Cognition 230 (2023) 105176

Contents lists available at ScienceDirect

Cognition

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

Discussion

No need to forget, just keep the balance: Hebbian neural networks for statistical learning

Ángel Eugenio Tovar [a,*], Gert Westermann [b]

[a] Facultad de Psicología, Universidad Nacional Autónoma de México, Av. Universidad 3004, 04510 Coyoacán, Mexico
[b] Department of Psychology, Lancaster University, Lancaster LA1 4YF, United Kingdom

ARTICLE INFO

Keywords
Statistical learning
Hebbian learning
Artificial neural networks
Language processing
Computational modeling

ABSTRACT

Language processing in humans has long been proposed to rely on sophisticated learning abilities including statistical learning. Endress and Johnson (E&J, 2021) recently presented a neural network model for statistical learning based on Hebbian learning principles. This model accounts for word segmentation tasks, one primary paradigm in statistical learning. In this discussion paper we review this model and compare it with the Hebbian model previously presented by Tovar and Westermann (T&W, 2017a; 2017b; 2018) that has accounted for serial reaction time tasks, cross-situational learning, and categorization paradigms, all relevant in the study of statistical learning. We discuss the similarities and differences between both models, and their key findings. From our analysis, we question the concept of "forgetting" in the model of E&J and their suggestion of considering forgetting as the critical ingredient for successful statistical learning. We instead suggest that a set of simple but well-balanced mechanisms including spreading activation, activation persistence, and synaptic weight decay, all based on biologically grounded principles, allow modeling statistical learning in Hebbian neural networks, as demonstrated in the T&W model which successfully covers learning of nonadjacent dependencies and accounts for differences between typical and atypical populations, both aspects that have not been fully demonstrated in the E&J model. We outline the main computational and theoretical differences between the E&J and T&W approaches, present new simulation results, and discuss implications for the development of a computational cognitive theory of statistical learning.

1. Introduction

Human language processing has long been proposed to rely on sophisticated learning abilities such as statistical learning. In a recent paper, Endress and Johnson (2021) (E&J) presented a neural network model for statistical learning mainly focusing on modeling transitional probabilities during word segmentation tasks (Saffran, Aslin, & Newport, 1996), which is one main experimental paradigm in this field. Here we comment on the E&J article and argue that in previous studies Tovar and Westermann (T&W; Tovar & Westermann, 2017a, 2017b; Tovar, Westermann, & Torres, 2018) have accounted for statistical learning phenomena with neural network models that operate in an analogous way to the E&J model. Particularly, T&W have simulated serial reaction time tasks (Tovar et al., 2018), another main paradigm of statistical learning (Hunt & Aslin, 2001; Nissen & Bullemer, 1987); cross-situational learning (Tovar & Westermann, 2017a); and learning of equivalence classes (Tovar & Westermann, 2017b), a paradigm which has long had a great impact in the study of language and symbolic behavior in the tradition of behavior analysis (Sidman, 1994). Through the modeling of these tasks, it is becoming clear that all of them present different sides of the same coin of statistical learning accounted for with the same learning principles. While both the E&J and T&W models support correlational (Hebbian) learning as a main force underlying statistical learning, below we identify key differences between both computational implementations that are crucial in the conceptualization of learning and forgetting in the models. Both models also differ in their recourse to biologically informed learning mechanisms, and in their ability to cover different experimental paradigms and to account for atypical populations. In the following, we compare both models and summarize their main findings and the key differences between them. In doing so, we aim to provide a framework for future computational developments in the field of statistical learning and to stress the importance of Hebbian learning as a domain general mechanism across different aspects of language development.

* Corresponding author at: Office 218, Facultad de Psicología, UNAM, Av. Universidad 3004, Mexico City, CP 04510, Mexico.
E-mail address: aetovar@unam.mx (Á.E. Tovar).

https://doi.org/10.1016/j.cognition.2022.105176
Received 18 November 2021; Received in revised form 15 April 2022; Accepted 16 May 2022
Available online 25 November 2022
0010-0277/© 2022 Elsevier B.V. All rights reserved.

2. The E&J model

During word segmentation tasks, the main focus of E&J, a participant is presented with a stream of novel sounds that represent fluent speech from which discrete words can be extracted. The statistical learning account of word segmentation (Saffran et al., 1996) proposes that sensitivity to transitional probabilities between syllables in the continuous stream of sound is key to recognizing units (words), because within a word transitional probabilities between syllables are high but between words they are low.

E&J suggested that an ability to succeed in this kind of statistical learning "follows naturally from a correlational learning mechanism such as Hebbian learning" (p. 2). They then provided a computational implementation of a generic neural network model that uses Hebbian learning to account for statistical learning.

E&J argued that no previous computational models captured "the sophistication of statistical learning abilities in their entirety". They cited a group of models that include connectionist, chunking, and information-theoretic models (Batchelder, 2002; Brent & Cartwright, 1996; Christiansen, Allen, & Seidenberg, 1998; Elman, 1990; Frank, Goldwater, Griffiths, & Tenenbaum, 2010; Orbán, Fiser, Aslin, & Lengyel, 2008; Perruchet & Vinter, 1998; Thiessen, 2017). Their main critique of these models was their inability to extract transitional probabilities in statistical learning tasks. Nonetheless, E&J did not explain the criteria for including models in their review. Remarkably, while E&J accounted for transitional probabilities with associative Hebbian learning, they did not review previous work using Hebbian algorithms that capture correlational or statistical learning (Ganis & Schendan, 1992; McMurray, Horst, & Samuelson, 2012). Critically, E&J overlooked two of our papers (Tovar et al., 2018; Tovar & Westermann, 2017b) in which we have presented Hebbian algorithms in artificial neural networks that have covered similar simulations to those described in their paper.

3. Analysis of T&W and E&J models

3.1. Similarities

Both the T&W and E&J models are implemented as neural networks composed of single layers of fully connected units (examples in Figs. 1 and 2). Representation in both models is symbolic: each unit (or artificial neuron) represents one and only one item from the environment (e.g., syllables, words, visual objects). Thus, the presentation of the syllable "*ba*" followed by the syllable "*by*" activates Neuron A (representing *ba*), followed by activation of Neuron B (representing *by*). This sequential presentation produces coactivation of units AB for a moment, which triggers Hebbian learning (Hebb, 1949): *Neurons that fire together wire together*. Frequent *baby* presentations strengthen the connection weight between A and B (W_{AB}), capturing a high transitional probability in *baby* and providing a simple neural model of statistical regularities.

Strong W_{AB} allows spreading activation from either A to B or from B to A in both models because both conceptualize connection weights as symmetrical, accounting for empirical evidence that participants respond with high accuracy to backwards BA dependencies after being exposed to AB training (Pelucchi, Hay, & Saffran, 2009; Sidman et al., 1982).

E&J analyzed the performance of their networks mainly from global unit activation patterns, whereas T&W mainly focused on connection weights. Since weights determine global activation, both models are also comparable in terms of how they were evaluated.

3.2. Differences

3.2.1. Hebbian learning

Hebb (1949) postulated one of the most influential neural theories of learning and memory; he hypothesized that synaptic efficacy increases from connected neurons firing together. His elegant and simple description has been formalized in numerous algorithms (Gerstner & Kistler, 2002). A multiplication of neuron activation states captures their firing correlations and drives changes in their connection weights W (Eqs. 1 and 2). A learning rate (β) and other multiplicative parameters control the amount of changes of W at each time step.

Changes in W

$$\Delta W_{AB} = \beta\ (act\_A{*}act\_B) \tag{1}$$

Updating of W

$$W_{AB(t+1)} = W_{AB(t)} + \Delta W_{AB} \tag{2}$$

One problem with the above equations is that there are no limits for weight increases; coactivation of AB leads to strengthening W_{AB}, which in turn propagates more activation between A and B in an endless loop, and this is not biologically realistic. Additionally, Hebb's original ideas did not consider activation-dependent decays in Ws. Current models of Hebbian learning have modified the original rule in order to control activation and weight overgrowth, and it is at this point that the T&W and E&J implementations differ.

The Hebbian algorithm in T&W implements both growth and decay in Ws as a function of the coactivation of units. This idea was motivated by neurophysiological long-term potentiation (LTP) and long-term depression (LTD) of synaptic efficacy in brain networks (Bliss, Collingridge, & Morris, 2007; Malenka & Bear, 2004). LTP and LTD are mathematically captured as a continuum of weight changes: if neuronal coactivation is strong and surpasses a threshold LTP occurs, but when neuronal activity falls below the threshold, LTD takes place (Bear, 1995; Bienenstock, Cooper, & Munro, 1982).

To capture this LTP/LTD continuum T&W included a threshold parameter (θ) that switches between positive and negative (LTP/LTD) weight changes depending on the amount of neural coactivation. Importantly, increasing the threshold θ allowed T&W to simulate and predict patterns of statistical learning in Down syndrome (Tovar et al., 2018; Tovar & Westermann, 2017a) which has been associated with atypical synaptic plasticity with increased LTD and limited LTP (Andrade-Talavera, Benito, Casañas, Rodríguez-Moreno, & Montesinos, 2015; Rueda, Flórez, & Martínez-Cué, 2012; Scott-McKean & Costa, 2011). Nonetheless, it should not be assumed that more LTP means more learning and more LTD means less learning or forgetting; instead, a balance between LTP and LTD is necessary to provide neural networks with the computational flexibility required for learning (Kemp & Manahan-Vaughan, 2007; Pinar et al., 2017).

The T&W algorithm includes a second parameter, *lambda* (λ), for limiting change to weights that are already strong, a mechanism inspired by metaplasticity in biological networks (Abraham, 2008). λ depends on the difference between the current W_{AB} and the current AB coactivation, tuning the amount and direction of weight changes, and it ensures that W_{AB} tracks the statistical co-occurrence of A and B. With θ and λ, frequent exposure to AB items leads to strengthen W_{AB}, while sparse exposure of AB keeps W_{AB} low.

In the T&W (2018) model, θ and λ are included through the following equations:

Computing λ in T&W

$$\text{if } (act\_A{*}act\_B) \rangle \theta,\ \lambda = (act\_A{*}act\_B) - W_{AB}$$

$$\text{else } \lambda = -W_{ij} \tag{3}$$

Hebbian learning in T&W

$$\Delta W_{AB} = \beta\ \lambda\ (act\_A{*}act\_B) \tag{4}$$

On the other hand, the Hebbian algorithm in the E&J model is the generic one, as depicted in Eqs. 1 and 2, without modulation of overgrowing connections. E&J did include a weight decay term, but it was set to zero during all simulations, which means that no decay of weights

was modeled. It is not clear therefore whether and how their model prevents excessive growth of connections.

The E&J network also includes inhibitory connections which control the overall activation and consequently modulate learning in the neural network. All units in the E&J model are fully connected with both excitatory and inhibitory connections. Excitatory connections undergo Hebbian learning, whereas inhibitory connections do not; instead, they all keep an arbitrary fixed weight that is determined by the modeler. E&J justified inhibitory connections "to keep the total activation in the network at a reasonable level" (p. 3). Nonetheless, they did not provide further details on the rationale of this implementation.

3.2.2. Spreading activation and recall

Activation spreads through the network via the weighted connections between units. Both networks compute the activation of each unit after adding up activations from external stimuli and lateral activation propagated from other units.

In the T&W model, spreading activation is conceptualized as a key component to recall important associations, akin to pattern completion (O'Reilly & Rudy, 2000). For example, after exposure to both AB and BC item-pairs, strong W_{AB} and W_{BC} weights result in spreading activation between the three-neuron assembly anytime the network is presented with either of these items (A, B or C). Through this mechanism learning of nonadjacent transitional probabilities (e.g., AC after AB and BC were presented) is mechanistically explained; in the above example, exposure to BC items also produces some level of activation in neuron A due to spreading activation from B to A, and this in turn produces coactivation of A and C, triggering Hebbian learning in W_{AC}. In the T&W model, incoming spreading activation is a weighted sum of the inputs coming from the already active units in the network:

Incoming spreading activation in T&W

$$\text{if net\_input} > x,\ \text{in\_act\_A} = \frac{1}{1 + exp^{-net\_input}}$$

$$\text{else in\_act\_A} = 0 \quad (5)$$

The equation uses a threshold x because otherwise negative and zero values are also transformed into positive incoming activation. The value of x can vary, restricting spreading activation in the network as x increases.

The E&J network also includes spreading activation. It is computed using the net input value as well, but for their simulations it was limited by an excitation coefficient parameter that multiplied the value of the net input times 0.7. Spreading activation in the E&J model exerts similar effects as in the T&W model, however E&J did not stress its role on statistical learning. Instead, they focused on forgetting, which they argued arises from activation decay.

3.2.3. Activation decay and forgetting

Forgetting for E&J is decisive in their conceptualization of statistical learning. E&J argued that, critically, "the sophisticated properties of statistical learning follow naturally from the combination of two simple mechanisms, namely correlational learning and forgetting" (p. 3).

Forgetting in the E&J model is implemented through unit activation decay. Let's say that item A activates neuron A at time step 1. Then, at time step 2, when item B activates neuron B, some degree of A activation must remain in the network to allow coactivation of AB and Hebbian learning of W_{AB}. For how long should A remain active? E&J ran parametric tests to determine the value of the exponential forgetting parameter: if forgetting was at its highest level (1 in the range 0–1), A was only active during one time step. If forgetting was zero, A remained active for the rest of the simulation. E&J found that intermediate forgetting values accounted better for learning transitional probabilities of both adjacent (AB) and nonadjacent (A_C) dependencies.

The T&W (2018) model also implements activation decay in simulations of sequence learning, but it controls neural activation decay in a three-step dynamic: full activation at step 1; 90% activation at step 2; and 0 activation from step 3 on. These decays apply only for the external source of activation.

While activation decay is used in both models, only E&J conceptualized it as forgetting and as "the critical ingredient for successful learning" (p. 7). We disagree with E&J's view of activation decay as forgetting, and argue that this may be a misleading interpretation. Activation values from external stimuli in both artificial and biological networks are non-persistent but are constantly updated in response to changes in the environment (Huber & O'Reilly, 2003). Rather than forgetting, we suggest that activation decay simply models the remaining neural activation from immediately past events, which is of course key to trigger Hebbian learning between sequentially perceived items. Nonetheless, we argue that there is more besides activation decay to provide Hebbian networks with efficient power to extract statistical information from the environment; notably, spreading activation and weight decay must be both highlighted as additional key components in the algorithm. In the next sections we provide evidence for this argument while we provide a chronological overview of the main tasks and simulation results reported by T&W and E&J. We also present new simulation results with the T&W model that challenge E&J's notion of activation decay as the key ingredient for Hebbian statistical learning.

4. What the models account for

4.1. Categorization in equivalence classes (T&W)

Tovar and Westermann (2017b) studied symbolic categorization through the simulation of equivalence class formation, a traditional categorization paradigm in behavior analysis (Sidman, 1994). One main finding in this field is that humans can derive relations between non directly trained items. This is usually done with conditional discrimination training, where participants are reinforced after selecting the correct comparison stimulus (e.g., B1, when B2, B3, and Bn are also present) in the presence of the related sample (e.g., A1). For example, after being trained on A1B1 and B1C1 conditional discriminations (e.g., A1 being a picture of a dog, B1 the written word "dog" and C1 the written word "chien" which means dog in French) participants go through a test phase of symmetry relations B1A1 and C1B1 (e.g., B1A1 requires selecting the picture of a dog [A1] from a pool of pictures when the word "dog" [B1] is presented as the sample stimulus), and transitive A1C1 relations (here, linking the dog picture with the word "chien" after seeing the picture paired with the word "dog", and the word "dog" paired with "chien"), which notably non-human animals fail to acquire. Tovar and Westermann (2017b) accounted for the trained (AB, BC), and derived symmetry (BA, CB) and transitive (AC) relations based on direct exposure to item-pairs and spreading activation in their network.

Notably, associative strengths between trained, symmetry and transitive relations in Tovar and Westermann (2017b) are equivalent to transitional probabilities between adjacent, backwards and nonadjacent dependencies in Endress and Johnson (2021), respectively, as we will review later. T&W showed that associative strength between derived relations is an inverse function of adjacency between items, simulating data from human participants (Spencer & Chase, 1996).

Tovar and Westermann (2017b) also simulated performance from populations with intellectual disabilities by unbalancing LTP/LTD in the weight adjustments of their model, and showed that associative learning of transitive (i.e., nonadjacent) items was comparably more impaired than learning of adjacent items, a learning pattern in line with empirical results (Devany, Hayes, & Nelson, 1986).

4.2. Cross-situational learning of word-object mappings (T&W)

Tovar and Westermann (2017a) also evaluated their version of the Hebbian algorithm during acquisition of word-object mappings in

simulations of cross-situational learning. For these tasks a word label is presented concurrently with different visual objects, including the correct visual match and other competitors. The visual competitors vary across training trials. Correct mappings between word labels and visual objects emerge from sensitivity to the auditory and visual statistical regularities across trials (Smith & Yu, 2008). T&W simulated this task in a neural network architecture originally proposed by McMurray et al. (2012) which was modified to include the T&W Hebbian algorithm. After training, the network was tested, using spreading activation, for lexical production (through correct activations of the auditory units when the corresponding visual items were presented) and lexical comprehension (through correct activations of visual units when the corresponding auditory items were presented). T&W analyzed whether it was possible to account for the empirical evidence of relatively preserved comprehension with impaired production in the lexical development of Down syndrome (Næss, Lyster, Hulme, & Melby-Lervåg, 2011), and showed that unbalancing LTP/LTD in the algorithm explained this atypical pattern of lexical development.

4.3. Serial reaction time (T&W)

In a subsequent paper, Tovar, Westermann and Torres (2018) simulated a serial reaction time task (SRT; Fig. 1a). In this study their model successfully predicted the performance of typical children and children with Down syndrome. The model and the participants were exposed to two streams of visual stimuli, one stream consisting of the fixed sequence A1-A2 followed by a Target object (AT trials), and the second stream consisting of the fixed sequence B1-B2-B3 followed by the Target (BT trials). Participants were instructed to respond as fast as possible to the target. In SRTs, sequence learning is confirmed if response times get faster for trials presenting fixed sequences compared with trials where A and B items are randomly ordered in the sequence (e.g., random: B2TA1B3A2B1T). The model predicted a critical difference between typical and atypical populations that was confirmed in a complementary empirical study with children: the typical group learned under a variety of familiarization schedules, while the with Down syndrome group only learned when the familiarization schedule did not interleave AT and BT trials; otherwise, their performance was impaired and described as a case of learning interference arising from two competing predictors (A and B) of the target (T). This study showed that, due to the reported LTP/LTD imbalance, competing predictors produce atypically stronger competition and disrupt learning in the model of Down syndrome.

4.4. Word segmentation (E&J)

E&J simulated a generic word segmentation task to later analyze transitional probabilities between items in a variety of conditions (i.e., forward, backward, and nonadjacent transitional probabilities). In E&J (2021) a total of 9 items (A, B, C… I) were presented to the model. Triplets were always presented sequentially together (e.g., A-B-C; G-H-I) which led to a strengthening of the connections between, for example, A and B, and B and C. This part of the process accounted for increases in transitional probabilities of the adjacent dependencies AB and BC, and thus the segmentation of words (triplets such as ABC) from the continuous stream of items. From this simulation E&J also accounted for learning of backward transitional probabilities (e.g., CBA after familiarization with ABC) comparable to empirical reports of sensitivity to these dependencies (Pelucchi et al., 2009). Learning of backward dependencies in E&J's model (2021) is analogous to the learning of symmetry relations reported in Tovar and Westermann (2017b).

4.5. Nonadjacent transitional probabilities (E&J)

E&J also reported learning of transitional probabilities between nonadjacent dependencies (e.g., between A and C, using the test-item AXC, after familiarization with ABC). These transitional probabilities were extracted during familiarization by the remaining activation of A (due to prior presentation of sequence AB) when sequence BC was presented in the next step. As detailed before, this happens when the network operates with intermediate forgetting. Equally to the description of transitive relations in Tovar and Westermann (2017b), E&J reported that final connection weights between adjacent dependencies (e.g., AB) were stronger than weights between nonadjacent dependencies (e.g., AC).

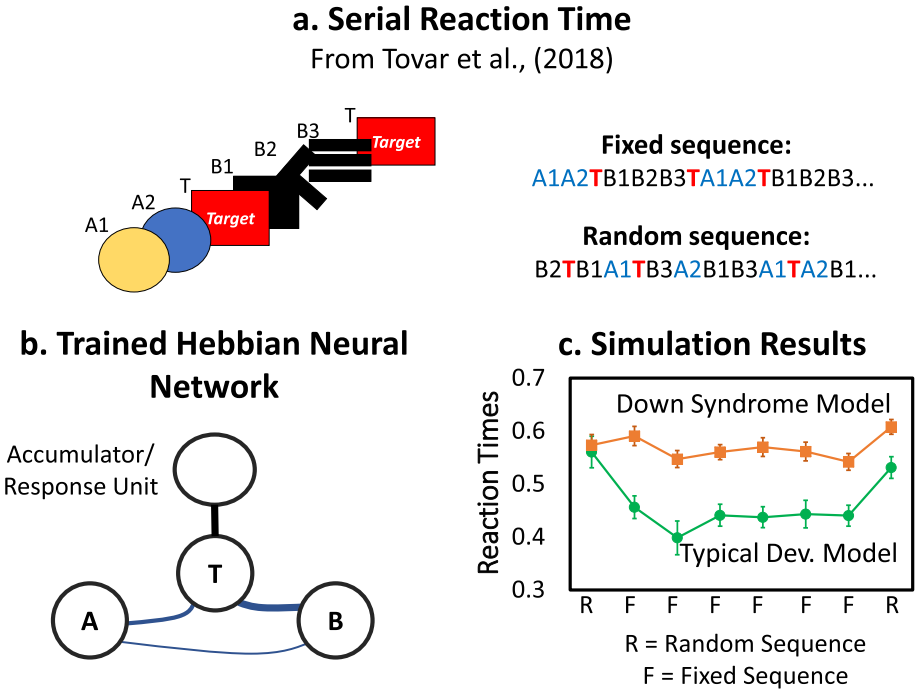

Fig. 1. (a) A schematic representation of the serial reaction time task used in Tovar et al. (2018). (b) Architecture of the Hebbian neural network used in this study. (c) Mean reaction times after 5 runs of the TD model (green circles) and the model of Down syndrome (orange squares) as reported in the original study. The U shape depicted by the TD model reveals learning of the fixed sequences, while the plane curve of the DS model reveals no learning of the fixed sequences. 1b and 1c are adapted from the original publication in Cognition, Vol 171, Tovar, Westermann & Torres, From altered synaptic plasticity to atypical learning: A computational model of Down syndrome. Copyright (2018), with permission from Elsevier. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

5. Revisiting word segmentation and nonadjacent transitional probabilities in the T&W Model

For this discussion paper we ran two additional simulations with the T&W model (a Matlab implementation of this model is available in https://osf.io/6jzg3/). Below we first describe the simulation of a word segmentation task that has not been covered in the T&W model before. We aimed to demonstrate the ability of the T&W model to account for empirical results with this classic paradigm that was the focus of Endress and Johnson (2021). We then discuss how the learning of nonadjacent dependencies reported in both models has not addressed an empirically relevant task that we cover here for the first time with the T&W model and argue on a weakness of E&J's approach to study this task.

5.1. Word segmentation in T&W

We simulated the word segmentation task used in the classic study of Saffran et al. (1996). The authors presented 8-month-old infants with a continuous stream of auditory syllables comprising a total of four 3-syllable nonsense words (Fig. 2a): *tupiro, golabu, bidaku, padoti,* repeated in random order (e.g., *bidakupadotigolatubidaku*…). After familiarization, infants were presented with 4 test items; two words: *tupiro* and *golabu*, and two nonwords: *dapiku* and *tilado*. Nonwords had the same syllables used in familiarization but not in the order they appeared as words. During the novelty preference tests, infants showed significant discrimination between words and nonwords, revealing sensitivity to the statistical structure of the continuous stream of syllables.

We simulated this task in a T&W neural network composed of 12 units fully connected. Each unit stood for one of the 12 syllables that composed the four 3-syllable words. Syllables were presented sequentially to the network, each syllable at each time step, with no pauses or markers between words. Familiarization consisted of continuous presentation of 180 words in random order, as in the original study. After familiarization, we analyzed connection weights between units that composed the test words: *tupiro* and *golabu*, and the nonwords: *dapiku* and *tilado*. The model was run 24 times in the same task, and it successfully simulated the main empirical result, as captured in stronger connection weights between syllables in test words ($M = 0.64$, $SD = 0.04$) compared with nonwords ($M = 0.33$, $SD = 0.04$, Fig. 2c). The weights in the model also revealed learning between nonadjacent word syllables (M = 0.52, SD = 0.04) and this was weaker than learning between adjacent word syllables (M = 0.70, SD 0.04, Fig. 2c). This pattern of results shows that the T&W model succeeds at simulating this classic word segmentation task and shows comparable results to those reported in the simulations of Endress and Johnson (2021).

5.2. Learning of nonadjacent dependencies in T&W

Both models have already reported learning of nonadjacent dependencies (Endress & Johnson, 2021; Tovar & Westermann, 2017b) when these are related through a highly frequent intermediate item (e.g., intermediate B during ABC). Nonetheless, several empirical studies have focused on the learning of nonadjacent elements when the intermediate element is unrelated during familiarization (e.g., AXC; Gómez, 2002; Newport & Aslin, 2004; Wang, Zevin, & Mintz, 2019; Wilson et al., 2020). Interestingly, many of these studies have shown negative or mixed results, revealing difficulties of participants to learn the distant dependencies. For example, Newport and colleagues described strong selectivity in the ability to learn nonadjacent regularities depending on the stimulus materials; human adults were poor learners of nonadjacent syllables but they did learn nonadjacent dependencies between consonants, and between vowels (Newport & Aslin, 2004). This pattern was not consistent across species: tamarin monkeys were good learners of nonadjacent dependencies between syllables, and between vowels, but not between consonants (Newport et al., 2004). A study by Gómez (2002) using word-like units in an artificial language, showed that learnability of nonadjacent words benefits when the adjacent dependencies become less relevant and more unpredictable.

For this simulation we modeled the influential study of Gómez (2002). In Experiment 1 she used three-item auditory word sequences (e.g., AXD: *pel-wadim-rud*). Strings from Language 1 (L1) took the form AXD, BXE, and CXF, and those in Language 2 (L2) were AXE, BXF, and CXD. Both languages had the same adjacent dependencies and were only distinguishable by their nonadjacent dependencies. In each language the first word predicted the final word (e.g., in L1 *pel* predicts *rud*). The A, B, C, D, E, and F elements were constant, while variability was manipulated by drawing X from a pool of either 2, 6, 12 or 24 words, depending on the experimental condition (Fig. 3a). During familiarization, participants were exposed to one of the 2 languages. Short pauses (250 ms) were used between words, and longer pauses (750 ms) between three-word sequences to make words and sequences distinguishable from each other. The three-word sequences were randomly ordered and the number of X elements varied between conditions (In Condition 1: 2, e.g., AXD trials were *pel-wadim-rud,* and *pel-kicey-rud*; Condition 2: 6; Condition 3: 12; and Condition 4: 24). The total number of familiarization trials in all conditions was 432 irrespective of variability in the X elements.

During tests, all participants were presented with strings used in Condition 1 (with only 2 X elements) of both L1 and L2 and were asked to discriminate whether the test sequence was legal or not according to the language they had previously heard. In this way tests evaluated discrimination of the two languages based only on their nonadjacent regularities. Participants' performance improved from Condition 1 to 4, as a function of increasing the number of intervening X elements during familiarization, and this suggested that higher variability of the adjacent (X) elements facilitates learnability of the nonadjacent (e.g., A_D) dependencies.

We simulated this experiment with the T&W model using 30 units fully connected. Each unit represented one of the 30 words required for the A, B, C, D, E, F and 24 X elements. Three-word sequences from L1 were presented to the model by activating each unit at each time step and with one-time-step pauses between sentences to allow distinguishing between words and sentences as in the empirical study. We modeled the four experimental conditions through 432 familiarization trials as in the original study.

To evaluate the model's performance, we obtained the mean weights across the three-word sequences of legal (L1) and illegal (L2) tests in each condition. These values are shown in Fig. 3b, and they capture the same tendency reported by Gómez (2002), because increasing the number of X elements resulted in higher discriminability between legal and illegal sequences in the model. We quantified discriminability through delta values as the difference of mean weights for legal sequences minus mean weights for illegal word sequences within each experimental condition. Higher values indicate better discriminability of legal from illegal sequences. Delta values increased through conditions as: Condition 1 $M = 0.015$, $SD = 0.029$, Condition 2 $M = 0.105$, $SD = 0.038$, Condition 3 $M = 0.132$, $SD = 0.029$, and Condition 4 $M = 0.144$, $SD = 0.024$.

We additionally analyzed the weights for each type of word-pairs: adjacent elements (e.g., *pel-wadim*), legal nonadjacent elements (e.g., *pel-rud*), and illegal nonadjacent elements (e.g., *pel-jic*); these values show that while the learning of adjacent elements becomes less relevant, the legal nonadjacent elements are strengthened, and this provides the key to discriminate between legal and illegal sequences (Fig. 3b).

To model the study of Gómez (2002) we used a low LTD/LTP threshold (θ) value. In previous studies (Tovar et al., 2018; Tovar & Westermann, 2017a,b), T&W used θ values between 0.65 and 0.7 to model typical performance, in this simulation we set θ to 0.6 because using higher values resulted in no learning and no differences between the four experimental conditions (neurophysiological theories and evidence support that θ is not fixed but it moves from the previous and incoming stimulation regularities; Bear, 1995; Bienenstock et al., 1982).

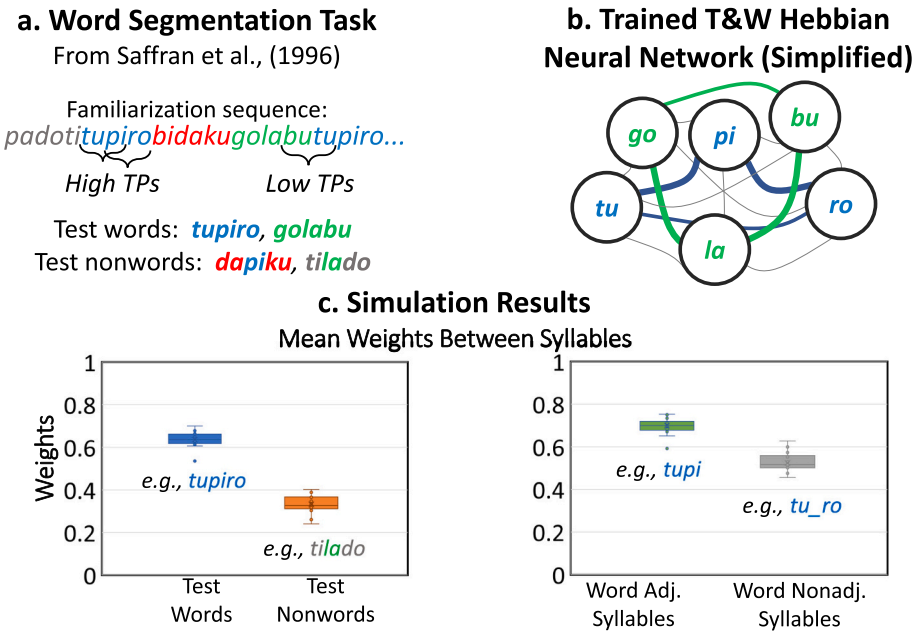

Fig. 2. (a) A schematic representation of the word segmentation task used in Saffran et al. (1996). (b) A (simplified) T&W Hebbian neural network trained with this task, thicker connections between processing units represent stronger syllable associations from higher transitional probabilities. Only some neurons are shown. (c left) Boxplots of the weight values between syllables in test words and nonwords. (c right) Boxplots of the weight values computed across all adjacent and nonadjacent word syllables.

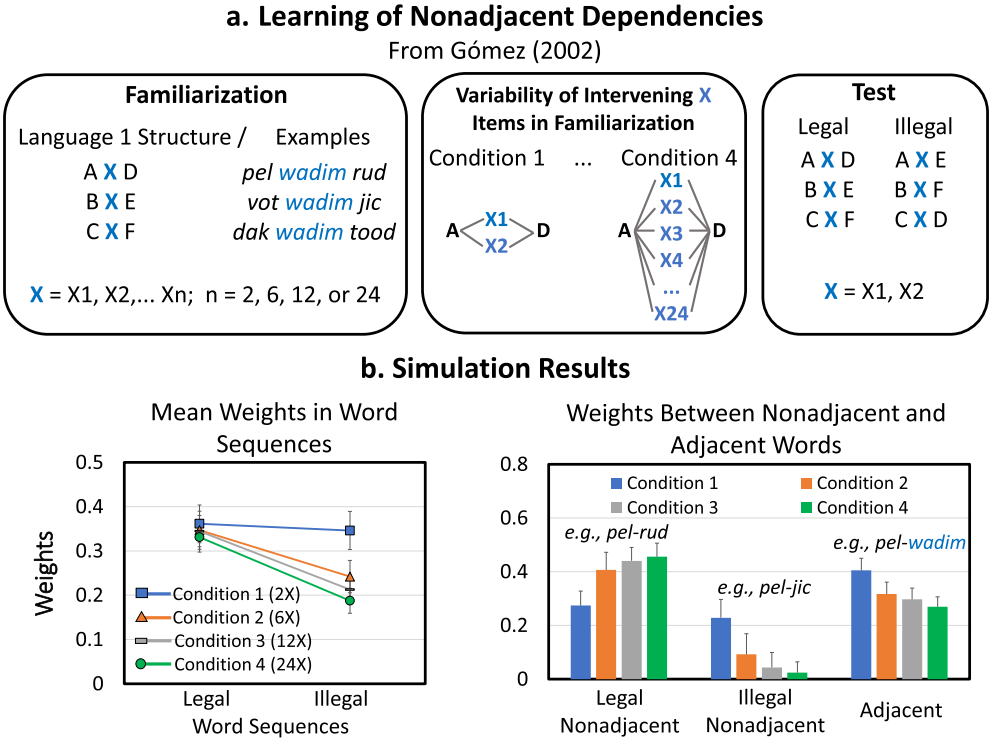

Fig. 3. (a) A schematic representation of the grammar task used in (Gómez, 2002). (b left) Mean connection weights for legal and illegal test sequences across the four experimental conditions simulated in the T&W model. (b right) Mean connection weights for nonadjacent words of legal and illegal sequences, and for adjacent words across the four experimental conditions simulated in the T&W model. Error bars show the standard deviations across 100 runs of the model.

This is interesting because in a network with spreading activation, lowering θ limits weight decay and favors learning from spreading activation, and consequently of nonadjacent regularities, as T&W have reported before (2017b). However, the present simulation shows that this processing restriction (θ) interacts with the external regularities so that only in determined scenarios the learning of nonadjacent dependencies gets favored. This result in turn captures the selectivities in the learning of nonadjacent words in artificial grammars empirically reported by Gómez (2002).

On the other hand, the approach of E&J has stressed the role of activation decay (i.e., forgetting) to account for learning of nonadjacent dependencies. We hypothesized that under E&J's assumptions, adjusting the persistence of neuronal activation through the decay parameter would result in either always learning the nonadjacent regularities and legal sequences (i.e., because A remains active by the time D is presented) or always failing in its learning (i.e., because A and D are not coactive at any time), irrespective to experimental conditions, which differs from the empirical evidence. We tested these hypotheses through

additional simulations of the study by Gómez (2002) using a Hebbian neural network implementing the equations presented by Endress and Johnson (2021). We evaluated the network performance with three forgetting values in the decay parameter (0, 0.5, and 1; corresponding to no forgetting, intermediate, and complete forgetting, respectively). We focused on simulations of the experimental conditions 1 and 4 only, because the main differences in learning the word sequences were reported between these conditions (Gómez, 2022).

We ran two sets of simulations.[1] For the first one we set excitation and inhibition coefficients to 0, to suppress in this way spreading activation in the model. This forced the weight updates in the network to result only from the interaction of correlation learning and activation decay (forgetting). For the second set, we set the excitation and inhibition coefficients to 0.7 and 0.4, respectively, which are the typical values used by Endress and Johnson (2021).

We evaluated discriminability (i.e., delta values) between legal and illegal sequences. Table 1 summarizes the results of these simulations, and it shows that maintaining any neural activation values in the network, through setting forgetting to 0 or 0.5, resulted in high discriminability of legal from illegal sequences in both conditions 1 and 4, but removing all activation values after each word presentation, through setting the forgetting parameter to 1, resulted in zero discriminability between legal and illegal sequences in both conditions 1 and 4. Critically, in the absence of spreading activation (set 1), this model fails to capture differences between conditions 1 and 4. Nonetheless, differences between conditions 1 and 4 did appear when spreading activation operated in the network (set 2) along with zero and intermediate forgetting values.

These simulations show that forgetting (implemented through activation decay/persistence) cannot be considered the only key ingredient besides correlation learning to account for complex performance in statistical learning tasks. From the present results it becomes clear that the role of additional components, such as spreading activation and weight decay, should also be regarded as critical to extend the scope of Hebbian algorithms in statistical learning.

6. Discussion

The models by T&W and E&J have demonstrated Hebbian learning as a promising mechanism for statistical learning. The parameter variations in these Hebbian networks are useful to test interactive effects of cognitive processes related to memory, retention, processing speed, and learning disabilities; all of these being important to explain individual and group differences in statistical learning (Arciuli, 2017).

Associative (Hebbian or correlation) learning has long been suggested to be a key component of statistical learning (Conway, 2020; Frost, Armstrong, Siegelman, & Christiansen, 2015; Pacton & Perruchet, 2008; Perruchet & Vinter, 1998). Remarkably, while it has been difficult to extend verbal theories (and other models) of correlation learning to account for complex phenomena such as the learning of nonadjacent and backwards dependencies, the T&W and E&J models provide formal means to cover these data just by slight adaptations of the general correlation learning theory. Notably, however, the present analysis shows that the conceptualization of these computational adaptations seems to be in opposing directions.

From our analysis, there are two key differences between the T&W and E&J models. One concerns associative weight changes: in T&W associative weight changes are either positive (LTP) or negative (LTD), they depend on detected and recalled items' co-occurrences, and mechanisms underlying weight changes are biologically motivated. In contrast, in the E&J model positive weight changes seem to be controlled in part by fixed inhibition, which we argue is a less biologically plausible implementation, and the E&J model does not include negative associative changes. Negative weight changes in the T&W model are a fundamental component for statistical learning because they emerge when environmental regularities become less relevant than they have been before. It is through the balance of strengthening and weakening (LTP/LTD) of weights that the neural network creates an efficient statistical model of the external regularities. Additionally, the effectiveness of spreading activation depends on a suitable LTP/LTD balance. Disturbance of this balance in the T&W model predicted atypical statistical learning, and this prediction was confirmed in Down syndrome (Tovar et al., 2018), a population in which the LTP/LTD imbalance exists (Rueda et al., 2012). The implementation of positive and negative weight changes in the T&W Hebbian algorithm provides a direct link between altered neural processing and atypical statistical learning.

The second difference between the models concerns activation decay: while both networks include activation decay, only E&J have conceptualized it as forgetting and the key ingredient for statistical learning. We argue that the concept of forgetting is misleading because activation decay merely represents typical accommodation of the network to the changing environment. Consequently, there is no need to postulate a high-level interpretation of this mechanism such as forgetting. In line with previous models (Huber & O'Reilly, 2003), we suggest the persistence of the neural response or the remaining neural activation as preferable interpretations of the processes captured by activation decay.

Finally, within the Hebbian theory there are relevant mechanisms that provide neural networks with efficient power to extract statistical regularities, and these mechanisms cannot be circumscribed to activation decay. In line with the original Hebbian theory, efficient statistical learning across domains (e.g., categorization, cross-situational learning, word segmentation) merely requires the concurrent or sequential neural activations through perceived items. The reviewed models have highlighted at least two mechanisms that explain how adjacent and distant items can produce concurrent neural activations: spreading activation and activation persistence (through activation decay parameters). We highlight the balance between spreading activation, activation decay, and synaptic weight decay, because too much or not enough of them disrupt statistical learning in Hebbian networks (Tovar et al., 2018; Tovar & Westermann, 2017a; Tovar & Westermann, 2017b). All of these components must exist and interact within a suitable balance to allow the networks to become efficient statistical learning models.

7. Future directions

Future computational research should test the scope and limitations of Hebbian learning. This point is important considering theories that suggest the need for multiple cognitive processes to account for complex statistical learning (Arciuli, 2017; Batterink, Reber, Neville, & Paller, 2015; Thiessen & Erickson, 2013). Two recent papers have stressed the importance of such perspectives; one comparative study evaluating humans and other primates that suggests that the human-only abilities during complex statistical learning may rely on verbal recoding strategies besides associative learning (Rey, Minier, Malassis, Bogaerts, & Fagot, 2019); and a paper by Conway (2020) that puts forward a neurocognitive theory of statistical learning. This theory postulates two sets of learning principles: the first is data-driven and automatic, based on general cortical plasticity (where Hebbian learning is well placed), and the second set is goal-directed and attention-based, composed by a modulatory executive system that allows learning and generalization of complex global patterns, including cross-modal dependencies. Notably,

[1] These simulations were run with 100 networks and 432 trials for each condition. Remaining neural activations were removed between trials (sequences) to make sequences distinguishable from each other as in the original study of Gómez (2002). Since there was no weight decay implemented in these simulations, connection weights increased without upper limits, therefore we normalized connection weights in the range [01] after finishing simulations of each condition to compute means and standard deviations in a comparable range across conditions.

Table 1
Discriminability of legal and illegal sequences in simulations of experimental conditions 1 and 4 from the study of Gómez (2002).

Neural Network	Parameters		Delta Values		
	Forgetting	Spreading Activation (SA)	Experimental Condition 1 **Mean** (S.D.) $n = 100$	Experimental Condition 4 **Mean** (S.D.) n = 100	Differences Between Conditions 1 and 4 *t*-test *p values*
Network with E&J's equations (Set 1)	0	without SA	**0.325** (0.003)	**0.325** (0.003)	0.454
	0.5		**0.304** (0.010)	**0.304** (0.009)	0.775
	1		**−0.001** (0.054)	**0.001** (0.043)	0.750
Network with E&J's equations (Set 2)	0	with SA	**0.023** (0.001)	**0.325** (0.003)	< **0.0001** *
	0.5		**0.197** (0.012)	**0.234** (0.030)	< **0.0001** *
	1		**0.004** (0.048)	**−0.003** (0.045)	0.247
T&W Model**			**0.015** (0.029)	**0.144** (0.024)	< **0.0001** *

Note: Single asterisks * indicate significant differences in delta values between conditions 1 and 4, these simulations capture the empirical results from Gómez (2002).
**Data from the last row is taken from the simulations run with the T&W model described in the main text.

Conway (2020) has suggested that learning of nonadjacent dependencies requires the second system. However, the T&W Hebbian network reviewed here has shown that this learning can be accounted for by principles from the first system only.

Both the T&W and E&J neural networks show that learning of challenging statistical regularities, such as backwards and nonadjacent dependencies, emerges in networks with balanced learning parameters, but in line with the hypothesis of verbal mediation (Rey et al., 2019) and the second learning system proposed by Conway (2020), empirical evidence has shown that acquiring backwards and nonadjacent regularities appears as a more complex process that does not occur as systematically as the learning of forward instances (Chartier & Rey, 2020; Wilson et al., 2020). Testing the reviewed models in more challenging learning scenarios is necessary to identify whether and what additional computations are required, besides associative Hebbian learning, for a comprehensive neurocomputational theory of statistical learning.

Acknowledgements

This study was supported by Consejo Nacional de Ciencia y Tecnología (CONACYT) [CB 285152], and by the Economic and Social Research Council (ESRC) International Centre for Language and Communicative Development (LuCiD) [ES/S007113/1 and ES/L008955/1].

References

Abraham, W. C. (2008). Metaplasticity: Tuning synapses and networks for plasticity. *Nature Reviews Neuroscience, 9*(5), 387. https://doi.org/10.1038/nrn2356
Andrade-Talavera, Y., Benito, I., Casañas, J. J., Rodríguez-Moreno, A., & Montesinos, M. L. (2015). Rapamycin restores BDNF-LTP and the persistence of long-term memory in a model of Down's syndrome. *Neurobiology of Disease, 82*, 516–525. https://doi.org/10.1016/j.nbd.2015.09.005
Arciuli, J. (2017). The multi-component nature of statistical learning. *Philosophical Transactions of the Royal Society, B: Biological Sciences, 372*(1711), 20160058. https://doi.org/10.1098/rstb.2016.0058
Batchelder, E. O. (2002). Bootstrapping the lexicon: A computational model of infant speech segmentation. *Cognition, 83*(2), 167–206. https://doi.org/10.1016/S0010-0277(02)00002-1
Batterink, L. J., Reber, P. J., Neville, H. J., & Paller, K. A. (2015). Implicit and explicit contributions to statistical learning. *Journal of Memory and Language, 83*, 62–78. https://doi.org/10.1016/j.jml.2015.04.004
Bear, M. F. (1995). Mechanism for a sliding synaptic modification threshold. *Neuron, 15*(1), 1–4.
Bienenstock, E. L., Cooper, L. N., & Munro, P. W. (1982). Theory for the development of neuron selectivity: Orientation specificity and binocular interaction in visual cortex. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience, 2*(1), 32–48.
Bliss, T. V. P., Collingridge, G. L., & Morris, R. G. M. (2007). Synaptic plasticity in the hippocampus. In P. Andersen, R. G. M. Morris, D. G. Amaral, T. V. P. Bliss, & J. O'Keefe (Eds.), *The Hippocampus book* (pp. 343–474). Oxford University Press.
Brent, M. R., & Cartwright, T. A. (1996). Distributional regularity and phonotactic constraints are useful for segmentation. *Cognition, 61*(1), 93–125. https://doi.org/10.1016/S0010-0277(96)00719-6
Chartier, T. F., & Rey, A. (2020). Is symmetry inference an essential component of language? *Learning & Behavior, 48*(3), 279–280. https://doi.org/10.3758/s13420-019-00405-5
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. https://doi.org/10.1080/016909698386528
Conway, C. M. (2020). How does the brain learn environmental structure? Ten core principles for understanding the neurocognitive mechanisms of statistical learning. *Neuroscience & Biobehavioral Reviews, 112*, 279–299. https://doi.org/10.1016/j.neubiorev.2020.01.032
Devany, J. M., Hayes, S. C., & Nelson, R. O. (1986). Equivalence class formation in language-able and language-disabled children. *Journal of the Experimental Analysis of Behavior, 46*(3), 243–257. https://doi.org/10.1901/jeab.1986.46-243
Elman, J. L. (1990). Finding structure in time. *Cognitive Science, 14*(2), 179–211.
Endress, A. D., & Johnson, S. P. (2021). When forgetting fosters learning: A neural network model for statistical learning. *Cognition, 104621*. https://doi.org/10.1016/j.cognition.2021.104621
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
Frost, R., Armstrong, B. C., Siegelman, N., & Christiansen, M. H. (2015). Domain generality versus modality specificity: The paradox of statistical learning. *Trends in Cognitive Sciences, 19*(3), 117–125. https://doi.org/10.1016/j.tics.2014.12.010
Ganis, G., & Schendan, H. (1992). Hebbian learning of artificial grammars. *Proceedings of the Cognitive Science Society, 14*, 838–843.
Gerstner, W., & Kistler, W. M. (2002). Mathematical formulations of Hebbian learning. *Biological Cybernetics, 87*(5), 404–415. https://doi.org/10.1007/s00422-002-0353-y
Gómez, R. L. (2002). Variability and detection of invariant structure. *Psychological Science, 13*(5), 431–436. https://doi.org/10.1111/1467-9280.00476
Hebb, D. O. (1949). *The organization of behavior; a neuropsychological theory*. Wiley.
Huber, D. E., & O'Reilly, R. C. (2003). Persistence and accommodation in short-term priming and other perceptual paradigms: Temporal segregation through synaptic depression. *Cognitive Science, 27*(3), 403–430. https://doi.org/10.1207/s15516709cog2703_4
Hunt, R. H., & Aslin, R. N. (2001). Statistical learning in a serial reaction time task: Access to separable statistical cues by individual learners. *Journal of Experimental Psychology: General, 130*(4), 658–680. https://doi.org/10.1037/0096-3445.130.4.658
Kemp, A., & Manahan-Vaughan, D. (2007). Hippocampal long-term depression: Master or minion in declarative memory processes? *Trends in Neurosciences, 30*(3), 111–118. https://doi.org/10.1016/j.tins.2007.01.002
Malenka, R. C., & Bear, M. F. (2004). LTP and LTD: An embarrassment of riches. *Neuron, 44*, 5–21. https://doi.org/10.1016/j.neuron.2004.09.012
McMurray, B., Horst, J. S., & Samuelson, L. K. (2012). Word learning emerges from the interaction of online referent selection and slow associative learning. *Psychological Review, 119*(4), 831–877. https://doi.org/10.1037/a0029872
Næss, K.-A. B., Lyster, S.-A. H., Hulme, C., & Melby-Lervåg, M. (2011). Language and verbal short-term memory skills in children with down syndrome: A meta-analytic review. *Research in Developmental Disabilities, 32*(6), 2225–2234. https://doi.org/10.1016/j.ridd.2011.05.014
Newport, E. L., & Aslin, R. N. (2004). Learning at a distance I. statistical learning of non-adjacent dependencies. *Cognitive Psychology, 48*(2), 127–162. https://doi.org/10.1016/S0010-0285(03)00128-2
Newport, E. L., Hauser, M. D., Spaepen, G., & Aslin, R. N. (2004). Learning at a distance II. Statistical learning of non-adjacent dependencies in a non-human primate. *Cognitive Psychology, 49*(2), 85–117. https://doi.org/10.1016/j.cogpsych.2003.12.002
Nissen, M. J., & Bullemer, P. (1987). Attentional requirements of learning: Evidence from performance measures. *Cognitive Psychology, 19*(1), 1–32. https://doi.org/10.1016/0010-0285(87)90002-8
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, 105*(7), 2745–2750. https://doi.org/10.1073/pnas.0708424105

O'Reilly, R. C., & Rudy, J. W. (2000). Computational principles of learning in the neocortex and hippocampus. *Hippocampus, 10*(4), 389–397. https://doi.org/10.1002/1098-1063(2000)10:4<389::AID-HIPO5>3.0.CO;2-P

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

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

Perruchet, P., & Vinter, A. (1998). PARSER: A model for word segmentation. *Journal of Memory and Language, 39*(2), 246–263. https://doi.org/10.1006/jmla.1998.2576

Pinar, C., Fontaine, C. J., Triviño-Paredes, J., Lottenberg, C. P., Gil-Mohapel, J., & Christie, B. R. (2017). Revisiting the flip side: Long-term depression of synaptic efficacy in the hippocampus. *Neuroscience & Biobehavioral Reviews, 80*, 394–413. https://doi.org/10.1016/j.neubiorev.2017.06.001

Rey, A., Minier, L., Malassis, R., Bogaerts, L., & Fagot, J. (2019). Regularity extraction across species: Associative learning mechanisms shared by human and non-human Primates. *Topics in Cognitive Science, 11*(3), 573–586. https://doi.org/10.1111/tops.12343

Rueda, N., Flórez, J., & Martínez-Cué, C. (2012). Mouse models of down syndrome as a tool to unravel the causes of mental disabilities. *Neural Plasticity, 2012*, Article 584071. https://doi.org/10.1155/2012/584071

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science, 274*(5294), 1926–1928. https://doi.org/10.1126/science.274.5294.1926

Scott-McKean, J. J., & Costa, A. C. S. (2011). Exaggerated NMDA mediated LTD in a mouse model of down syndrome and pharmacological rescuing by memantine. *Learning & Memory, 18*(12), 774–778. https://doi.org/10.1101/lm.024182.111

Sidman, M. (1994). *Equivalence relations and behavior: A research story*. Boston: Authors Cooperative.

Sidman, M., Rauzin, R., Lazar, R., Cunningham, S., Tailby, W., & Carrigan, P. (1982). A search for symmetry in the conditional discriminations of rhesus monkeys, baboons, and children. *Journal of the Experimental Analysis of Behavior, 37*(1), 23–44. https://doi.org/10.1901/jeab.1982.37-23

Smith, L., & Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition, 106*(3), 1558–1568. https://doi.org/10.1016/j.cognition.2007.06.010

Spencer, T. J., & Chase, P. N. (1996). Speed analyses of stimulus equivalence. *Journal of the Experimental Analysis of Behavior, 65*(3), 643–659. https://doi.org/10.1901/jeab.1996.65-643

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, B: Biological Sciences, 372*(1711), 20160056. https://doi.org/10.1098/rstb.2016.0056

Thiessen, E. D., & Erickson, L. C. (2013). Beyond word segmentation: A two- process account of statistical learning. *Current Directions in Psychological Science, 22*(3), 239–243. https://doi.org/10.1177/0963721413476035

Tovar, Á. E., & Westermann, G. (2017a). Computational exploration of lexical development in down syndrome. In G. Gunzelmann, A. Howes, T. Tenbrink, & E. J. Davelaar (Eds.), *Proceedings of the 39th Annual Conference of the Cognitive Science Society* (pp. 1199–1204).

Tovar, Á. E., & Westermann, G. (2017b). A Neurocomputational approach to trained and transitive relations in equivalence classes. *Frontiers in Psychology, 8*. https://doi.org/10.3389/fpsyg.2017.01848

Tovar, Á. E., Westermann, G., & Torres, A. (2018). From altered synaptic plasticity to atypical learning: A computational model of down syndrome. *Cognition, 171*, 15–24. https://doi.org/10.1016/j.cognition.2017.10.021

Wang, F. H., Zevin, J., & Mintz, T. H. (2019). Successfully learning non-adjacent dependencies in a continuous artificial language stream. *Cognitive Psychology, 113*, Article 101223. https://doi.org/10.1016/j.cogpsych.2019.101223

Wilson, B., Spierings, M., Ravignani, A., Mueller, J. L., Mintz, T. H., Wijnen, F., … Rey, A. (2020). Non-adjacent dependency learning in humans and other animals. *Topics in Cognitive Science, 12*(3), 843–858. https://doi.org/10.1111/tops.12381