Transitional Probabilities, Prediction and Backward Associations in Language

Laura Lazartigues, Frederic Lavigne, Fabien Mathy

Transitional Probabilities, Prediction and Backward Associations in Language

Probabilités transitionnelles, Prédiction et Associations Rétrospectives dans le Langage

Laura Lazartigues [a*], Frédéric Lavigne [b], Fabien Mathy [b]

[a] Univ. Lille, CNRS, UMR 9193 - SCALab - Sciences Cognitives et Sciences Affectives, F-59000 Lille, France

[b] Bases, Corpus, Langage (BCL, UMR 7320), Université Côte d'Azur and CNRS, Nice, France

Author Note

Laura Lazartigues https://orcid.org/0000-0002-5361-647X

Frédéric Lavigne https://orcid.org/0000-0003-4493-430X

Fabien Mathy https://orcid.org/0000-0002-7705-5689

The authors declare that they have no conflict of interest to disclose. Correspondence concerning this article should be addressed to Laura Lazartigues, Université de Lille, SCALab, UMR 9193, Université de Lille, Rue du barreau, BP 60149, 59653 Villeneuve d'Ascq Cedex. Email: Laura.lazartigues@univ-lille.fr

Transitional Probabilities, Prediction and Backward Associations in Language

A typical cognitive mechanism involved in language comprehension relates to the processing of sequences of stimuli. To learn sequences, statistical learning enables the computation of transitional probabilities (TP). A TP corresponds to the probability to encounter successive events in a sequence. While forward TP (FTP: probability to encounter B after A in an AB sequence) have been studied extensively, backward TP (BTP, probability to encounter A before B) remain elusive, as few studies have investigated how they are learned. Moreover, these investigations were based only on offline measures, thus leaving aside the issue of how BTP are used in real-time sequential processing, such as in language. We therefore attempted to synthesize the evolution of various statistical concepts in language processing, from cooccurrences in semantic priming to TP, and to highlight the lack of studies on BTP in the context of sequential processing.

Keywords: Statistical learning, Language, Prediction, Transitional probabilities, Backward transitional probabilities

Probabilités transitionnelles, Prédiction et Associations Rétrospectives dans le Langage

Le traitement de séquences de stimuli apparait comme étant central dans la compréhension du langage. Au sein de séquences, l'apprentissage statistique permet de calculer des probabilités transitionnelles (PT). Une PT correspond à la probabilité de rencontrer successivement deux stimuli dans une séquence. Alors que les PT prospectives (PTP : probabilité de rencontrer B après A dans AB) ont été largement étudiées, les PT rétrospectives (PTR, probabilité de rencontrer A avant B) restent méconnues. De plus, les quelques études portant sur les PTR ne se sont basées que sur des mesures hors ligne, laissant ainsi de côté la question de leur utilisation dans le traitement séquentiel. Dans cet article, nous synthétisons l'évolution de divers concepts statistiques dans le traitement du langage, des cooccurrences dans l'amorçage sémantique jusqu'au PT, afin de mettre en évidence le manque d'études sur les PTR dans le contexte du traitement séquentiel.

Mots-clefs : Apprentissage statistique, Langage, Prédiction, Probabilité transitionnelle, Probabilité transitionnelle rétrospective

Sequence processing is a key cognitive mechanism that plays a central role in language. Sequential information helps to predict a stimulus based on previous one(s) more effectively. Accurate predictions result in reduced processing times and fewer errors on stimuli that can be anticipated (Brunellière et al., 2022; DeLong et al., 2005; Willems et al., 2016). Moreover, enhanced predictions lead to higher language skills, including smoother reading aloud (Gavard & Ziegler, 2022), facilitated speech perception in poor listening conditions (Conway et al., 2010), and accelerated reading with more predictable words being read more rapidly (Albrengues et al., 2019; Frisson et al., 2005; Zang et al., 2023). However, although prediction can improve comprehension (Pickering & Gambi, 2018), it is not a necessary component of language learning and comprehension (Huettig & Mani, 2016). Since the ability to predict sequential events relies primarily on the statistical regularities between the ordered stimuli of a sequence (Erickson & Desimone, 1999), the main issue is to establish which statistical regularities can be learned. In this article, we first review the literature that primarily focused on the frequency of cooccurrence of two words. The words may be related (frequently associated) or unrelated (not frequently associated) in the semantic priming paradigm. We posit that transitional probabilities play a crucial role in providing a more nuanced description of the associations between words. This should contribute to a deeper comprehension of the connections between sequence processing and statistical regularities. The second part of this article examines the researches on transitional probabilities related to statistical learning in the context of sequence processing, especially backward associations.

Semantic priming: the role of association strength in prediction and underlying mechanisms

Prediction effects in language have been highlighted using the semantic priming paradigm, in which a pair of words is presented successively (Meyer & Schvaneveldt, 1971;

see Neely, 1991). The pair can either be semantically related or unrelated (e.g., cat – mouse *vs*. sun – mouse). The degree of relationship corresponds to the strength of association between the words which, at this stage, is calculated by the frequency of the word pair. In this paradigm, participants must perform a lexical decision task to decide whether the target word is a real word or a pseudoword. Results have shown that participants identify target words faster when they are preceded by semantically related words (Brunel & Lavigne, 2009; Neely, 1991, for reviews). This suggests that the processing of a word is more efficient when the latter is strongly associated with a previously processed word. In addition to the semantic priming effect, a syntactic priming effect also seems to play a role in tasks such as lexical decision (Goodman et al., 1981) and reading (Gavard & Ziegler, 2022). This effect is characterized by shorter response times when a target word is preceded by a syntactically appropriate word (e.g., "this mouse") compared to a priming word that is not syntactically appropriate (e.g., "she mouse"). However, a key difference between semantic and syntactic priming is that syntactic priming effects disappear with a short presentation time, whereas semantic priming effects persist (Gavard & Ziegler, 2022).

Although semantic priming effects remain observable at shorter presentation times, their magnitude can be influenced by the time lapse between the presentation of the two words (technically called *stimulus onset asynchrony*). This delay influences the strength of preactivation: priming effects are present with short SOAs (250ms), yet more significant with longer SOAs (1,000ms) in the classical semantic priming paradigm (Coney, 2002). The type of association between words also plays a role. As evidenced by both behavioral and EEG data, a direct association, like ‘tiger – stripes’ (step-1 association), is reported to yield a larger priming effect than an indirect one, such as ‘lion – stripe’ (Chwilla et al., 2000). The latter is referred to as a step-2 association and is considered indirect, as it requires an intermediate

concept to link the two words: lion – (tiger) – stripe. Overall, the more direct the association between two concepts, the stronger the preactivation (Brunel & Lavigne, 2009 for a review).

To better understand the interaction between these factors (association strength and SOA), three mechanisms acting concomitantly were described by Neely (1991). *(i)* The first mechanism involves automatic spreading activation between associated concepts, possibly linked to the priming effect observed within short SOAs. According to this mechanism, two stimuli presented close together in time are associated in memory (Hebb, 1949, 1961). Consequently, the activation of the first word leads to automatic propagation of this activation towards stimuli associated with it in long-term memory. This is a top-down mechanism, since it is the preactivation in long-term memory that leads to faster processing of stimuli related to the first word. Facilitation also depends on the strength of association between the two words (Luka & Van Petten, 2014). This mechanism is fast, does not require conscious effort, and does not inhibit stimuli unrelated with the priming word. *(ii)* The second mechanism is expectation-based priming and is also top-down. The presentation of the first word generates an expectation on the set of stimuli likely to follow, facilitating the quicker processing of a target in this set. Unlike the first mechanism, this one operates more slowly and is associated with the late priming effect observed with long SOAs. Neely (1991) suggested that this mechanism tends to inhibit stimuli which are not associated with the first word and not part of the expected set. These two mechanisms have been modeled together based on unified neuronal processes (Brunel & Lavigne, 2009; Lavigne et al., 2011, 2012, 2013). *(iii)* The third mechanism is the postlexical priming mechanism (or semantic matching), which can operate after the presentation of two words. Although it is not predictive per se, it allows for a plausibility check of the association between the first word and the target *a posteriori*. This mechanism makes it possible to check whether the predictions made before the encoding of the target have occurred or not, akin to the retrodiction process described by Jones and

Pashler (2007). This postlexical mechanism requires consideration of the target after its encoding, whereas the two previous processes are active even before the target is encoded. Unlike top-down predictive mechanisms, this third one (explored under the terms 'integration' or 'semantic integration', Hagoort et al., 2004) is considered to be a bottom-up process that assesses the plausibility of the context in light of the target word (Mantegna et al., 2019; Nieuwland et al., 2020). Distinguishing between a predictive and an integration process is challenging as both processes are expected to yield similar results in certain cases. For instance, a related word presented after a prime is assumed to be both well predicted and integrated. However, the existence of these two processes in language has been highlighted by distinct temporal patterns in event-related brain potentials (ERP), with earlier effects observed in predictive processes (Nieuwland et al., 2020).

To illustrate these three mechanisms, let us consider the sentence "It is on the desk". As we hear "It is on the…", the first two mechanisms come into play to anticipate the next word. The first mechanism rapidly preactivates all potential words corresponding to this sentence, such as "table" or "cupboard". The second mechanism generates expectations about what the next word might be but inhibits words that are not related to the sentence, such as "sky" or "cat". Finally, when hearing the end of the sentence "It is on the desk", the third mechanism allows us to assess the plausibility of the word "desk" given the context. Regarding the first two processes, Perruchet (1985) showed that behavioral responses depend on the associative strength between the stimuli and the associated implicit predictions (as described by the first mechanism above) rather than on explicit expectations (as in the second mechanism above). The eponym 'Perruchet effect' underlines the primary importance of association strength in learning (Perruchet, 2015). However, a more recent study by Destrebecqz et al. (2019) highlighted the influence of association strength, expectations and

motor priming on performance, in line with the processes described by Neely (1991). Altogether, all these studies highlight the primary importance of association strength.

In standard semantic priming studies, the strength of association is determined by the frequency of word cooccurrences. However, when examining the neural basis of associative learning, it becomes clear that cooccurrence frequency is not sufficient to fully describe association as a general mechanism. If we refer to the neural foundation of associative learning as described by Hebb (1949, 1961), an association is created by repeated occurrences of words, leading to simultaneous (or close in time) activation of two populations of neurons (coding for stimuli in the environment).

Let us take the example of an AB association, in which each letter corresponds to a specific word. If AB frequently appears in the environment, A and B will activate their respective populations of neurons and this coactivation will lead to long-term potentiation of the synapse between A and B. This corresponds to an increase in synaptic efficacy between the two populations of A and B neurons (Lavigne et al., 2011), which constitutes the core mechanism underlying an association. This first process, illustrated by the expression “Neurons that fire together wire together”, is consistent with the idea that associations are linked to the frequency of word cooccurrences: the more frequently a pair of words is repeated, the stronger the association. However, Hebb's theory includes another crucial process. When a stimulus from a previously established association appears alone or in a different association, the asymmetry of activation — where one population of neurons is active while another is not — leads to long-term depression. For instance, after learning AB, encountering AC in the environment activates the populations coding for A and C, but not those for B. In the synaptic links between A and B, the asymmetry between the activated A and the non-activated B induces long-term depression, thus weakening the AB association.

Consequently, an association between two words cannot be solely explained by cooccurrence alone: it also requires consideration of the occurrences of all connected words in their different contexts. Therefore, the frequency of cooccurrences must be examined in light of the frequency of each word separately. Importantly, the value of the association between two stimuli calculated on the basis of these Hebbian mechanisms corresponds to a specific statistical regularity: transitional probability.

Transitional probabilities

Transitional probabilities (TP) make it possible to precisely define an association by considering not only the frequency of cooccurrence but also the frequency of each stimulus. In this respect, the study of their impact on prediction is in line with studies on semantic priming. In addition, we posit that TP can be associated with the three mechanisms proposed by Neely (1991). The first two predictive mechanisms can be associated with forward TP (FTP), enabling the quantification of the strength with which A predicts B in an AB sequence (*p(B|A)*). An FTP is therefore the number of times B appears after A divided by the number of times A appears. TP thus provides new information about association strength since it refers to the absolute frequency of A, which was not the case before. Conversely, the third mechanism, based on the verification of the plausibility of a context given a target, can be linked to backward TP (BTP). BTP corresponds to the probability of encountering an A stimulus *before* a B stimulus (*p(*A|B*)*). Consequently, it corresponds to the process of assessing the plausibility of context *A* given the presence of stimulus B.

$$\text{FTP:}\quad p(\text{B}|\text{A}) = \frac{frequency(\text{AB})}{frequency(\text{A})} \qquad \text{BTP:}\quad p(\text{A}|\text{B}) = \frac{frequency(\text{AB})}{frequency(\text{B})}$$

The literature has tended to put emphasis on FTP, to the extent that the generic term 'transitional probability' has been used to designate them. In contrast, BTP have scarcely been addressed, although both types of TP are thought to be involved in the same associations. In the present article, we first focus on FTP and then highlight the lack of studies on BTP, which represents the current challenge in the field of statistical learning.

Forward transitional probabilities

Seminal studies

The seminal studies that initially introduced the term TP focused on the role of FTP in speech segmentation in both adults (Saffran, Newport, et al., 1996) and eight-month-old infants (Saffran, Aslin, et al., 1996), using an artificial language paradigm. They laid the methodological foundations that have since been widely applied to investigate the statistical learning of TP.

In the adult study, Saffran and colleagues (1996) used an artificial language consisting of a set of six nonsense words, each comprising three syllables. FTP values had been fixed at a specific value between .31 and 1 between syllables forming a word, e.g., the word "pabiru", where "pa" was always followed by "bi", which in turn was always followed by "ru" (p(bi|pa) = 1 and p(ru|bi) = 1). However, some syllables appeared in several words, leading to lower TP. In words such as "gobaku", and "tibada", the syllable "ba" appeared in various contexts, leading to lower TP: p(ba|go) = .5 and p(ba|ti) = .5, respectively. Words were presented using an electronic voice in a pseudorandom order to avoid them being repeated twice in a row. This pseudorandom order resulted in low FTP between syllables not belonging to the same word, ranging between .1 and .2. The familiarization phase consisted of a 21-minutes passive listening to continuous speech (three blocks of seven minutes). Then, learning was tested with

a two-alternative forced-choice (2AFC) test. Participants heard two tri-syllabic sequences, one being a word heard during the familiarization phase, the other being a lure. Two types of lures were used: nonwords and partwords. A nonword corresponded to a new sequence comprising three syllables present in the familiarization phase but never presented altogether ($p = 0$ for the FTP between the syllables making up the nonword, e.g., "pabalu" if "**pa**biru", "go**ba**ku" and "tiva**lu**" were the words presented in the familiarization phase). A partword was composed of the final syllable of a word and the first two syllables of another word, like "rugoba" ("pabi**ru**" and "**goba**ku"), or of the last two syllables of a word and the first syllable of another word, like "birugo" ("pa**biru**" and "**go**baku"). Participants had to indicate which of the two strings sounded more like a word heard during the familiarization phase. The results indicated that participants chose significantly more words than nonwords or partwords. They also performed better with words involving higher TP. These results highlight the involvement of statistical learning in speech segmentation, and more particularly the use of TP to determine boundaries between words.

The involvement of statistical learning and TP in speech segmentation was then studied by the same team in eight-month old infants (Saffran, Aslin, et al., 1996). In this experiment, the simplified material consisted of four words repeated in a pseudorandom order. Consequently, the FTP between words were 1/3 ($p = .33$). Saffran et al. (1996) also used the familiarization-preference procedure (Jusczyk & Aslin, 1995). Participants were familiarized with a two-minute continuous speech stream. After that, learning was tested with a head-turn preference procedure. The test phase included words and nonwords (Experiment 1) or words and partwords (Experiment 2). The results indicated a longer listening time for nonwords and partwords than for words. The study by Saffran et al. (1996) thus showed for the first time that participants can use TP to determine which triplets of syllables do or do not correspond to a word.

These studies used the term ‘statistical learning’ to describe the mechanism for extracting statistical regularities from the environment, such as TP. The term statistical learning has since been used for various types of stimuli (artificial speech, Bonatti et al.; 2005; shapes, Brady et al., 2009; shapes sequences, Bertels et al., 2012; tactile sequences, Conway & Christiansen, 2005; musical sequences, Tillmann & McAdams, 2004) to describe more broadly the neurocognitive mechanism that enables us to detect and learn a range of regularities, leading to persistent memories (Kóbor et al., 2017). However, the link between statistical learning and language has raised particular interest, as illustrated by a few examples.

Role in language

Speech segmentation. Since the initial studies by Saffran et al. (1996), subsequent research has deepened our understanding of the role of FTP, demonstrating the rapid learning of TP after only one or two exposures (Batterink, 2017; Hao Wang et al., 2023), assessing a specific effect of TP while controlling for frequencies (Aslin et al., 1998), indicating potential generalization across variations (Graf Estes, 2012) or across other structures (Gomez & Gerken, 1999), and also extending findings to natural language. The study by Pelucchi et al. (2009b) used Italian sentences to investigate whether monolingual English-learning infants could extract words from natural language speech based on TP. Target words were repeated in the speech and their constituting syllables were always presented together, resulting in a TP of 1 between them. Infants were tested using a similar procedure as in the study by Saffran, Aslin, et al. (1996). The results indicated a preference for target words over novel words (Italian words previously unheard), thus demonstrating that statistical learning occurs not only in a highly controlled artificial language but also in more natural speech. The limits of speech segmentation based on statistical learning were then tested by including two distinct artificial

languages to be learned (two sets of pseudowords) within a single speech flow. The results highlighted the robustness of statistical learning, since participants (under specific conditions) were able to extract distinct sets of TP from a single stimulus stream (Weiss et al., 2009). However, when two artificial languages were presented sequentially (first language followed by the second), a primacy effect appeared and only the first set was learned (Bulgarelli et al., 2017; Karuza et al., 2016). To teach a second language successfully using a sequential presentation, it is necessary to use a contextual cue marking the change of language (such as a pause between the two artificial languages) and a duration of presentation longer than that required for learning a single language (Gebhart et al., 2009). The difficulty of learning a second language may be attributed to an anchoring effect consisting in remaining fixated on the first language while ignoring the second. Learning a first artificial language thus hinders the learning of the second. However, this anchoring effect can be mitigated by introducing language changes at the same time as the first language is learned (Bulgarelli & Weiss, 2016). The overall results of these studies support the idea that statistical learning is a prevalent mechanism in handling minimal aspects of language acquisition (Romberg & Saffran, 2010). However, the recent meta-analysis by Isbilen and Christiansen (2022) warned that most studies used a TP equal to 1, whereas statistical regularities in language are far more varied. They argued that future studies need to test more diverse values of TP to deepen our understanding of how statistical learning operates.

Reading. Statistical learning ability has also been shown to be generally related to reading ability (Apfelbaum et al., 2013; Arciuli & Monaghan, 2009; Treiman & Kessler, 2006). A recent meta-analysis by Ren et al. (2023) including 42 studies indicated a correlation between statistical learning and reading in both adults and children (yet with stronger effects in adults), for all statistical learning paradigms and across modalities. Interestingly, the link between statistical learning and reading is more significant in deep-orthography languages

(i.e., languages with a complex mapping between phonemes and graphemes) than in shallow-orthography ones (Ren et al., 2023). The authors explain this effect by the necessity to use more resources to efficiently learn the more complex grapheme/phoneme mapping existing in deep-orthography languages. The general link between statistical learning and reading was illustrated by Arciuli and Simpson (2012), who found a correlation between a reading test (reading aloud words from the WRAT-4, Wilkinson & Robertson, 1993) and a statistical learning task using nonlinguistic visual stimuli governed by FTP, as described in Saffran, Aslin, et al. (1996). Moreover, studies investigating the effect of various predictors (including TPs) on eye movements during reading have shown an effect of FTP, with shorter fixation times with higher TP (Demberg & Keller, 2008; McDonald & Shillcock, 2003), but no effect of FTP when words were presented one by one on the screen (Onnis et al., 2022). However, the effect of BTP during reading is inconsistent across studies, with higher BTP leading to either faster (McDonald & Shillcock, 2003; Onnis et al., 2022) or slower reading (Demberg & Keller, 2008). Since this effect remains elusive, further studies on BTP are required.

Syntax. A final example of the involvement of statistical learning in language is the acquisition of syntax. FTP associate grammatical groups, which allows for better learning of grammatical rules (Saffran, 2001; Thompson & Newport, 2007). The study by Saffran and Wilson (2003) used a classical familiarization and head-turn preference procedure with 11- to 12-month-old infants, who had to listen to 16 sentences of an artificial language, each sentence consisting of five disyllabic nonsense words. After the familiarization phase, grammatical and ungrammatical sentences were tested. The results showed that participants were able to distinguish between both type of sentences. According to the authors, these results suggest that infants can acquire two levels of structure: the word level (extracting words from speech) and the syntactic level (extracting grammatical rules). The role of TP in a

sentence-sized sequence thus confirms the importance of considering much more than the association between two stimuli alone.

More context for better prediction

Until 1996, FTP had mostly been studied using a single stimulus predicting another (i.e., *p*(B|A) in an AB sequence). However, more complex associations in larger sequences are sometimes present in the environment (as in languages) and should also be taken into account. The predictive benefit offered by a richer context including several stimuli (or several words) has been tested by the *n*-gram model used in automatic natural language processing (Bahl et al., 1983). This model is derived from Markov chains and accounts for the role of bigrams (set of two stimuli), trigrams (set of three stimuli) and n-grams (set of n stimuli) in prediction. The *n*-gram model shows that trigrams provide the best prediction within a sequence of four stimuli (Tremblay & Tucker, 2011).

However, the *n*-gram model only considers the absolute frequency of *n*-gram, while TP are more specific. As explained earlier, a TP corresponds to the absolute frequency of a cooccurrence divided by the frequency of a specific element, thus providing the probability of a stimulus *given* another. FTP should therefore help to predict the next stimulus in the sequence. However, the standard paradigm of statistical learning generally uses offline measures and does not make it possible to account for the use of FTP in prediction, or for the evolution of learning in real time.

Various tasks can be used to test various types of mechanisms (Siegelman et al., 2017). Offline tasks, such as the classical 2AFC used in the statistical learning field, mostly rely on recognition processes, while online tasks rely essentially on prediction (Lazartigues et al., 2023). The use of online measures thus seems crucial to test the involvement of TP in

sequence processing and to better understand the trajectory of learning and its evolution in real time (Siegelman et al., 2018).

Online measures of sequential processing and prediction

Initially, online measures were obtained using a serial-response task (SRT) (Nissen & Bullemer, 1987), during which sequences were presented on a screen and each stimulus was associated with a key on a computer keyboard. Participants were asked to press a key corresponding to the stimulus as quickly as possible, while making as few errors as possible. For each trial, response times and accuracy were recorded, thus tracking the evolution of learning as the sequence was processed.

Studies using the SRT paradigm showed quicker response times as the task unfolded, particularly when the sequences were constructed in such a way that the next stimulus could be predicted thanks to FTP (Cleeremans & McClelland, 1991). This prediction applied to both perfectly predictable sequences ($p = 1$) and moderately predictable ones ($p = .5$), although faster response times were observed for the highest TP (Hunt & Aslin, 2001). This paradigm can be used to examine the effect of a broader context on prediction, as shown by the study of second-order TP, which is the probability of a stimulus given the combination of two stimuli: in an ABC sequence, stimulus *C* can be predicted by both stimuli *A* and *B* separately (such as p(C|A) and p(C|B)) but also by the *AB* set corresponding to the combination of two stimuli (such as p(C|AB)). To clarify the differences between these cases, the prediction of a stimulus by a single other stimulus has been called first-order TP, while the prediction of a stimulus by a combination of two stimuli has been called second-order TP (also called second-order dependencies, Gomez, 1997 or SOCs for Second-Order Conditionals, Deroost et al., 2010; Reed & Johnson, 1994). Using an artificial language paradigm, it has been shown that second-

order TP are more difficult to learn than first-order ones (Gomez, 1997). However, the use of online measures has indicated that second-order TP may yield better prediction of the target.

Schvaneveldt and Gomez (1998) aimed at better understanding the role of attention in learning first-order TP (Experiment 1) and second-order TP (Experiment 2) in a SRT task. They used two conditions in their experiments: the SRT task alone or the SRT task plus tone counting. In Experiment 1, testing first-order TP, two four-item sequences were used (1-2-4-3 and 1-3-4-2), with one sequence being probable (p = .80) and the other improbable (p = .20). The results showed that higher FTP led to shorter response times in both conditions, suggesting that first-order FTP can be processed without using attention, although attentional load led to higher errors. Experiment 2 investigated second-order TP with the same protocol but involved two 12-item sequences for perfect dissociation of first- and second-order TP (1-2-1-3-4-2-3-1-4-3-2-4 and 1-2-3-4-1-3-2-1-4-2-4-3). In these sequences, a single stimulus did not predict the next one, but a combination of two stimuli did (e.g., in the first sequence, 2 can predict 1, 3 and 4, as in 1-**2-1**-3-4-**2-3**-1-4-3-**2-4**, but in the same sequence, each pair of stimuli predicts the next stimulus with certainty, e.g., 2 and 1 together predict 3, as in 1-**2-1-3**-4-2-3-1-4-3-2-4). Once again, one sequence was probable (p = .90) while the other was improbable (p = .10). The results showed the use of second-order TP to better anticipate the next stimulus. However, attention load led to higher errors. This suggested a more difficult learning of second-order than first-order TP, as also observed in online measures with a separation between the probable and improbable sequences requiring more trials than in Experiment 1. Such measures are useful to test the evolution of learning, although SRT tasks have the disadvantage of requiring the learning of an association between a stimulus and a response key.

To tackle this issue, Minier et al. (2016) used an alternative paradigm in which monkeys did not need to learn a stimulus/touch association. The authors created a task in

which the monkeys only had to touch stimuli on a touchscreen. During the experiment, the screen was divided into a virtual 3×3 grid of nine locations filled with a cross. Each stimulus corresponded to a red dot replacing a cross on a specific grid position. As soon as a red dot was touched, it disappeared and another red dot appeared in one on the remaining locations. Response times (RTs) were recorded between each stimulus and a decrease in RTs over the course of the experiment was assumed to reflect the evolution of TP learning. Shorter RTs could also suggest an effect of TP during prediction. In their study, Minier et al. (2016) used three deterministic sequences, each composed of three items, repeated over 2,000 trials. A trial corresponded to the three sequences presented in random order. Their results showed a decrease in RTs throughout the experiment, suggesting efficient statistical learning of the sequences. According to the authors, this paradigm helped to track pattern extraction in real time. It seems to provide the most direct measure of prediction in behavioral experiments.

This protocol has already been used to investigate statistical learning of both first-order and second-order FTP in both humans and monkeys. Lazartigues et al. (2021) tested the interactive effects of first-order TP, second-order TP and frequency in prediction. They used a set of four three-item sequences in which all second-order TP were deterministic ($p = 1$), while first-order TP were not ($p < 1$). Different frequency values were assigned to each sequence by manipulating the number of repetitions of each sequence in the experiment, thus leading to a modification of the values of first-order TP (since frequency is a part of the calculation of TP). The experimental design also produced a partial dissociation between frequency and first-order TP, allowing us to test the respective effect of these factors. Results showed that second-order TP could be used to make predictions and that participants preferentially used FTP rather than frequency. Using the same paradigm in their Experiment 1, Lazartigues et al. (2023) tested various conditions involving first-order or second-order TP. Each condition included four three-item sequences in which TP were controlled. In the first-

order TP condition, one stimulus predicted the next one, while in the second-order TP condition, only the combination of the two first stimuli predicted the last stimulus of a sequence. Results showed a larger decrease in RTs throughout the experiment with first-order TP than with second-order TP, suggesting an easier learning of first-order than of second-order FTP in sequence processing in humans.

Interestingly, the condition used by Lazartigues et al. (2023) to test the learning of second-order TP was also used by Rey et al. (2022) in monkeys (*Guinea baboons papio papio*). Using a similar protocol, their results showed that monkeys could also use second-order FTP. Similar statistical learning across species suggests similar mechanisms. According to Rey et al. (2022), Hebb-based computational models (as in Endress & Johnson, 2021 or Tovar & Westermann, 2023) may provide a plausible explanation for the mechanisms underlying statistical learning in both humans and monkeys. This idea of a common mechanism between species could therefore be confirmed by online measures. The simplicity of the touch-screen pointing protocol not only made it possible to test different species using the same method, but also made it possible to explore various regularities (i.e., TP and frequency) in prediction and statistical learning. We therefore believe that studying the effect of backward TP in sequence processing using these online measure protocols should improve our understanding of the mechanisms involved in statistical learning, sequence processing and language processing.

Backward transitional probabilities

In a pair of AB stimuli, we have shown the importance of FTP from A to B in prediction. However, in the same pair, another type of association strength is present: backward TP. Since FTP depends on the absolute frequency of A (p(B|A) = frequency(AB)/frequency(A)) and BTP depends on the frequency of *B* (p(A|B) =

frequency(AB)/frequency(B)), FTP and BTP can be different in the same pair of stimuli. In some cases, BTP may thus provide more information than FTP: in the sequence "the cat", FTP is low because "the" can be followed by several words, but the BTP of having "the" before "cat" is very high. The construction of English language itself thus generally leads to high BTP and low FTP (Onnis & Thiessen, 2013). Since BTP provides more information than FTP, the BTP may be an interesting regularity to use during language processing.

Based on this idea, the Chunk-Based Learner computational model (CBL; McCauley & Christiansen, 2019) uses BTP as a main parameter to study language comprehension and production. It aims to account for the identification and use of multi-word units in children's initial comprehension and production. To this end, the model creates an inventory of chunks – called ‘chunkatory’ – with one chunk corresponding to one or several words. To create a chunk, the model computes the BTP between two input words. If the BTP is greater than the threshold corresponding to the running average, the two words are chunked together, placed in the chunkatory and kept there whenever repeated. Otherwise, a boundary is placed between the two words and the model moves on to the next word. When processing speech, the model computes BTP but also uses the chunks available in the chunkatory to predict the following word. The chunkatory is then used to assess comprehension and production based on multi-word chunks. In this respect, BTP is assumed to be a fundamental parameter in language. In addition, computational models such as TRACX (truncated recursive auto-associative chunk extractor) can extract words from a speech stream based on the BTP (French et al., 2011). TRACX is a three-layer auto-associator with a hidden layer built for implicit chunk recognition. This process automatically groups frequently encountered elements together. Simulations have shown that TRACX can chunk sequences and recognize them as words based on both FTP and BTP (French et al., 2011).

Although these simulations have indicated a possible or necessary use of BTP during language processing, very few behavioral studies have been conducted. Previous studies have indicated a possible backward association learning within nonlinguistic stimuli in both humans (Arcediano et al., 2003) and animals (Chartier & Fagot, 2022; Soares Filho et al., 2016), but the question of learning BTP via a specific mechanism of statistical learning with linguistic material has been addressed only a few times in recent decades.

Language studies

The various studies on BTP related to language can be classified into four different categories: (1) speech segmentation in artificial language and (2) in natural language, (3) BTP in reading, and (4) studies investigating the link between participants' first language and the ability to learn BTP.

Speech segmentation in artificial language. Perruchet and Desaulty (2008) were the first to test the effect of BTP on speech segmentation. They used a standard statistical learning paradigm comprising an eight-minute familiarization phase during which participants passively listened to the artificial language. They then performed a 2AFC task in which they had to choose the most familiar sequence between a word and a partword. In their Experiment 1, two conditions were used to assess the effect of one type of TP (BTP or FTP): in Condition 1, within-word FTP was equal to 1 and between-word FTP was equal to .11, but both BTP were equal to .33. Conversely, Condition 2 involved the exact opposite pattern, with within-word BTP equal to 1 and between-word BTP equal to .11, while both FTP were set at .33. In summary, Condition 1 tested FTP at a fixed value of BTP while Condition 2 tested BTP at a fixed value of FTP. The results indicated that participants chose words over partwords, with performance above chance in both conditions. These results are consistent with previous studies on FTP, and therefore suggest that BTP can be used in speech segmentation. Perruchet

and Desaulty (2008) replicated these results in their Experiment 2, in which they controlled syllable frequency to ensure that the results of Experiment 1 were indeed due to TP rather than frequency. This control led to a modification of TP in each condition. In Condition 1, within-word FTP was equal to 1, but between-word FTP was equal to .20, while within-word BTP was equal to .20 and between-word BTP was equal to .33. In Condition 2, the pattern was opposite (within-word BTP = 1, between-word BTP = 20, within-word FTP = .20 and between-word FTP = .33). The results of Experiment 2 showed that, even with controlled frequency, participants were able to distinguish words from partial words on the basis of FTP and BTP, once again suggesting the effective learning of both types of TP.

Speech segmentation in natural language. To further investigate the role of BTP in speech segmentation, Pelucchi et al. (2009a) carried out a study to test the statistical learning of BTP using natural language in infants. English monolingual infants were exposed to a language consisting of correct Italian sentences in which BTP and FTP were manipulated by repeating four target words six times. Two of these words were created as high-TP words and their constituent syllables did not appear in other words, so the TP (both FTP and BTP) in these words was equal to 1. For instance, “fuga” and “melo” were high-TP words, so the syllables “fu”, “ga”, “me” and “lo” never appeared in other words. The other two target words, for instance “bici” and “casa”, were low-TP words, of which the last syllables “sa” and “ci” appeared in other words, leading to a BTP equal to .33 while FTP remained equal to 1. Therefore, the only difference between high-TP words and low-TP words was their BTP. The infants listened to this language for three minutes, then completed a task using the head-turn preference procedure (Saffran, Aslin, et al., 1996). Results indicated a preference for high-TP words, suggesting a sensitivity to BTP, although FTP were equal to 1 in both conditions. These results are consistent with those of Perruchet and Desaulty (2008) and support the theory that BTP are used in speech segmentation. Moreover, Pelucchi and colleagues (2009a)

demonstrated that statistical regularities such as FTP and BTP can be successfully extracted from natural speech, which is much richer and more complex than an artificial language, thus providing more ecological data for the comprehension of statistical learning.

Another natural language study by Hay et al. (2011) tested the effect of FTP and BTP in a word-meaning mapping task. The authors used the same protocol as Pelucchi et al. (2009a): English monolingual children were tested with Italian speech including four target words that could either be high-TP (p = 1) or low-TP (p = .33). After this speech segmentation task, the children were administered a label-object association task in which they were familiarized with associating a label, corresponding to one of the words presented during the segmentation phase, with a 3D object. During this task, the object was presented and moved across the screen to retain the children's attention while the label was broadcast in loudspeakers. The habituation phase stopped either after 20s or after the children had looked away from the screen for one second. Two conditions were used: the first once used high-TP words as labels while the second used low-TP words. After the habituation phase, the children were tested and their looking time was recorded. Two types of trials were used: same-test trials, involving learned label-object associations, and switch trials, in which the labels of the two objects were switched. The hypothesis was that if children actually learned the label-object association, the results should show a difference in looking times between same-test and switch trials. Results indicated a longer looking time in the switch than the same-test trials, both with high-TP and with low-TP words in Experiment 1 (high-TP: FTP and BTP = 1; low-TP: FTP = .33, BTP = 1), but only with high-TP words in Experiment 3 (high-TP words: FTP and BTP = 1; low-TP words: FTP and BTP = .33). These results suggest that TP (both FTP and BTP) do enable children to segment speech to extract words. When the TP are sufficiently high (high-TP words), children manage to associate the word with an object. Conversely, if the TP within the word are too low (low-TP words), children fail to associate

the label with an object. In this respect, the TP make it possible to determine which set of syllables is likely to be a word and can therefore be associated with a meaning. These results also suggest that BTP plays a role in speech segmentation and learning new words, since children were able to learn a label-object association when BTP were equal to 1 (Experiment 1), but not when they were equal to .33 (Experiment 3).

Reading. We previously focused on the ability to learn BTP, but its actual involvement in language and more generally in processing sequences is another crucial issue. So far, three studies have addressed this question in the field of reading. Onnis et al. (2022) tested prediction (with the predictive strength being determined by a calculation involving FTP) and integration (determined by a calculation involving BTP) using a word-by-word reading task in which each word of a given sentence appeared on a screen. Participants had to touch the space bar to move on to the next word. An effect of BTP was observed, with higher BTP leading to shorter reading time, but no significant FTP effect appeared. However, in a more natural reading task, the results were different, with a significant effect of both FTP and BTP, leading to a shorter reading time with higher TP. Demberg and Keller (2008) and McDonald and Shillcock (2003) also investigated the role of certain predictors on eye movements during reading, including TP. Regarding BTP, the results are inconsistent between studies. Both McDonald and Shillcock (2003) and Onnis et al. (2022) found that high BTP led to faster reading with shorter fixation time, while Demberg and Keller (2008) found that high BTP led to higher reading time. The results of Onnis et al. (2022) on natural reading also showed that the effect of BTP was higher when the FTP were low, suggesting that participants used BTP more when FTP were not helpful. Interestingly, although they used the same simple reading task, the calculations used to determine TP were not strictly comparable between these three studies. However, these calculations difference do not explain the inconsistent results between McDonald and Shillcock (2003) and Demberg and Keller (2008).

The role of BTP in reading and sequence processing therefore needs to be further investigated.

First language and ability to learn BTP. Although the above-mentioned studies suggest that BTP can be learned, the study by Onnis and Thiessen (2013) investigated the possibility of a bias induced by prior language on statistical learning. By using a large corpus, the authors first established that Korean and English involve different patterns of TP. The most informative TP seems to be BTP in English and FTP in Korean. Onnis and Thiessen (2013) posited that participants might be biased by their first language when learning FTP and BTP. To test their hypothesis, they performed experiments using a classical statistical learning paradigm, involving a familiarization phase in which BTP and FTP were manipulated, and a 2AFC task. Their results indicated a preference for BTP among English speakers and for FTP among Korean speakers, suggesting a first-language bias. Another study by Thiessen et al. (2019) indicated that although no first-language preference exists at seven months, it appears during the first year of life, as early as 13 months. This highlights the need to carefully consider the results of other studies in light of this bias.

Future research directions using online measures

Overall, the studies presented in this article suggest that BTP can be learned by both children and adults to segment speech, and that it may be involved in reading, even though results are inconsistent. However, there have been few experiments on the topic (Table 1). We believe that two limitations can be underline in the field of statistical learning of BTP.

First, most of the above-mentioned studies used offline measures, which involve a passive familiarization phase before the test phase. This type of measure provides no information on the evolution of learning or on the actual use of BTP *during* sequence processing. Online measures could address both of these issues, making it possible both to

observe the evolution of learning in real time for comparison with FTP and to assess whether BTP is used in sequential processing. Such a test of BTP during sequential processing could help achieve a better understanding of both the statistical regularities used and the processes involved in sequential processing (i.e., the mechanisms of Neely, 1991, and prediction and integration mechanisms). Using a different paradigm seems essential to better investigate the mechanisms involved in statistical learning. The use of different tasks, although carried out with the same material, leads to highlighting different processes and differential use of the factor tested (Lazartigues et al., 2023). In addition, the finding that BTP is involved in a recognition mechanism, as tested by the 2AFC task, does not necessary mean that BTP is used during sequence processing. Alternative statistical learning paradigms may be used, such as the SRT (Nissen & Bullemer, 1987), the alternating SRT (Howard Jr & Howard, 1997) or the pointing task (Lazartigues et al., 2021; Minier et al., 2016), which all involve online measures to track BTP learning in real time.

Second, the first-language learning bias when testing BTP using linguistic materials is another limitation (Onnis & Thiessen, 2013; Thiessen et al., 2019). Most of the studies presented in Table 1 involve English- or French-speaking[1] participants, which are languages that bias speakers in favor of learning BTP (Onnis & Thiessen, 2013). The results obtained for these languages may thus not be generalizable, and this bias must be taken into account when analyzing the results of the studies. The link between the statistical properties of a language and the ability of speakers to learn certain statistical regularities should also be further investigated. The fact that one's first language can lead to sensitivity to specific statistical regularities has strong theoretical implications. This may suggest that statistical learning

[1] To our knowledge, no corpus study has yet been carried out to demonstrate this point. Nevertheless, given the structural similarities in word order between English and French (Östling, 2015), we believe that French is also oriented towards BTP.

ability is shaped at the individual level by one's environment, as is the case with the perception of phonemes as children develop (Pena et al., 2012).

Finally, the concepts of backward associations and BTP may be confusing. It is important to distinguish between backward association, which is the idea that if you learn A – B you should also automatically learn B – A, and BTP, which is a statistical regularity corresponding to the probability of having stimulus A before stimulus B in an AB sequence. Chartier and Dautriche (2023) present backward association in the learning of word-object associations. They highlight that studies on this subject can lead to spurious order effects by presenting the word and object sequentially, which may not account for real phenomena. In fact, it is much more common for the word and the object to be presented simultaneously in the environment. The authors thus emphasize that backward associations may say nothing about language (or at least about the learning of word-meaning associations). BTP is not simply a backward association created by a manipulation as part of an experiment. In the case of language, a spoken word is a sequence of syllables presented in fixed order and repeated in different sentences. A spoken word thus intrinsically involves a BTP between syllables. Although backward association does not appear not to be involved in language (Chartier & Dautriche, 2023), BTP may well be and more research on the topic is needed.

Table 1

Studies testing backward transitional probabilities in natural language and artificial language. Études testant les probabilités transitionnelles rétrospectives dans le langage naturel et le langage artificiel.

Study	**Sample**	**Material**	**Task**	**Measure**
Perruchet & Desaulty, 2008	French adults	Artificial words	2AFC	Offline
Pelucchi et al., 2009a	English infants	Natural language (Italian)	Head-turn preference procedure	Offline
Hay et al.,2011	English infants	Natural language (Italian)	Looking time	Offline
McDonald & Shillcock, 2003	English adults	Natural language (English)	Reading	Online
Demberg & Keller, 2008	English portion of the Dundee Corpus (50,000 words read by 10 English participants, see Kennedy and Pynte, 2005)	Natural language (English)	Reading	Online
Onnis et al., 2022	English adults	Natural language (English)	Reading	Online
Onnis & Thiessen, 2013	English and Korean students	- Artificial words - Visual sequences (shapes) - Nonlinguistic sequences (tones)	2AFC	Offline
Thiessen et al., 2019	English and Korean infants aged 7 months and 13 months	Artificial words	Head-turn preference procedure	Offline

Conflict of interest

The authors declare no conflict of interest.

References

Albrengues, C., Lavigne, F., Aguilar, C., Castet, E., & Vitu, F. (2019). Linguistic processes do not beat visuo-motor constraints, but they modulate where the eyes move regardless of word boundaries : Evidence against top-down word-based eye-movement control during reading. *PloS One*, *14*(7). https://doi.org/10.1371/journal.pone.0219666

Apfelbaum, K. S., Hazeltine, E., & McMurray, B. (2013). Statistical learning in reading : Variability in irrelevant letters helps children learn phonics skills. *Developmental Psychology*, *49*(7), 1348-1365. https://doi.org/10.1037/a0029839

Arcediano, F., Escobar, M., & Miller, R. R. (2003). Temporal integration and temporal backward associations in human and nonhuman subjects. *Animal Learning & Behavior*, *31*, 242-256. https://doi.org/10.3758/BF03195986

Arciuli, J., & Monaghan, P. (2009). Probabilistic cues to grammatical category in English orthography and their influence during reading. *Scientific Studies of Reading*, *13*(1), 73-93. https://doi.org/10.1080/10888430802633508

Arciuli, J., & Simpson, I. C. (2012). Statistical learning is related to reading ability in children and adults. *Cognitive Science*, *36*(2), 286-304. https://doi.org/10.1111/j.1551-6709.2011.01200.x

Aslin, R. N., Saffran, J. R., & Newport, E. L. (1998). Computation of conditional probability statistics by 8-month-old infants. *Psychological Science*, *9*(4), 321-324. https://doi.org/10.1111/1467-9280.00063

Bahl, L. R., Jelinek, F., & Mercer, R. L. (1983). A maximum likelihood approach to continuous speech recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *2*, 179-190. https://doi.org/10.1109/TPAMI.1983.4767370

Batterink, L. J. (2017). Rapid statistical learning dupporting word extraction from continuous speech. *Psychological Science*, *28*(7), 921-928. https://doi.org/10.1177/0956797617698226

Bertels, J., Franco, A., & Destrebecqz, A. (2012). How implicit is visual statistical learning? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(5), 1425. https://doi.org/10.1037/a0027210

Bonatti, L. L., Pena, M., Nespor, M., & Mehler, J. (2005). Linguistic constraints on statistical computations : The role of consonants and vowels in continuous speech processing. *Psychological Science*, *16*(6), 451-459. https://doi.org/10.1111/j.0956-7976.2005.01556.x

Brady, T. F., Konkle, T., & Alvarez, G. A. (2009). Compression in visual working memory : Using statistical regularities to form more efficient memory representations. *Journal of Experimental Psychology: General*, *138*(4), 487. https://doi.org/10.1037/a0016797

Brunel, N., & Lavigne, F. (2009). Semantic priming in a cortical network model. *Journal of Cognitive Neuroscience*, *21*(12), 2300-2319. https://doi.org/10.1162/jocn.2008.21156

Brunellière, A., Vincent, M., & Delrue, L. (2022). Oscillatory correlates of linguistic prediction and modality effects during listening to auditory-only and audiovisual sentences. *International Journal of Psychophysiology*, *178*, 9-21. https://doi.org/10.1016/j.ijpsycho.2022.06.003

Bulgarelli, F., Benitez, V., Saffran, J., Byers-Heinlein, K., & Weiss, D. J. (2017). Statistical learning of multiple structures by 8-month-old infants. *Proceedings of the... Annual Boston University Conference on Language Development. Boston University*

Conference on Language Development, *41*, 128. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5576994/

Bulgarelli, F., & Weiss, D. J. (2016). Anchors aweigh : The impact of overlearning on entrenchment effects in statistical learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *42*(10), 1621. https://doi.org/10.1037/xlm0000263

Chartier, T. F., & Dautriche, I. (2023). Do backward associations have anything to say about language? *Cognitive Science*, *47*(4), e13282. https://doi.org/10.1111/cogs.13282

Chartier, T. F., & Fagot, J. (2022). Simultaneous learning of directional and non-directional stimulus relations in baboons (Papio papio). *Learning & Behavior*, 1-13. https://doi.org/10.3758/s13420-022-00522-8

Chwilla, D. J., Kolk, H. H., & Mulder, G. (2000). Mediated priming in the lexical decision task : Evidence from event-related potentials and reaction time. *Journal of Memory and Language*, *42*(3), 314-341. https://doi.org/10.1006/jmla.1999.2680

Cleeremans, A., & McClelland, J. L. (1991). Learning the structure of event sequences. *Journal of Experimental Psychology: General*, *120*(3), 235. https://doi.org/10.1037/0096-3445.120.3.235

Coney, J. (2002). The effect of associative strength on priming in the cerebral hemispheres. *Brain and Cognition*, *50*(2), 234-241. https://doi.org/10.1016/S0278-2626(02)00507-9

Conway, C. M., Bauernschmidt, A., Huang, S. S., & Pisoni, D. B. (2010). Implicit statistical learning in language processing : Word predictability is the key. *Cognition*, *114*(3), 356-371. https://doi.org/10.1016/j.cognition.2009.10.009

Conway, C. M., & Christiansen, M. H. (2005). Modality-constrained statistical learning of tactile, visual, and auditory sequences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(1), 24. https://doi.org/10.1037/0278-7393.31.1.24

DeLong, K. A., Urbach, T. P., & Kutas, M. (2005). Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nature Neuroscience*, *8*(8), 1117-1121. https://doi.org/10.1038/nn1504

Demberg, V., & Keller, F. (2008). Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, *109*(2), 193-210. https://doi.org/10.1016/j.cognition.2008.07.008

Deroost, N., Zeischka, P., Coomans, D., Bouazza, S., Depessemier, P., & Soetens, E. (2010). Intact first-and second-order implicit sequence learning in secondary-school-aged children with developmental dyslexia. *Journal of Clinical and Experimental Neuropsychology*, *32*(6), 561-572. https://doi.org/10.1080/13803390903313556

Destrebecqz, A., Vande Velde, M., San Anton, E., Cleeremans, A., & Bertels, J. (2019). Saving the Perruchet effect : A role for the strength of the association in associative learning. *Quarterly Journal of Experimental Psychology*, *72*(6), 1379-1386. https://doi.org/10.1177/1747021818791079

Endress, A. D., & Johnson, S. P. (2021). When forgetting fosters learning : A neural network model for statistical learning. *Cognition*, *213*, 104621. https://doi.org/10.1016/j.cognition.2021.104621

Erickson, C. A., & Desimone, R. (1999). Responses of macaque perirhinal neurons during and after visual stimulus association learning. *Journal of Neuroscience*, *19*(23), 10404-10416. https://doi.org/10.1523/JNEUROSCI.19-23-10404.1999

French, R. M., Addyman, C., & Mareschal, D. (2011). TRACX : A recognition-based connectionist framework for sequence segmentation and chunk extraction. *Psychological Review*, *118*(4), 614. https://doi.org/10.1037/a0025255

Frisson, S., Rayner, K., & Pickering, M. J. (2005). Effects of contextual predictability and transitional probability on eye movements during reading. *Journal of Experimental*

Psychology: Learning, Memory, and Cognition, *31*(5), 862. https://doi.org/10.1037/0278-7393.31.5.862

Gavard, E., & Ziegler, J. C. (2022). The effects of semantic and syntactic prediction on reading aloud. *Experimental Psychology*, *69*(6), 308-319. https://doi.org/10.1027/1618-3169/a000568

Gebhart, A. L., Aslin, R. N., & Newport, E. L. (2009). Changing structures in midstream : Learning along the statistical garden path. *Cognitive Science*, *33*(6), 1087-1116. https://doi.org/10.1111/j.1551-6709.2009.01041.x

Gomez, R. L. (1997). Transfer and complexity in artificial grammar learning. *Cognitive Psychology*, *33*(2), 154-207. https://doi.org/10.1006/cogp.1997.0654

Gomez, R. L., & Gerken, L. (1999). Artificial grammar learning by 1-year-olds leads to specific and abstract knowledge. *Cognition*, *70*(2), 109-135. https://doi.org/10.1016/S0010-0277(99)00003-7

Goodman, G. O., McClelland, J. L., & Gibbs., R. W. (1981). The role of syntactic context in word recognition. *Memory & Cognition*, *9*(6), 580-586. https://doi.org/10.3758/BF03202352

Graf Estes, K. (2012). Infants generalize representations of statistically segmented words. *Frontiers in Psychology*, *3*, 447. https://doi.org/10.3389/fpsyg.2012.00447

Hagoort, P., Hald, L., Bastiaansen, M., & Petersson, K. M. (2004). Integration of word meaning and world knowledge in language comprehension. *Science*, *304*(5669), 438-441. https://doi.org/10.1126/science.1095455

Hao Wang, F., Luo, M., & Wang, S. (2023). Statistical word segmentation succeeds given the minimal amount of exposure. *Psychonomic Bulletin & Review*. https://doi.org/10.3758/s13423-023-02386-z

Hay, J. F., Pelucchi, B., Estes, K. G., & Saffran, J. R. (2011). Linking sounds to meanings : Infant statistical learning in a natural language. *Cognitive Psychology*, *63*(2), 93-106. https://doi.org/10.1016/j.cogpsych.2011.06.002

Hebb, D. O. (1949). The first stage of perception : Growth of the assembly. *The Organization of Behavior*, *4*(60), 78-60.

Hebb, D. O. (1961). Distinctive features of learning in the higher animal. *Brain Mechanisms and Learning*, *37*, 46.

Howard Jr, J. H., & Howard, D. V. (1997). Age differences in implicit learning of higher order dependencies in serial patterns. *Psychology and Aging*, *12*(4), 634. https://doi.org/10.1037/0882-7974.12.4.634

Huettig, F., & Mani, N. (2016). Is prediction necessary to understand language? Probably not. *Language, Cognition and Neuroscience*, *31*(1), 19-31. https://doi.org/10.1080/23273798.2015.1072223

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. https://doi.org/10.1037/0096-3445.130.4.658

Isbilen, E. S., & Christiansen, M. H. (2022). Statistical learning of language : A meta‐analysis into 25 years of research. *Cognitive Science*, *46*(9), e13198. https://doi.org/10.1111/cogs.13198

Jones, J., & Pashler, H. (2007). Is the mind inherently forward looking? Comparing prediction and retrodiction. *Psychonomic Bulletin & Review*, *14*(2), 295-300. https://doi.org/10.3758/BF03194067

Jusczyk, P. W., & Aslin, R. N. (1995). Infants′ detection of the sound patterns of words in fluent speech. *Cognitive Psychology*, *29*(1), 1-23. https://doi.org/10.1006/cogp.1995.1010

Karuza, E. A., Li, P., Weiss, D. J., Bulgarelli, F., Zinszer, B. D., & Aslin, R. N. (2016). Sampling over nonuniform distributions : A neural efficiency account of the primacy effect in statistical learning. *Journal of Cognitive Neuroscience*, *28*(10), 1484-1500. https://doi.org/10.1162/jocn_a_00990

Kennedy, A., & Pynte, J. (2005). Parafoveal-on-foveal effects in normal reading. *Vision Research*, *45*(2), 153-168. https://doi.org/10.1016/j.visres.2004.07.037

Kóbor, A., Janacsek, K., Takács, Á., & Nemeth, D. (2017). Statistical learning leads to persistent memory : Evidence for one-year consolidation. *Scientific Reports*, *7*(1), 760. https://doi.org/10.1038/s41598-017-00807-3

Lavigne, F., Chanquoy, L., Dumercy, L., & Vitu, F. (2013). Early dynamics of the semantic priming shift. *Advances in Cognitive Psychology*, *9*(1), 1.

Lavigne, F., Dumercy, L., Chanquoy, L., Mercier, B., & Vitu-Thibault, F. (2012). Dynamics of the semantic priming shift : Behavioral experiments and cortical network model. *Cognitive Neurodynamics*, *6*(6), 467-483. https://doi.org/10.1007/s11571-012-9206-0

Lavigne, F., Dumercy, L., & Darmon, N. (2011). Determinants of multiple semantic priming : A meta-analysis and spike frequency adaptive model of a cortical network. *Journal of Cognitive Neuroscience*, *23*(6), 1447-1474. https://doi.org/10.1162/jocn.2010.21504

Lazartigues, L., Mathy, F., & Lavigne, F. (2021). Statistical learning of unbalanced exclusive-or temporal sequences in humans. *Plos One*, *16*(2). https://doi.org/10.1371/journal.pone.0246826

Lazartigues, L., Mathy, F., & Lavigne, F. (2023). Probability, dependency, and frequency are not all equally involved in statistical learning. *Experimental Psychology*. https://doi.org/10.1027/1618-3169/a000561

Luka, B. J., & Van Petten, C. (2014). Prospective and retrospective semantic processing : Prediction, time, and relationship strength in event-related potentials. *Brain and Language*, *135*, 115-129. https://doi.org/10.1016/j.bandl.2014.06.001

Mantegna, F., Hintz, F., Ostarek, M., Alday, P. M., & Huettig, F. (2019). Distinguishing integration and prediction accounts of ERP N400 modulations in language processing through experimental design. *Neuropsychologia*, *134*, 107199. https://doi.org/10.1016/j.neuropsychologia.2019.107199

McCauley, S. M., & Christiansen, M. H. (2019). Language learning as language use : A cross-linguistic model of child language development. *Psychological Review*, *126*(1), 1-51. https://doi.org/10.1037/rev0000126

McDonald, S. A., & Shillcock, R. C. (2003). Low-level predictive inference in reading : The influence of transitional probabilities on eye movements. *Vision Research*, *43*(16), 1735-1751. https://doi.org/10.1016/S0042-6989(03)00237-2

Meyer, D. E., & Schvaneveldt, R. W. (1971). Facilitation in recognizing pairs of words : Evidence of a dependence between retrieval operations. *Journal of Experimental Psychology*, *90*(2), 227-234. https://doi.org/10.1037/h0031564

Minier, L., Fagot, J., & Rey, A. (2016). The temporal dynamics of regularity extraction in non-human primates. *Cognitive Science*, *40*(4), 1019-1030. https://doi.org/10.1111/cogs.12279

Neely, J. H. (1991). *Semantic priming effects in visual word recognition : A selectivereview of current findings and theory. in d. besner & gw humphreys (eds.), Basic processes in reading: Visual word recognition (pp. 264-336)*. Hillsdale, NJ: Erlbaum.

Nieuwland, M. S., Barr, D. J., Bartolozzi, F., Busch-Moreno, S., Darley, E., Donaldson, D. I., Ferguson, H. J., Fu, X., Heyselaar, E., Huettig, F., Matthew Husband, E., Ito, A., Kazanina, N., Kogan, V., Kohút, Z., Kulakova, E., Mézière, D., Politzer-Ahles, S.,

Rousselet, G., … Von Grebmer Zu Wolfsthurn, S. (2020). Dissociable effects of prediction and integration during language comprehension : Evidence from a large-scale study using brain potentials. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *375*(1791), 20180522. https://doi.org/10.1098/rstb.2018.0522

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

Onnis, L., Lim, A., Cheung, S., & Huettig, F. (2022). Is the mind inherently predicting? Exploring forward and backward looking in language processing. *Cognitive Science*, *46*(10), e13201. https://doi.org/10.1111/cogs.13201

Onnis, L., & Thiessen, E. (2013). Language experience changes subsequent learning. *Cognition*, *126*(2), 268-284. https://doi.org/10.1016/j.cognition.2012.10.008

Östling, R. (2015). Word order typology through multilingual word alignment. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, 205-211. https://aclanthology.org/P15-2034.pdf

Pelucchi, B., Hay, J. F., & Saffran, J. R. (2009a). 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

Pelucchi, B., Hay, J. F., & Saffran, J. R. (2009b). Statistical learning in a natural language by 8-month-old infants. *Child Development*, *80*(3), 674-685. https://doi.org/10.1111/j.1467-8624.2009.01290.x

Pena, M., Werker, J. F., & Dehaene-Lambertz, G. (2012). Earlier speech exposure does not accelerate speech acquisition. *Journal of Neuroscience*, *32*(33), 11159-11163. https://doi.org/10.1523/JNEUROSCI.6516-11.2012

Perruchet, P. (1985). A pitfall for the expectancy theory of human eyelid conditioning. *The Pavlovian Journal of Biological Science*, *20*(4), 163-170. https://doi.org/10.1007/BF03003653

Perruchet, P. (2015). Dissociating conscious expectancies from automatic link formation in associative learning : A review on the so-called Perruchet effect. *Journal of Experimental Psychology: Animal Learning and Cognition*, *41*(2), 105. https://doi.org/10.1037/xan0000060

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

Pickering, M. J., & Gambi, C. (2018). Predicting while comprehending language : A theory and review. *Psychological Bulletin*, *144*(10), 1002-1044. https://doi.org/10.1037/bul0000158

Reed, J., & Johnson, P. (1994). Assessing implicit learning with indirect tests : Determining what is learned about sequence structure. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(3), 585. https://doi.org/10.1037/0278-7393.20.3.585

Ren, J., Wang, M., & Arciuli, J. (2023). A meta-analysis on the correlations between statistical learning, language, and reading outcomes. *Developmental Psychology*, *59*(9), 1626. https://doi.org/10.1037/dev0001577

Rey, A., Fagot, J., Mathy, F., Lazartigues, L., Tosatto, L., Bonafos, G., Freyermuth, J., & Lavigne, F. (2022). Learning higher-order transitional probabilities in nonhuman primates. *Cognitive Science*, *46*(4). https://doi.org/10.1111/cogs.13121

Romberg, A. R., & Saffran, J. R. (2010). Statistical learning and language acquisition. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*(6), 906-914. https://doi.org/10.1002/wcs.78

Saffran, J. R. (2001). The use of predictive dependencies in language learning. *Journal of Memory and Language*, *44*(4), 493-515. https://doi.org/10.1006/jmla.2000.2759

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

Saffran, J. R., Newport, E. L., & Aslin, R. N. (1996). Word segmentation : The role of distributional cues. *Journal of Memory and Language*, *35*(4), 606-621. https://doi.org/10.1006/jmla.1996.0032

Saffran, J. R., & Wilson, D. P. (2003). From syllables to syntax : Multilevel statistical learning by 12-month-old infants. *Infancy*, *4*(2), 273-284. https://doi.org/10.1207/S15327078IN0402_07

Schvaneveldt, R. W., & Gomez, R. L. (1998). Attention and probabilistic sequence learning. *Psychological Research*, *61*(3), 175-190. https://doi.org/10.1007/s004260050023

Siegelman, N., Bogaerts, L., Christiansen, M. H., & Frost, R. (2017). Towards a theory of individual differences in statistical learning. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*(1711). https://doi.org/10.1098/rstb.2016.0059

Siegelman, N., Bogaerts, L., Kronenfeld, O., & Frost, R. (2018). Redefining "learning" in statistical learning : What does an online measure reveal about the assimilation of visual regularities? *Cognitive Science*, *42*(S3), 692-727. https://doi.org/10.1111/cogs.12556

Soares Filho, P. S., Silva, Á. J., Velasco, S. M., Barros, R. S., & Tomanari, G. Y. (2016). Assessing symmetry by comparing the acquisition of symmetric and nonsymmetric

conditional relations in a Capuchin Monkey. *International Journal of Psychological Research*, *9*(2), 30-39. https://doi.org/10.21500/20112084.2320

Thiessen, E. D., Onnis, L., Hong, S.-J., & Lee, K.-S. (2019). Early developing syntactic knowledge influences sequential statistical learning in infancy. *Journal of Experimental Child Psychology*, *177*, 211-221. https://doi.org/10.1016/j.jecp.2018.04.009

Thompson, S. P., & Newport, E. L. (2007). Statistical learning of syntax : The role of transitional probability. *Language Learning and Development*, *3*(1), 1-42. https://doi.org/10.1080/15475440709336999

Tillmann, B., & McAdams, S. (2004). Implicit learning of musical timbre sequences : Statistical regularities confronted with acoustical (dis) similarities. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(5), 1131. https://doi.org/10.1037/0278-7393.30.5.1131

Tovar, Á. E., & Westermann, G. (2023). No need to forget, just keep the balance : Hebbian neural networks for statistical learning. *Cognition*, *230*. https://doi.org/10.1016/j.cognition.2022.105176

Treiman, R., & Kessler, B. (2006). Spelling as statistical learning : Using consonantal context to spell vowels. *Journal of Educational Psychology*, *98*(3), 642. https://doi.org/10.1037/0022-0663.98.3.642

Tremblay, A., & Tucker, B. V. (2011). The effects of N-gram probabilistic measures on the recognition and production of four-word sequences. *The Mental Lexicon*, *6*(2), 302-324. https://doi.org/10.1075/ml.6.2.04tre

Weiss, D. J., Gerfen, C., & Mitchel, A. D. (2009). Speech segmentation in a simulated bilingual environment : A challenge for statistical learning? *Language Learning and Development*, *5*(1), 30-49. https://doi.org/10.1080/15475440802340101

Wilkinson, G. S., & Robertson, G. J. (1993). Wide range achievement test 4. *Journal of Clinical and Experimental Neuropsychology*. https://doi.org/10.1037/t27160-000

Willems, R. M., Frank, S. L., Nijhof, A. D., Hagoort, P., & Van den Bosch, A. (2016). Prediction during natural language comprehension. *Cerebral Cortex*, *26*(6), 2506-2516. https://doi.org/10.1093/cercor/bhv075

Zang, C., Fu, Y., Du, H., Bai, X., Yan, G., & Liversedge, S. P. (2023). Processing multiconstituent units : Preview effects during reading of Chinese words, idioms, and phrases. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. https://doi.org/10.1037/xlm0001234