Vickery & Jiang, 2009), we thus focus on sequential statistical learning, and use these computational principles to illustrate some of the critical results in the statistical learning literature.

3. Results

3.1. High- vs. low-TP items, tested forwards and backwards

We first explore the discrimination of high vs. low TP items after exposure to a sequence of 4 units of 3 items each (e.g., 4 words of 3 syllables). These units are randomly concatenated into a familiarization stream so that each unit occurs 100 times. We then present the network with test-items (see below) and record the total network activation while each item is presented. We hypothesize that the total activation provides us with a measure of the network's familiarity with the unit.[2]

This cycle of familiarization and test will be repeated 100 times, representing 100 participants.

While keeping the parameters for self-excitation and mutual inhibition constant (α and β in Supplementary Material A), we used forgetting rates (λ_a in Supplementary Material A) between 0 and 1. As forgetting in our model is exponential, a forgetting rate of zero means no forgetting, a forgetting rate of 1 implies the complete disappearance of activation on the next time step (unless a population of neurons receives excitatory input from other populations), and a forgetting rate of 0.5 implies the decay of half of the activation.

Before presenting our results, it is useful to outline possible psychological interpretations of the forgetting parameter. Similar forgetting parameters are widely used in related models (e.g., Bays, Singh-Curry, Gorgoraptis, Driver, & Husain, 2010; Endress & Szabó, 2020; Gottlieb, 2007; Knops, Piazza, Sengupta, Eger, & Melcher, 2014; Roggeman, Fias, & Verguts, 2010), and seem plausible at least at the single neuron level (e.g., Whitmire & Stanley, 2016). Forgetting functions have also been proposed at the macroscopic, cognitive level (e.g., Rubin & Wenzel, 1996; Wixted & Ebbesen, 1991), though the specific forgetting functions are debated. However, the psychological mechanisms underlying "forgetting" have a considerable history of controversy. While forgetting is time-based in our model, many authors argue that, psychologically speaking, there is no forgetting over time unless there are other stimuli that interfere with the memory items (e.g., Baddeley & Scott, 1971; Berman, Jonides, & Lewis, 2009; Nairne, Whiteman, & Kelley, 1999). Here, we do not attempt to decide between these possibilities; in fact, the model equations in Supplementary Material A make it plausible that our interference parameter might well mimic the role of forgetting (see Endress & Szabó, 2020). Our point simply is that the (time-based or interference-based) mechanisms that lead to forgetting are critical for learning to occur.

3.1.1. Adjacent and non-adjacent forward TPs

We first evaluate the network' s sensitivity to forward TPs among adjacent and non-adjacent items. These simulations are inspired by the paradigm by Saffran, Aslin, and Newport (1996) and Saffran, Newport, and Aslin (1996), among many others. After familiarization as described above, the network will be tested on units such as *ABC* and "part-units." Part-units are created either by taking the last two items from one unit and the first item from the next unit (e.g., *BC:D*, where the colon indicates the former unit boundary but is not present in the stimuli) or by taking the last item from one unit and the first two items from the next unit (e.g., *C:DE*). As a result, part-units have occurred during the familiarization sequence but straddled a unit boundary and thus have relatively weak TPs. We thus expect the network to be more familiar with units than with part-units.

The demonstration of a sensitivity to TPs among *non*-adjacent items is inspired by the paradigm by Endress and Bonatti (2007). Specifically, our high non-adjacent TP test-items take their first and the last item from the same unit, but the middle item from a different unit (e.g., *AGC*, where *A* and *C* come from the unit *ABC*, while *G* was the first item of the

[1] To foreshadow the results below, forgetting will also play role in learning. For example, if forgetting is so slow that, say, the representations of *A* and *B* are still active while *a* and *b* are presented, all shapes will form associations, and the network will not preferentially recognize certain pairs.

[2] We also report simulations where we consider only those network activation in the items that are part of the current test-item rather than the global network activation. For example, when a unit *ABC* is presented, we assess the network's familiarity with the items by recording the activation in *A*, *B* and *C* – rather than the activation in *all* items. Intuitively, one would expect the results to be similar, as the active items will mainly be those that have been stimulated. These simulations are reported in Supplementary Material D.

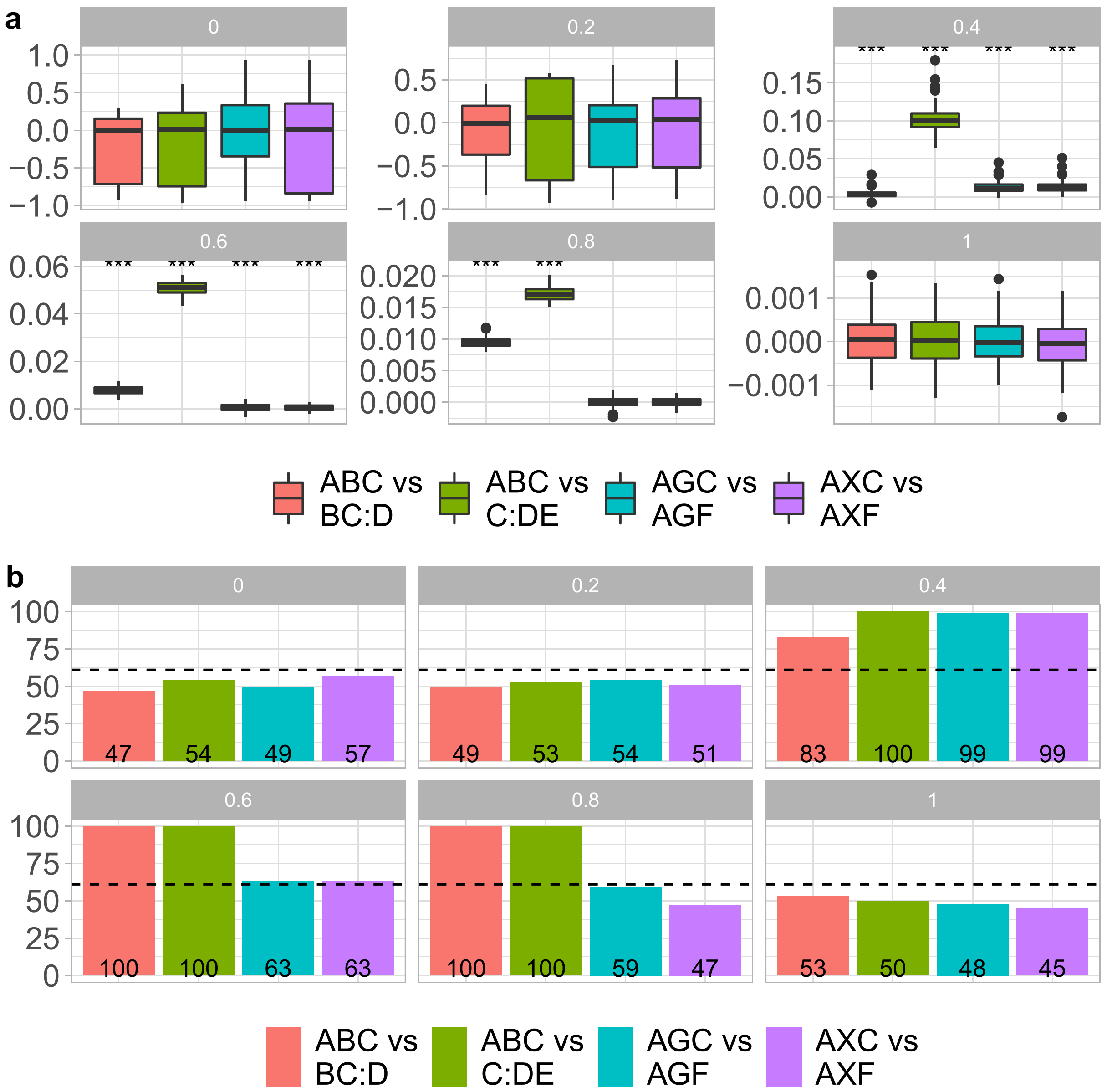

Fig. 3. Results for items presented in forward **order**, different forgetting rates (0, 0.2, 0.4, 0.6, 0.8 and 1), and for the different comparisons (Unit vs. Part-Unit: *ABC* vs. *BC:D* and *ABC* vs. *C:DE*; Rule-Unit vs. Class-Unit: *AGC* vs. *AGF* and *AXC* vs. *AXF*). (a) Difference scores. The 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 the target items. 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.

unit *GHI*). By analogy to Endress and Bonatti (2007), we call these items *rule-units*.

Our low non-adjacent TP test-items take their first and the last items from different units and take the middle item from yet another unit (e.g., *AGF*, where *A* is the first item from *ABC*, *F* is the last item from *DEF*, while *G* was the first item of the unit *GHI*). By analogy to Endress and Bonatti (2007), we call these items *class-units*. The critical difference between the rule-units and the class-units is that the TP between the first and the last item is 1.0 in rule-units and 0 in class-units.

We will also test a second rule-unit vs. class-unit contrast where the middle item is novel and did not appear in the familiarization stream (e.g., *AXC* vs. *AXF*, where *X* has never appeared in the familiarization stream).

For each comparison, we will create normalized difference scores to evaluate the model performance:

$$d = \frac{\text{Item}_1 - \text{Item}_2}{\text{Item}_1 + \text{Item}_2}$$

We then evaluate these difference scores against the chance level of zero using Wilcoxon tests. An alternative evaluation metric is to count the number of simulations (each representing a participant) preferring the target items, 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 Fig. 3a and b. For low forgetting rates (0 and 0.2), the network fails for all comparisons. This is unsurprising as low forgetting rates mean that all items remain active for many time steps, so that the network indiscriminately forms associations among virtually all items, and thus fails to track the statistical structure of the familiarization stream. Likewise, for the maximum forgetting rate, the network fails

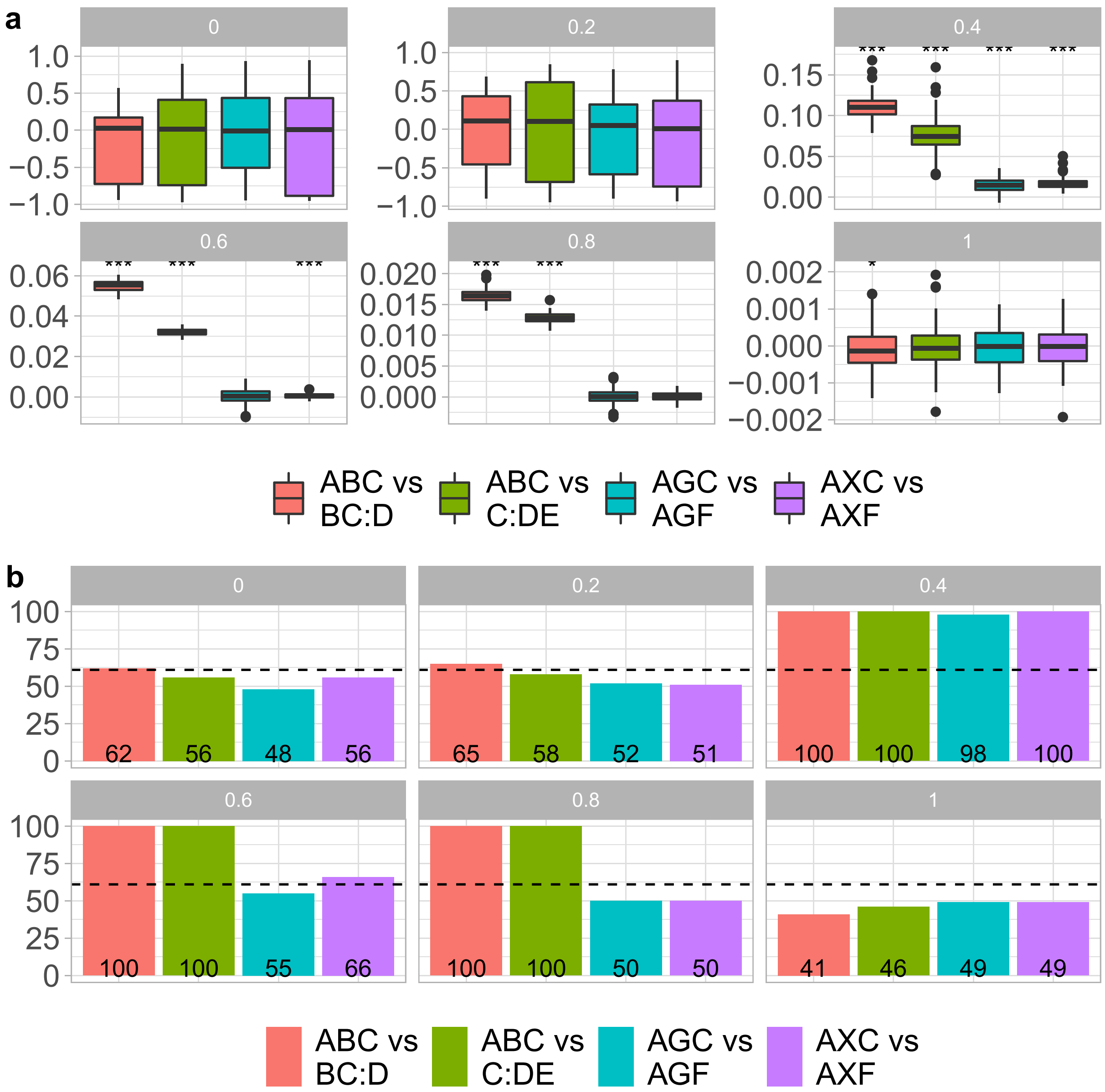

Fig. 4. Results for items presented in **backward order**, different forgetting rates (0, 0.2, 0.4, 0.6, 0.8 and 1), and for the different comparisons (Unit vs. Part-Unit: *ABC* vs. *BC:D* and *ABC* vs. *C:DE*; Rule-Unit vs. Class-Unit: *AGC* vs. *AGF* and *AXC* vs. *AXF*). (a) Difference scores. The 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 the target items. 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.

on all discriminations as well; this is again unsurprising, as no associations can be formed among items if forgetting is so strong that there is no overlap in activation between items.

Critically, for intermediate forgetting rates, the network performed well above chance for all comparisons. It performed somewhat better when contrasting units with *C:DE* part-units than when contrasting units with *BC:D* part-unit, a result that has been observed in human participants with syllables (Saffran, Newport, & Aslin, 1996), tone sequences (Saffran et al., 1999) and visual shapes (Fiser & Aslin, 2002b). Importantly, all difference scores are clearly above chance, and between 83% and 100% of the simulations yielded positive difference scores (though only 63% yielded positive difference scores for forgetting rate 0.6 and non-adjacent TP comparisons). Further, adjacent TPs support higher forgetting rates than non-adjacent TPs, because activations need to last longer for non-adjacent TPs to be formed; while a sensitivity to TPs among adjacent items is maintained for a forgetting rate of 0.8, there is no such sensitivity to non-adjacent TPs.

3.1.2. *Adjacent and non-adjacent backward TPS*

There is considerable evidence that participants are not only sensitive to forward TPs, but also to backward TPs. They track TPs when the only informative TPs are backward rather than forward TPs (Pelucchi et al., 2009; Perruchet & Desaulty, 2008), and discriminate high-TP items from low-TP items when the test-items are played in reverse order (Endress & Wood, 2011; Turk-Browne & Scholl, 2009).

Here, we test the network's ability to track backward TPs by familiarizing the network with the same streams as in the previous section, but playing the test-items in reverse order (e.g., *CBA* instead of *ABC*).

As shown in Fig. 4a and b, the network performance with reversed items essentially mirrors that with forward items, with similar performance for both forward and backward items, with the main difference that the performance asymmetry between *C:DE* and *BC:D* part-units was

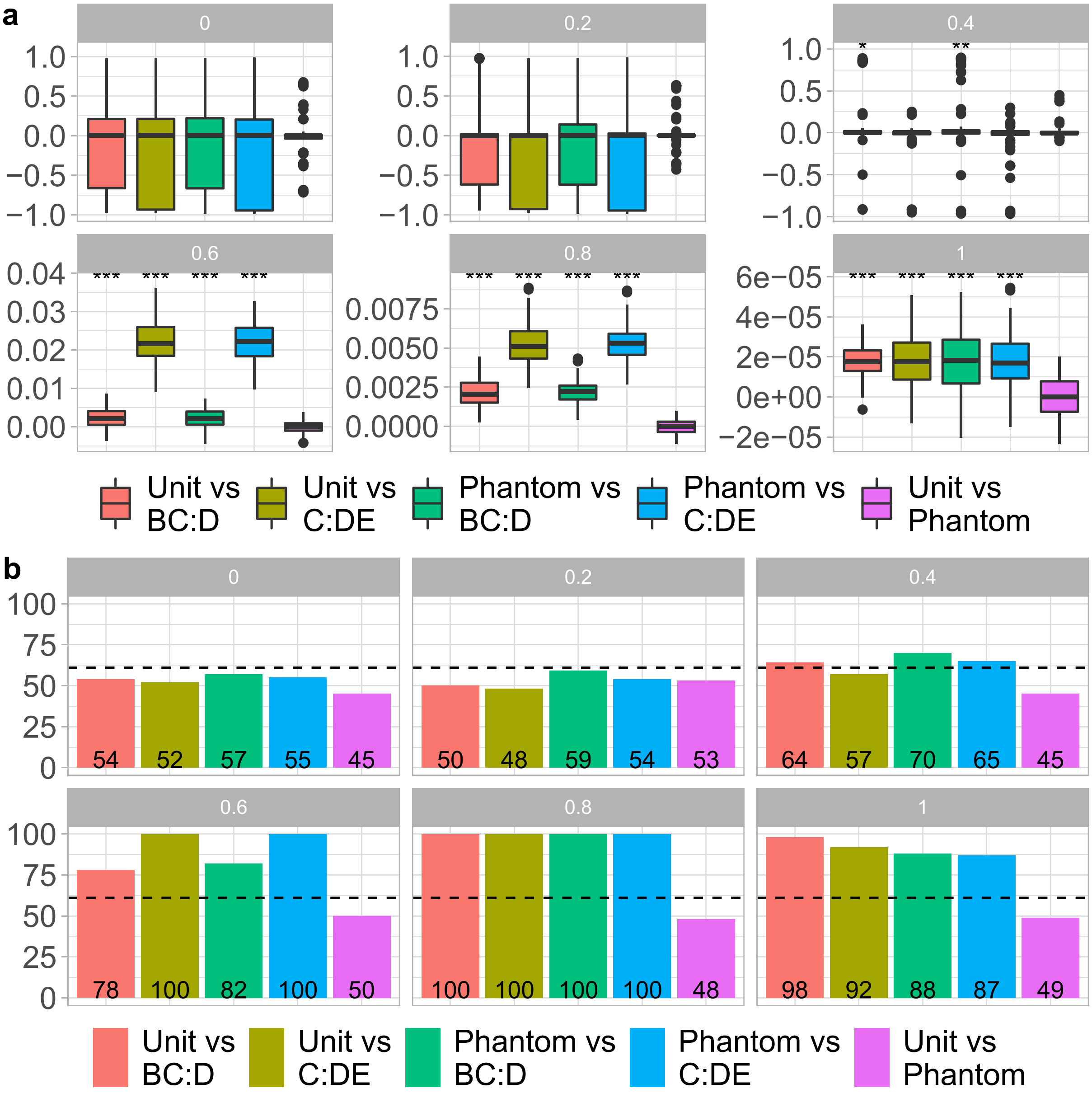

Fig. 5. Results of the simulations comprising phantom-units, for items presented in **forward order**, different forgetting rates (0, 0.2, 0.4, 0.6, 0.8 and 1), and for the different comparisons (Unit vs. Part-Unit: *ABC* vs. *BC:D* and *ABC* vs. *C:DE*; Phantom-Unit vs. Part-Unit: Phantom-Unit vs. *BC:D* and Phantom-Unit vs. *C:DE*; Unit vs. Phantom-Unit). (a) Difference scores. 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 the target items. The simulations are assessed based on the global activation. The dashed line shows the minimum percentage of simulations that is significant based on a binomial test.

reversed.

3.2. The role of frequency of occurrence

The experiments presented so far confound TPs and frequency of occurrence: Units do not only have stronger TPs than part-units, but they also occur more frequently.

This problem was initially noted by Aslin et al. (1998). They addressed it by having infants "choose" between units and part-units that were matched in frequency (see Aslin et al., 1998 for more details on the design).

Endress and Mehler (2009) and Endress and Langus (2017) presented a more "extreme" control experiment. In their experiments, high-TP units were matched in terms of TPs to high-TP *phantom-units* that had the same TPs as units but never occurred in the familiarization stream and thus had a frequency of occurrence of zero (see Supplementary Material B for more details on the design). Participants preferred (i.e., better recognized) high-TP units to low-TP part-units that had occurred in the familiarization stream, they preferred high-TP phantom-units to low-TP part-units despite the difference in frequency of occurrence, and they failed to discriminate between units and phantom-units (but see Perruchet & Poulin-Charronnat, 2012, for evidence that units and phantom-units might sometimes be discriminated).

Here, we expose the network to a six unit stream inspired by Endress and Mehler (2009) and Endress and Langus (2017). Following this, we test the network on units, phantom-units and part-units.

The results are shown in Fig. 5a and b. As in the experiments reported above, the network failed on all comparisons for low forgetting rates as it indiscriminately learned associations among all items.

For medium and, in this experiment, high forgetting rates, the network preferred units and, critically, also phantom-units over part-units roughly to the same extent; we also replicate the somewhat better performance when the part-unit is of *C:DA* type compared to part-units of *BC:D* type. As the participants in Endress and Mehler (2009) and Endress and Langus (2017), the network is thus more sensitive to differences in TPs than to differences in frequency of occurrence, and recognizes TPs even in items it has never encountered before.

In contrast, the network does not seem to discriminate between units and phantom-units, replicating Endress and Mehler's (2009) and Endress and Langus's (2017) results, and suggesting again that the network is more sensitive to TPs than to frequency of occurrence.

4. Discussion

Identifying recurrent units in a continuous signal is an important problem, especially for language acquisition. Observers might potentially solve this problem by tracking co-occurrence statistics among items, assessing the predictiveness of different items. Indeed, humans have sophisticated statistical learning abilities, allowing them to encode and recognize Transitional Probabilities (TPs) irrespective of whether items are played forward or backwards, whether the items are temporarily adjacent or non-adjacent, and whether the units in which the TPs occur are known or entirely novel.

We show that a simple neural network accounts for all of these phenomena based on correlational (i.e., Hebbian) learning. Interestingly, the critical ingredient for successful learning seems to be forgetting: If forgetting is too weak, indiscriminate associations are formed that are, therefore, uninformative; conversely, if forgetting is too strong, no associations are formed.

Given that statistical learning has been observed in human adults and infants as well as non-human animals, it is interesting to ask how forgetting develops over the lifespan. However, as mentioned above, interference can likely mimic the effects of forgetting, and developmental changes in interference might appear as developmental changes in forgetting. For example, younger infants might have more broadly tuned representations (e.g., Pascalis, de Haan, & Nelson, 2002; Vouloumanos, Hauser, Werker, & Martin, 2010) that are more likely to overlap and thus to interfere with each other; in fact, at least in working memory, items that do not have categorical representations are less well retained (e.g., Olsson & Poom, 2005). Conversely, infants might also experience *less* interference if they have fewer representations; after all, if there are fewer representations, there are fewer representations that can interfere with any other representation (see Mani & Plunkett, 2011, for data consistent with this possibility). As a result, we need to know more about the nature of the underlying representations before being able to make specific developmental predictions.

Our results lead to a counterintuitive conclusion about the computational function of statistical learning. While our model presents a rather simple and straightforward mechanistic explanation for our sophisticated statistical learning abilities, these TP-based mechanisms are only partially compatible with the presumed function of statistical learning – namely to store recurrent units in memory. Ultimately, a mechanism that recognizes items played backwards or items it has not encountered at all can hardly be said to maintain faithful memory representations of the relevant items. Conversely, recognizing backwards or unheard items is inconsistent with models that actually store items in memory (Batchelder, 2002; Perruchet & Vinter, 1998; Thiessen, 2017). As a result, it is important to find out what the function of statistical learning is during language acquisition.

Similar dissociations between statistical learning abilities and memory for specific episodes between amnesic and Parkinson's patients have led to the conclusion that humans have a (cortical) declarative memory system that is independent of a (neostriatal) system for forming associations (Knowlton, Mangels, & Squire, 1996; Poldrack et al., 2001). Statistical learning might be used for predictive processing rather than memory per se (Goujon, Didierjean, & Thorpe, 2015; Turk-Browne, Scholl, Johnson, & Chun, 2010), and statistical predictive processing might even *impair* memory encoding (Sherman & Turk-Browne, 2020). Our model is consistent such results: it learns the statistical structure of a sequence (and thus to predict elements in the sequence elements) in the absence of a memory store in which units could be placed.[3]

Together with our model, such results suggest that statistical learning, powerful as it is, might not be sufficient for placing recurring units in memory. After all, we clearly have declarative memories of such items, and know that we know the word *learning* rather than a backwards version such as *gninrael*. As a result, a critical question for future research is to find out how the power of predictive processes such as statistical learning is harnessed to form declarative memories of recurring units in sequences or whether other cues and mechanisms[4] are required.

Acknowledgements

An R implementation of the model is available at http://doi.org/10.25383/city.11359376. This research was support by NIH grant R01-HD073535 to SPJ.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cognition.2021.104621.

References

Aslin, R. N., Saffran, J. R., & Newport, E. L. (1998). Computation of conditional probability statistics by 8-month-old infants. *Psychological Science, 9*, 321–324.

Baddeley, A. D., & Scott, D. (1971). Short term forgetting in absence of proactive interference. *The Quarterly Journal of Experimental Psychology, 23*, 275–283.

Batchelder, E. O. (2002). Bootstrapping the lexicon: A computational model of infant speech segmentation. *Cognition, 83*(2), 167–206.

Bays, P. M., Singh-Curry, V., Gorgoraptis, N., Driver, J., & Husain, M. (2010). Integration of goal- and stimulus-related visual signals revealed by damage to human parietal cortex. *Journal of Neuroscience, 30*, 5968–5978. https://doi.org/10.1523/JNEUROSCI.0997-10.2010.

Berman, M. G., Jonides, J., & Lewis, R. L. (2009). In search of decay in verbal short-term memory. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 35*(2), 317–333. https://doi.org/10.1037/a0014873.

Brent, M., & Cartwright, T. (1996). Distributional regularity and phonotactic constraints are useful for segmentation. *Cognition, 61*(1–2), 93–125.

Brentari, D., González, C., Seidl, A., & Wilbur, R. (2011). Sensitivity to visual prosodic cues in signers and nonsigners. *Language and Speech, 54*(1), 49–72.

Chen, J., & Ten Cate, C. (2015). Zebra finches can use positional and transitional cues to distinguish vocal element strings. *Behavioural Processes, 117*, 29–34. https://doi.org/10.1016/j.beproc.2014.09.004.

Christiansen, M. H., Allen, J., & Seidenberg, M. S. (1998). Learning to segment speech using multiple cues: A connectionist model. *Language & Cognitive Processes, 13*(2–3), 221–268.

Christophe, A., Dupoux, E., Bertoncini, J., & Mehler, J. (1994). Do infants perceive word boundaries? An empirical study of the bootstrapping of lexical acquisition. *Journal of the Acoustical Society of America, 95*(3), 1570–1580.

Christophe, A., Mehler, J., & Sebastian-Galles, N. (2001). Perception of prosodic boundary correlates by newborn infants. *Infancy, 2*(3), 385–394.

[3] While Parkinson's patients were initially thought to be impaired in associative learning in general (Knowlton et al., 1996), further research revealed that, for many tasks, such patients have intact associative learning abilities, and that their impairment might depend on the need to integrate probabilistic feedback across learning episodes (Smith & McDowall, 2006). Be that as it might, statistical learning does not seem to lead to declarative knowledge of specific events even in studies that link it to the Medial Temporal Lobe (Turk-Browne et al., 2010).

[4] As mentioned above, such cues and mechanisms might include rhythmic, prosodic and phonotactic cues (e.g., Christophe, Peperkamp, Pallier, Block, & Mehler, 2004; Johnson & Jusczyk, 2001; Johnson & Seidl, 2009; McQueen, 1998; Mehler et al., 1981; Norris et al., 1997; Salverda et al., 2007), especially those that available across languages (e.g., Brentari et al., 2011; Christophe et al., 1994, 2001; Endress & Hauser, 2010; Pilon, 1981; Shukla et al., 2007, 2011).

Christophe, A., Peperkamp, S., Pallier, C., Block, E., & Mehler, J. (2004). Phonological phrase boundaries constrain lexical access I. Adult data. *Journal of Memory and Language, 51*(4), 523–547.

Creel, S. C., Newport, E. L., & Aslin, R. N. (2004). Distant melodies: Statistical learning of nonadjacent dependencies in tone sequences. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 30*(5), 1119–1130. https://doi.org/10.1037/0278-7393.30.5.1119.

Cutler, A., Mehler, J., Norris, D., & Segui, J. (1986). The syllable's differing role in the segmentation of french and english. *Journal of Memory and Language, 25*(4), 385–400.

Cutler, A., Mehler, J., Norris, D., & Segui, J. (1992). The monolingual nature of speech segmentation by bilinguals. *Cognitive Psychology, 24*(3), 381–410.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science, 14*(2), 179–211.

Endress, A. D. (2010). Learning melodies from non-adjacent tones. *Acta Psychologica, 135* (2), 182–190.

Endress, A. D., & Bonatti, L. L. (2007). Rapid learning of syllable classes from a perceptually continuous speech stream. *Cognition, 105*(2), 247–299.

Endress, A. D., & Hauser, M. D. (2010). Word segmentation with universal prosodic cues. *Cognitive Psychology, 61*(2), 177–199.

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

Endress, A. D., & Mehler, J. (2009). 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.

Endress, A. D., & Szabó, S. (2020). Sequential presentation protects memory from catastrophic interference. *Cognitive Science, 44*(5). https://doi.org/10.1111/cogs.12828.

Endress, A. D., & Wood, J. N. (2011). From movements to actions: Two mechanisms for learning action sequences. *Cognitive Psychology, 63*(3), 141–171.

Fiser, J., & Aslin, R. N. (2002a). Statistical learning of higher-order temporal structure from visual shape sequences. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 28*(3), 458–467.

Fiser, J., & Aslin, R. N. (2002b). Statistical learning of new visual feature combinations by infants. *Proceedings of the National Academy of Sciences of the United States of America, 99*(24), 15822–15826. https://doi.org/10.1073/pnas.232472899.

Fiser, J., & Aslin, R. N. (2005). Encoding multielement scenes: Statistical learning of visual feature hierarchies. *Journal of Experimental Psychology. General, 134*(4), 521–537. https://doi.org/10.1037/0096-3445.134.4.521.

Frank, M. C., Goldwater, S., Griffiths, T. L., & Tenenbaum, J. B. (2010). Modeling human performance in statistical word segmentation. *Cognition, 117*(2), 107–125. https://doi.org/10.1016/j.cognition.2010.07.005.

Gervain, J., & Guevara Erra, R. (2012). The statistical signature of morphosyntax: A study of Hungarian and Italian infant-directed speech. *Cognition, 125*(2), 263–287. https://doi.org/10.1016/j.cognition.2012.06.010.

Gottlieb, J. (2007). From thought to action: The parietal cortex as a bridge between perception, action, and cognition. *Neuron, 53*, 9–16.

Goujon, A., Didierjean, A., & Thorpe, S. (2015). Investigating implicit statistical learning mechanisms through contextual cueing. *Trends in Cognitive Sciences, 19*, 524–533. https://doi.org/10.1016/j.tics.2015.07.009.

Hauser, M. D., Newport, E. L., & Aslin, R. N. (2001). Segmentation of the speech stream in a non-human primate: Statistical learning in cotton-top tamarins. *Cognition, 78*(3), B53–B64.

Johnson, E. K., & Jusczyk, P. W. (2001). Word segmentation by 8-month-olds: When speech cues count more than statistics. *Journal of Memory and Language, 44*(4), 548–567.

Johnson, E. K., & Seidl, A. H. (2009). At 11 months, prosody still outranks statistics. *Developmental Science, 12*(1), 131–141. https://doi.org/10.1111/j.1467-7687.2008.00740.x.

Knops, A., Piazza, M., Sengupta, R., Eger, E., & Melcher, D. (2014). A shared, flexible neural map architecture reflects capacity limits in both visual short-term memory and enumeration. *Journal of Neuroscience, 34*(30), 9857–9866. https://doi.org/10.1523/JNEUROSCI.2758-13.2014.

Knowlton, B. J., Mangels, J. A., & Squire, L. R. (1996). A neostriatal habit learning system in humans. *Science, 273*, 1399–1402.

Mani, N., & Plunkett, K. (2011). Phonological priming and cohort effects in toddlers. *Cognition, 121*, 196–206. https://doi.org/10.1016/j.cognition.2011.06.013.

McQueen, J. M. (1998). Segmentation of continuous speech using phonotactics. *Journal of Memory and Language, 39*(1), 21–46.

Mehler, J., Dommergues, J., Frauenfelder, U., & Segui, J. (1981). The syllable's role in speech segmentation. *Journal of Verbal Learning and Verbal Behavior, 20*(3), 298–305.

Nairne, J. S., Whiteman, H. L., & Kelley, M. R. (1999). Short-term forgetting of order under conditions of reduced interference. *The Quarterly Journal of Experimental Psychology, 52*, 241–251.

Newport, E. L., & Aslin, R. N. (2004). Learning at a distance I. Statistical learning of non-adjacent dependencies. *Cognitive Psychology, 48*(2), 127–162.

Newtson, D. (1973). Attribution and the unit of perception of ongoing behavior. *Journal of Personality and Social Psychology, 28*(1), 28–38.

Norris, D., McQueen, J. M., Cutler, A., & Butterfield, S. (1997). The possible-word constraint in the segmentation of continuous speech. *Cognitive Psychology, 34*(3), 191–243. https://doi.org/10.1006/cogp.1997.0671.

Olsson, H., & Poom, L. (2005). Visual memory needs categories. *Proceedings of the National Academy of Sciences of the United States of America, 102*(24), 8776–8780. https://doi.org/10.1073/pnas.0500810102.

Orbán, G., Fiser, J., Aslin, R. N., & Lengyel, M. (2008). Bayesian learning of visual chunks by human observers. *Proceedings of the National Academy of Sciences of the United States of America, 105*(7), 2745–2750. https://doi.org/10.1073/pnas.0708424105.

Otake, T., Hatano, G., Cutler, A., & Mehler, J. (1993). Mora or syllable? Speech segmentation in Japanese. *Journal of Memory and Language, 32*(2), 258–278.

Pacton, S., & Perruchet, P. (2008). An attention-based associative account of adjacent and nonadjacent dependency learning. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 34*(1), 80–96. https://doi.org/10.1037/0278-7393.34.1.80.

Pascalis, O., de Haan, M., & Nelson, C. A. (2002). Is face processing species-specific during the first year of life? *Science, 296*(5571), 1321–1323. https://doi.org/10.1126/science.1070223.

Pelucchi, B., Hay, J. F., & Saffran, J. R. (2009). Learning in reverse: Eight-month-old infants track backward transitional probabilities. *Cognition, 113*(2), 244–247. https://doi.org/10.1016/j.cognition.2009.07.011.

Peña, M., Bonatti, L. L., Nespor, M., & Mehler, J. (2002). Signal-driven computations in speech processing. *Science, 298*(5593), 604–607. https://doi.org/10.1126/science.1072901.

Perruchet, P., & Desaulty, S. (2008). A role for backward transitional probabilities in word segmentation? *Memory and Cognition, 36*(7), 1299–1305. https://doi.org/10.3758/MC.36.7.1299.

Perruchet, P., & Poulin-Charronnat, B. (2012). Beyond transitional probability computations: Extracting word-like units when only statistical information is available. *Journal of Memory and Language, 66*(4), 807–818. https://doi.org/10.1016/j.jml.2012.02.010.

Perruchet, P., & Vinter, A. (1998). PARSER: A model for word segmentation. *Journal of Memory and Language, 39*, 246–263.

Pilon, R. (1981). Segmentation of speech in a foreign language. *Journal of Psycholinguistic Research, 10*(2), 113–122.

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

Roggeman, C., Fias, W., & Verguts, T. (2010). Salience maps in parietal cortex: Imaging and computational modeling. *NeuroImage, 52*, 1005–1014. https://doi.org/10.1016/j.neuroimage.2010.01.060.

Rubin, D. C., & Wenzel, A. E. (1996). One hundred years of forgetting: A quantitative description of retention. *Psychological Review, 103*(4), 734–760.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science, 274*(5294), 1926–1928.

Saffran, J. R., & Griepentrog, G. J. (2001). Absolute pitch in infant auditory learning: Evidence for developmental reorganization. *Developmental Psychology, 37*(1), 74–85.

Saffran, J. R., Johnson, E., Aslin, R. N., & Newport, E. L. (1999). Statistical learning of tone sequences by human infants and adults. *Cognition, 70*(1), 27–52.

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

Saksida, A., Langus, A., & Nespor, M. (2017). Co-occurrence statistics as a language dependent cue for speech segmentation. *Developmental Science, 20*(3). https://doi.org/10.1111/desc.12390.

Salverda, A. P., Dahan, D., Tanenhaus, M. K., Crosswhite, K., Masharov, M., & McDonough, J. (2007). Effects of prosodically modulated sub-phonetic variation on lexical competition. *Cognition, 105*(2), 466–476. https://doi.org/10.1016/j.cognition.2006.10.008.

Sherman, B. E., & Turk-Browne, N. B. (2020). 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. https://doi.org/10.1073/pnas.2013291117.

Shukla, M., Nespor, M., & Mehler, J. (2007). An interaction between prosody and statistics in the segmentation of fluent speech. *Cognitive Psychology, 54*(1), 1–32. https://doi.org/10.1016/j.cogpsych.2006.04.002.

Shukla, M., White, K. S., & Aslin, R. N. (2011). Prosody guides the rapid mapping of auditory word forms onto visual objects in 6-mo-old infants. *Proceedings of the National Academy of Sciences of the United States of America, 108*(15), 6038–6043. https://doi.org/10.1073/pnas.1017617108.

Smith, J. G., & McDowall, J. (2006). When artificial grammar acquisition in parkinson's disease is impaired: The case of learning via trial-by-trial feedback. *Brain Research, 1067*, 216–228. https://doi.org/10.1016/j.brainres.2005.10.025.

Sohail, J., & Johnson, E. K. (2016). How transitional probabilities and the edge effect contribute to listeners' phonological bootstrapping success. *Language Learning and Development*, 1–11. https://doi.org/10.1080/15475441.2015.1073153.

Thiessen, E. D. (2017). What's statistical about learning? Insights from modelling statistical learning as a set of memory processes. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, 372*. https://doi.org/10.1098/rstb.2016.0056.

Toro, J. M., Sinnett, S., & Soto-Faraco, S. (2005). Speech segmentation by statistical learning depends on attention. *Cognition, 97*(2), B25–B34. https://doi.org/10.1016/j.cognition.2005.01.006.

Toro, J. M., & Trobalón, J. B. (2005). Statistical computations over a speech stream in a rodent. *Perception & Psychophysics, 67*(5), 867–875.

Turk-Browne, N. B., Jungé, J., & Scholl, B. J. (2005). The automaticity of visual statistical learning. *Journal of Experimental Psychology. General, 134*(4), 552–564. https://doi.org/10.1037/0096-3445.134.4.552.

Turk-Browne, N. B., & Scholl, B. J. (2009). Flexible visual statistical learning: Transfer across space and time. *Journal of Experimental Psychology. Human Perception and Performance, 35*(1), 195–202.

Turk-Browne, N. B., Scholl, B. J., Johnson, M. K., & Chun, M. M. (2010). Implicit perceptual anticipation triggered by statistical learning. *Journal of Neuroscience, 30*, 11177–11187. https://doi.org/10.1523/JNEUROSCI.0858-10.2010.

Vickery, T. J., & Jiang, Y. V. (2009). Associative grouping: Perceptual grouping of shapes by association. *Attention, Perception, & Psychophysics, 71*(4), 896–909. https://doi.org/10.3758/APP.71.4.896.

Vouloumanos, A., Hauser, M. D., Werker, J. F., & Martin, A. (2010). The tuning of human neonates' preference for speech. *Child Development, 81*(2), 517–527. https://doi.org/10.1111/j.1467-8624.2009.01412.x.

Whitmire, C. J., & Stanley, G. B. (2016). Rapid sensory adaptation redux: A circuit perspective. *Neuron, 92*, 298–315. https://doi.org/10.1016/j.neuron.2016.09.046.

Wixted, J. T., & Ebbesen, E. B. (1991). On the form of forgetting. *Psychological Science, 2*(6), 409–415.

Zacks, J. M., & Swallow, K. M. (2007). Event segmentation. *Current Directions in Psychological Science, 16*(2), 80–84.