Martin Zettersten

Learning by predicting: How predictive processing informs language development

Abstract: An increasingly influential attempt to provide a unified theory of the mind is grounded in the notion of prediction. On this account, our minds are prediction engines, continuously matching incoming input to top-down expectations. Higher-level predictions or expectations are generated by internal cognitive models at multiple hierarchical levels that jointly serve to minimize prediction error at lower levels in the information processing hierarchy. In language research, prediction has become an increasingly influential approach to understanding how language comprehension unfolds in real time. But how can predictive processing inform our understanding of how we come to learn language in the first place? In this review, I consider how prediction-based theories of the mind can aid in explaining how language development unfolds. First, I review research in perception and language on predictive processes and assess the degree to which they are found in infancy. Next, I consider how prediction-based mechanisms contribute to our understanding of learning, as well as the kinds of patterns that models grounded in prediction can learn. I review research on infants' prodigious ability to track novel patterns and relate these statistical learning abilities to prediction-based explanations. Finally, I sketch how prediction-based accounts fit within current theoretical positions and debates in the field of language development and suggest directions for future research into how predictive processes support language learning.

1 Introduction

At the heart of cognitive development lie two fundamental mysteries:[1] What is the nature of the infant mind, and how does an infant mind develop into an adult mind? The answers to these questions have historically diverged radically, with

1 This work was supported by NSF-GRFP DGE-1256259 to MZ. I am grateful to Jenny Saffran, Gary Lupyan, Viridiana Benitez, and the participants in the 2017 Kavli Summer Institute in Cognitive Neuroscience for helpful discussion.

Martin Zettersten, University of Wisconsin-Madison, Department of Psychology, Madison, WI 53706, USA, zettersten@wisc.edu

https://doi.org/10.1515/9783110596656-010

some suggesting the infant mind initially encounters the world as a "blooming, buzzing confusion" (James 1890: 488), while others suggest that the infant mind is from the beginning endowed with rich, adult-like cognitive structure and knowledge (Chomsky 1959, 1980). Modern approaches to cognitive development in general and language development in particular explore different solutions that lie somewhere between these two extremes, either by positing strong continuities in knowledge and ability between the infant and the adult mind (Spelke and Kinzler 2007; Baillargeon and Carey 2012; Dehaene-Lambertz and Spelke 2015) or by exploring how adult-like cognitive structure and knowledge can emerge despite apparently humble cognitive beginnings (Elman et al. 1996; McClelland et al. 2010; Smith and Thelen 2003). Yet the underlying questions are still fundamentally unresolved.

What makes the mystery of early cognition and development so difficult is that the most basic question in psychology itself remains elusive: how does the mind work? What are the general organizing principles underlying how we learn about and engage with the world? One increasingly influential proposal is grounded in the notion of prediction (Clark 2013; Hohwy 2014; Friston 2010; Bar 2009). In these accounts, the brain is conceptualized as "proactive", in that it "continuously generates predictions that anticipate the relevant future" (Bar 2009: 1235). On this view, our minds are essentially prediction engines, continuously deploying top-down expectations to anticipate what will occur next and reduce errors that occur when these expectations do not match incoming input.

The view of the mind as a prediction engine has particularly gained traction in the study of language processing (Kuperberg and Jaeger 2016; Huettig 2015; Pickering and Garrod 2007; Pickering and Clark 2014; Kutas, Federmaier and Urbach 2014). Prediction-based accounts of cognitive processing are intuitively appealing in the domain of language, because they make sense of what otherwise appears to be an almost impossible task: as each sentence unfolds, a language comprehender must parse a continuous stream of incoming fluctuations in sound into a coherent collection of phonemes, syllables, words, and sentences while decoding their meaning within mere fractions of a second. The only hope for the hearer would seem to involve bringing to bear a strong set of expectations about incoming linguistic elements in order to arrive at the speaker's intended message quickly.

Given its merit as key explanatory principle in the functioning of the mind, can prediction theory help us solve the fundamental questions in language development? In the following article, I explore what prediction can tell us about how infants develop and learn language. In the first part, I review evidence from sensory and language processing for prediction both in adults and in infants. Next, I investigate the relationship between prediction and learning, particularly

how they relate to infants' powerful pattern-learning abilities. In the final section, I consider where prediction-based approaches fit within classic theoretical debates in language development.

2 Minds as the products of predicting brains: The (abridged) case for prediction

2.1 What is prediction?

Predictive processing accounts seek to explain the key mechanisms governing how the brain works. There are many different families of predictive processing accounts (Friston 2010; Rao and Ballard 1999; Hawkins and Blakeslee 2004; O'Reilly, Wyatte and Rohrlich 2014; Clark 2013; Hohwy 2014; Bar 2009; Kuperberg and Jaeger 2016), but they all share the basic idea that prediction is the key principle to how the brain – as well as the mind – functions. The goal of the brain is to predict incoming input – what will occur in the very next moment. In order to make these predictions, the brain develops a hierarchical model, with each level attempting to predict the input it receives from the level below. The overarching goal of the system is to reduce prediction error as best as possible. Operating under these constraints, the brain builds ever richer and more precise models of its environment, all in service of making efficient and accurate predictions.

Prediction-based accounts come in many flavors. One particularly influential account introduces the principle of predictive coding (Rao and Ballard 1999; Friston 2010; Clark 2013). What gives this account its name is how it reconceptualizes the nature of neural signals. Rather than representing information about the currently processed stimulus, neural signals encode prediction errors: discrepancies between predicted and actual input. These prediction errors then feed forward, becoming the input to the next level of cortical hierarchy. This level in turn attempts to predict incoming signals and passes error signals to the next level of the hierarchy, and so on. In this sense, neural responses are "signaling the news" (Clark 2015: 18), passing along unexplained or unpredicted information in the incoming signal.

In this paper, I will not seek to weigh different prediction accounts against one another, but instead use basic ideas shared among all of them to illuminate language development. I will focus in particular on how prediction-based accounts have begun to be applied to language processing and learning. There are two major ways in which language researchers have used the term prediction, in a broad sense and a narrower sense. In the broad sense, prediction refers to the

idea that the mind is constantly engaged in a process of probabilistic inference. As Kuperberg and Jaeger (2016) put it in an influential review:

> ... prediction implies that, at any given time, we use high-level information within our representation of context to probabilistically infer upcoming information at this same higher-level representation. (Kuperberg and Jaeger 2016: 47)

On this view, the mind is continuously engaged in a cycle of updating its beliefs and expectations at multiple levels of its hierarchical representation of incoming language input.

A related, but somewhat more narrow sense focuses on the timing of activations in cognitive processing. Kutas, Federmaier and Urbach (2014) define prediction as encompassing any form of cognitive processing "involv[ing] the activation of or information about likely upcoming stimuli, prior to their receipt, that plays a causal role in stimulus processing" (Kutas, Federmaier and Urbach 2014: 649). Note that predictions, on this definition, can take many forms: they can be consciously or unconsciously generated, they may be explicit or implicit, they can be more fine-grained or more coarse-grained, and they can be generated at multiple levels. For instance, when processing a sentence such as *I love* ..., prediction may involve an expectation[2] for the specific word that will come next (*babies*, for instance), for a particular meaning to be expressed ('something cute or lovable'), for a particular grammatical structure (e.g. a noun phrase), for a particular phonetic feature (e.g. that the next word will begin with a voiced consonant), and so on.

One of the recurring difficulties in interpreting the psychological literature is determining what "counts" as evidence for prediction, in particular disentangling prediction effects from effects of integration or facilitation (Kutas, Federmaier and Urbach 2014; Kuperberg and Jaeger 2016). In language processing, for instance, if participants respond faster in a serial reading task to a more predictable word as compared to a less predictable word, this finding can often be explained in two ways: it could be that the previous linguistic context is allowing participants to begin activating information relevant to the word before encountering it (pre-activation) or it could be that the word, once encountered, is more easily integrated with the previous linguistic context (integration). In the broad

2 While it may not always be practical to distinguish between predictive behaviour and the cognitive construct of prediction, I will generally use the terms "expectation" or "expectancy" to refer to the cognitive processing that generates a prediction, and "anticipation" to refer to behavioural responses that reflect these predictions (Haith, Hazan and Goodman 1988; Canfield et al. 1997).

sense of prediction in terms of probabilistic inference, this distinction collapses somewhat, since probabilistic expectations should lead to both pre-activating and integrating cognitive processes further down the hierarchical processing stream. In the following sections, I will review evidence that supports an explanation in terms of predictive processing, both in the broad sense of probabilistic inference and in the narrow sense of pre-activation, but this evidence will also be interwoven with a broader psychological literature, some of which could be interpreted as effects of pre-activation or as effects of integration. In all cases, I hope to show that adopting the lens of prediction leads to a fruitful interpretation of a wide variety of experimental evidence.

What makes predictive processing accounts of the brain so powerful is their ability to unify a vast amount of what we know about behaviour across a variety of domains, including classically perceptual processes such as vision and the higher-level cognitive processes involved in language comprehension. Why does the brain respond to the *absence* of expected stimuli? Why are we subject to "garden-path" effects in language processing? Why are infants drawn to regularities, and why do they seem to automatically detect patterns in their environment? In the following sections, I briefly review how prediction unifies these disparate phenomena, to provide an intuition as to why prediction is a useful unifying framework for understanding cognition. While by no means a comprehensive treatment of predictive processing accounts, the goal is to offer a glimpse of the explanatory breadth and depth of this family of accounts and to motivate why prediction is an attractive lens through which to consider the development of the mind.

2.2 Processing expected and unexpected sensory input

Some of the most compelling evidence that the brain is consistently developing expectations about what it will encounter in the world comes from studying what happens when predictions go wrong. A vast number of studies have studied cortical responses to unexpected events (e.g. den Ouden et al. 2009; Bendixen et al. 2014; Chennu et al. 2013; Wacongne et al. 2011). Of particular interest are cases in which an expected stimulus does not appear. A purely bottom-up account of cortical processing predicts that early sensory areas should show little or no activation when an expected stimulus is absent, since the sensory system is not receiving any bottom-up input from the world. Instead, studies that investigate neural response patterns in sensory cortex to withheld stimuli find very different results: early sensory areas show strong activation in the absence of bottom-up input when a sensory stimulus is expected. For instance, when processing temporal auditory sequences, sensory cortices show strong activation when expected

items in the sequence are omitted, i.e. even in the absence of a stimulus (Wacongne et al. 2011; Wacongne, Changeux and Dehaene 2012). This has led researchers to reinterpret early cortical responses associated with unexpected events as signatures of prediction. For instance, Wacongne et al. (2012) offer a model of the mismatch negativity (MMN) – an event-related potential that registers roughly 100–200 ms after an infrequent unexpected auditory event[3] – as indexing a prediction violation, which explains why this characteristic signature is found in response to the omission of expected input.

A converse result is that cortical activity can disappear even in the presence of a stimulus, provided that the stimulus is highly predictable. For instance, in functional magnetic resonance imaging (fMRI) studies, cortical activity elicited by a stimulus is increasingly reduced each time that stimulus is repeated, a phenomenon known as repetition suppression (Grill-Spector, Henson and Martin 2006). Recent studies suggest that the reduction in cortical activity results from the stimulus becoming progressively more precisely predicted (Todorovic and de Lange 2012; Todorovic et al. 2011; Andics et al. 2013; Summerfield et al. 2011; Summerfield et al. 2008). The key finding is that repetition suppression is modulated by how predictable the repetition is: when a repetition is more frequent or predictable, cortical activity is suppressed more strongly, presumably reflecting more accurate and precise predictions (e.g. Summerfield et al. 2008).

Further evidence that the cortex is generating active predictions about incoming perceptual input comes from studies that show that participants' expectations bias early visual representations (Kok et al. 2013; Kok, Failing and de Lange 2014; Kok, Jehee and de Lange 2012). For instance, in one study (Kok et al. 2013), participants' expectations about the orientation of an upcoming visual stimulus was manipulated with an auditory cue played shortly before the visual input – different auditory cues systematically predicted specific orientations. The central result was that information about the orientation of the stimulus could be reconstructed from early visual areas prior to the actual onset of the visual stimulus, demonstrating that predictions about the upcoming visual input reshaped early visual representations. The picture emerging from these studies is that the

3 Event-related potentials (ERPs) are changes in electrical brain activity time-locked to a specific sensory or cognitive event that are measured through electroencephalography (EEG) – an electrophysiological method used in cognitive neuroscience to detect electrical activity in the brain using electrodes placed on the scalp. ERPs can be described and analysed as waveforms with peaks and troughs that are thought to index different cognitive processes. The MMN is a particular ERP component that is typically found 100–200 ms after the onset of an infrequent or surprising ("oddball") element in a sequence of stimuli (usually a visual or an auditory sequence).

perceptual system rapidly generates predictions – in the sense of pre-activation – about upcoming sensory input.

2.3 A garden of forking paths in language processing

Contextual effects pervade language (Kuperberg and Jaeger 2016). For instance, a classic finding in language comprehension is the so-called "garden-path" phenomenon. If an ambiguous phrase such as (1) is resolved into a less frequent syntactic parse such as (2), as opposed to a more salient syntactic interpretation such as (3), this leads to processing difficulty that manifests as slower reading times or worse comprehension (MacDonald, Just and Carpenter 1992; Ferreira and Clifton Jr. 1986; Ferreira and Patson 2007).

(1) *The researcher expected to finish the paper …*
(2) *… fell asleep.*
(3) *… by the end of the day.*

Another example is that people react faster to and spend less time processing predictable than unpredictable words across a number of paradigms (Stanovich and West 1979, McClelland and O'Regan 1981; Staub 2015).

A longstanding controversy in the field is whether these contextual effects are best understood as effects of prediction or as effects of integration (Kuperberg and Jaeger 2016; Kutas, Federmaier and Urbach 2014). Responses to garden-path sentences might be slower because the system must explain prediction error, or because the system must engage more cognitive resources to integrate the end of the sentence with the preceding linguistic context. However, recent studies have provided strong evidence that the language comprehension is consistently engaged in prediction.

First, relatively abstract linguistic expectations can modulate sensory processing in its earliest stages. For instance, expectations about the form of words belonging to different syntactic categories can affect visual processing at early stages when reading sentences (Dikker et al. 2010). Second, recent research has provided compelling evidence that words become pre-activated prior to their occurrence during language comprehension. A classic finding from electroencephalography (EEG) studies is that semantically unexpected words generate a characteristic neural response about 200–500 ms post word onset, the N400 (Kutas

and Hillyard 1980).[4] Interestingly, the amplitude of the N400 correlates strongly with how expected a word is based on the preceding context (Kutas and Hillyard 1984; Kutas and Federmaier 2011). In one of the strongest demonstrations that the N400 reflects prediction, rather than integration, DeLong et al. (2005) presented participants with sentences which generated expectations for specific nouns (e.g. *The day was breezy so the boy went outside to fly …*). Crucially, the form of the indefinite article preceding the noun (*a* or *an*) could be consistent or inconsistent with the expected noun (in this example *kite*), but both articles were equally easy to integrate with the preceding context. DeLong et al. found an N400 effect in response to the inconsistent articles (*an*), before encountering an unexpected noun (*airplane*), an effect that could only be found if participants were pre-activating the corresponding noun.[5] This study, along with many others using a similar design, show that – at least in some contexts – language comprehenders are generating expectations about upcoming words (Wicha et al. 2003; Wicha, Moreno and Kutas 2004; Van Berkum et al. 2005; Brothers, Swaab and Traxler 2015; Wicha, Moreno and Kutas 2003) and word classes (Szewczyk and Schriefers 2013). These results are key highlights within a converging literature suggesting that prediction – in the sense of generating expectations about upcoming language input – is integral to language processing (Kuperberg and Jaeger 2016).

3 Infants as predictors

There is a substantial amount of evidence that has accrued for the predictive processing account in adults. But how well does this account mesh with existing evidence in developmental research? Are babies' brains fruitfully construed as prediction engines? None other than Jean Piaget noted that "anticipatory function … is to be found over and over again, at every level of the cognitive mechanisms and

4 The N400 is a component of an event-related potential (ERP) – see also fn. 3 – first observed in response to semantically unexpected words. The name is derived from the fact that the component is associated with a negative deflection in the ERP waveform around 400 ms after the onset of an unexpected word/stimulus.

5 There is currently some controversy surrounding the specific anticipatory results from Delong et al. (2005) following recent failures to replicate this result (Ito, Martin, and Nieuwland 2017; Nieuwland et al. 2018) and subsequent rebuttals from the original authors (DeLong, Urbach, and Kutas 2017a, 2017b). Regardless of the final determination regarding this particular result about pre-activation on the phonological level of words, there is a broad literature supporting evidence for the pre-activation of words more generally across different contexts and languages (see the studies cited in the text and Kuperberg and Jaeger 2016 for a review).

at the very heart of the most elementary habits, even of perception" (Piaget 1971: 19). In the following section, I invite the reader to see the developmental literature through the lens of prediction. Prediction casts new light on a vast number of phenomena spanning all domains of cognitive developmental research, including perception and language, and even the very methods cognitive development researchers employ to understand the infant mind.

3.1 Looking to predict: Infant looking behaviour

Some of the most important and influential insights in the field of infant development stem from measurements of infants' gaze and looking preferences (Hespos and Spelke 2004; Wynn 1992; Baillargeon, Spelke and Wasserman 1985; Baillargeon and Carey 2012; Gergely et al. 1995; Fantz 1961; 1963). Yet there is still substantial debate in the field as to what various behavioural measurements reflect in terms of infants' processing, leaving open the question of "what's in a look" (Aslin 2007). A particularly vexing question is why infants sometimes show novelty preferences, looking longer to events that are more surprising or less consistent with previous experience, but on other occasions show familiarity preferences, looking longer to events that are more expected or consistent with previous experience

The traditional view of infant looking times is that they are reactions to visual or auditory experience, that may be driven by exogenous factors (e.g. how salient a stimulus is) or endogenous factors (e.g. how robustly a stimulus is encoded in memory; Aslin 2014). More recently, infant looking behaviour has begun to be reconceptualized as a more active process (Kidd, Piantadosi and Aslin 2012; Kidd and Hayden 2015). On this model, infants' looking behaviour may reflect an active attempt to sample information from the environment. This perspective is consistent with a predictive processing account, whereby infants' looking behaviour should reflect a continuous process of collecting information about the visual environment to reduce uncertainty (Itti and Baldi 2009; Henderson 2017).

A key result in understanding infants' gaze behaviour as a more active process is the so-called *Goldilocks effect* (Kidd, Piantadosi and Aslin 2012, 2014). Both in the visual and in the auditory domain infants appear to prefer events that are "just right" in terms of their predictability: neither perfectly predictable nor completely predictable. For instance, in Kidd et al. (2012), infants viewed objects disappearing and reappearing behind a screen. By varying how predictable the pattern of reappearance of an object was from behind a particular screen, Kidd et al. obtained a measure of a given event's predictability or complexity. For example, if an object can appear from behind one of two screens, an extremely predictable

event is one in which an object appears from behind screen 1 after having appeared repeatedly from screen 1 on previous events (e.g. creating the sequence 1–1–1–**1**). On the other end of the continuum, if an object suddenly appears from behind screen 2 after having only appeared from behind screen 1 (i.e. the sequence 1–1–1–**2**), then the event is much more surprising. An event can also lie in between these two extremes, creating a pattern that has some variability, but is also somewhat predictable (e.g. 1–2–1–**2**). Crucially, the predictability or complexity of a particular event within a pattern influenced how long infants would continue to watch the event sequence. Infants showed a U-curve preference, with infants looking longest to patterns that were neither too predictable nor too unpredictable (i.e. events such as 1–2–1–2 in the example above). This U-shaped curve held for every individual infant, not just for the group of participants overall (Piantadosi, Kidd and Aslin 2014).

These results lend themselves to an account of infant looking behaviour based on prediction: if infants organize their gaze behaviour around minimizing prediction error, their looking behaviour will depend on how successfully they can reduce prediction error for a given visual event. If an event is highly predictable, the visual system will rapidly learn to predict upcoming events and will move on from the current event sequence to make predictions about other aspects of the environment. If, on the other hand, the event is too unpredictable, the system will quickly plateau in its ability to reduce prediction error and therefore seek out other events where prediction error can be reduced more efficiently. When stimuli lie between these two extremes, they will hold infant gaze longer to the extent to which longer looking continues to reduce prediction error. Sequential patterns that will hold gaze the longest are those that lie at the "sweet-spot" of informativeness, where continuing to look improves infants' predictions regarding the task currently in focus (e.g. in the case of the Kidd et al. task, predicting where an object will appear next in a sequence).

This view of infant looking behaviour offers a principled way to predict when infants will show novelty or familiarity preferences. Looking preferences will ultimately depend on the relative effectiveness with which prediction error can be reduced for novel and familiar stimuli. This is consistent with the fact that infants often show novelty preferences in studies with lengthy habituation phases, e.g. in statistical learning studies (Saffran, Aslin and Newport 1996; Aslin 2014): infants in these studies have minimized prediction error to the familiar stimulus and thus spend more time looking at the novel stimulus to reduce prediction error. It also explains why infants often show familiarity preferences in studies in which infants listen to their native language without an extended habituation phase (Jusczyk and Aslin 1995). Fluent speech provides ample opportunity for an

infant's processing system to attempt to reduce prediction error. Similarly, this explains why infants show a preference for speech rather than a variety of non-speech stimuli (Vouloumanos et al. 2010; Vouloumanos and Werker 2004) and for their native language rather than a non-native language (Moon, Cooper and Fifer 1993). Infants' predictive models of their auditory environment in these cases are more effective in reducing prediction error for speech, and particularly for their native tongue, which even in the absence of a habituation phase is a rich source of prediction error that can be productively reduced.

On a predictive processing account, looking times are more than simply measures of a learning outcome or an infant's ability to discriminate two different stimuli. Longer looking times reflect an active process of predicting upcoming stimuli and integrating information about the outcome of these predictions into a dynamically updated model of the world. This view of looking times brings into focus that these looking events are themselves learning events.

3.2 Vision and multimodal sensory processing

From a young age, infants rapidly build expectations and anticipate what will occur in their environment. When viewing a video in which engaging stimuli occur in one of two possible locations (either on the left or the right side of a screen), infants as young as 2–3 months of age begin to anticipate the onset of an upcoming visual stimulus, as measured by fixation shifts to the likely location of the stimulus that begin before an eye movement could be programmed in reaction to the onset of the stimulus (Canfield and Haith 1991; Haith, Hazan and Goodman 1988; Canfield et al. 1997). The extent to which infants show anticipatory shifts depends on the predictability of the sequence: by 3 months, infants will show more anticipatory shifts when a sequence is regular (e.g. when the visual stimulus alternates between two locations) than when it is irregular (e.g. when the next stimulus location cannot be predicted from the previous two or three events in the sequence; Canfield and Haith 1991). By 12 months of age, infants show regular anticipatory looks even to more probabilistic event sequences, and their anticipatory fixations become increasingly accurate (Romberg and Saffran 2013).

The fact that infants will reliably attempt to predict upcoming visual events is exploited by various research paradigms that measure infants' learning and knowledge in terms of anticipatory behaviour. In anticipatory eye movement paradigms, researchers expose infants to associations between a cue (e.g. an auditory cue such as a word) and a reinforcing event occurring in a particular location (e.g. a circle appearing on the left or on the right side of the screen). By 6 months, infants will regularly anticipate the reinforcing event's location on perceiving the

cue. Researchers can then infer that infants distinguish two cues (e.g. the different words) if they differentially predict where the reinforcing stimulus will appear on the screen based on the specific cue presented. This paradigm has used infants' anticipatory behaviour to demonstrate the types of auditory categories 6-month-olds form (McMurray and Aslin 2004) or to investigate how 7-month-olds rapidly and flexibly learn to distinguish speech patterns (Kovács and Mehler 2009a, 2009b). Though not always considered in the context of predictive processing accounts of the mind, these studies reveal that infants form expectations about where visually interesting events will occur after only brief exposure to predictive cues, and actively orient their attention in anticipation of visual events.

While these studies show that infants anticipate *where* perceptual events will occur, it leaves open the question of whether infants predict *what* they will see, i.e. the perceptual content of the events themselves. Recent work investigating the neural processing in cross-modal priming events provides compelling evidence that by 6 months of age, infants' perceptual processing reflects sensory expectations (Emberson, Richards and Aslin 2015; Kouider et al. 2015). In one study, Emberson and colleagues (2015) measured changes in blood oxygenation using functional near-infrared spectroscopy (fNIRS) while 6-month-old infants watched movies in which novel sounds and visual stimuli were paired. After establishing the mutual predictability of sound and visual stimuli, infants saw a series of trials either consistent with the induced sensory expectation (i.e. in which both auditory and visual stimuli appeared; about 80% of the trials) or inconsistent, such that the expected visual stimulus was omitted (about 20% of trials). The striking finding was that infants' occipital cortex responded not only when a visual stimulus appeared, but also when an expected visual stimulus was omitted. Importantly, infants' occipital cortex did not show similar levels of activation in a control condition in which infants did not learn an association between visual and auditory stimuli. Infants' cortical responses in this condition reflected the type of incoming input: when an auditory stimulus was presented without a visual stimulus, temporal cortex, but not occipital cortex, showed changes in blood oxygenation level. Infants who had formed associations between auditory and visual stimuli, on the other hand, showed a strong occipital response to the exact same auditory stimulus. Infants' cortical responses do not simply reflect bottom-up visual input; instead, early cortical responses reflect in part what infants expect to see.

Infants also rapidly form expectations about patterns in upcoming auditory input. For instance, in studies measuring event-related potentials in infants, 3-month-old infants exposed to sequences of repeated auditory stimuli (such as the syllable *i*, i.e. *i*–*i*–*i*–*i*) will show an early cortical mismatch response analogous

to the adult MMN when a novel auditory oddball (such as the syllable *a*) breaks this repetition (Dehaene-Lambertz and Dehaene 1994). Moreover, infants show a later cortical response (the late negative slow wave, NSW) depending on whether the sequence as a whole (regardless of local deviations) is expected (Basirat, Dehaene and Dehaene-Lambertz 2014). In light of predictive coding explanations of early mismatch responses in adults (Wacongne, Changeux and Dehaene 2012; Wacongne et al. 2011), these patterns of cortical responses suggest that infants form both local (about the next syllable in a sequence) and global (about the frequency of a sequence as a whole) predictions about auditory sequences (Basirat, Dehaene and Dehaene-Lambertz 2014).

Together, these results provide diverse evidence that infants' early visual and auditory processing is future-oriented: from an early age and across a variety of tasks, infants generate predictions about upcoming perceptual input and organize their behaviour in anticipation of expected perceptual events.

3.3 Language

By the latter half of their first year, infants have begun to form expectations about the words they commonly hear in their environment and their meanings (Bergelson and Swingley 2012). Forming word-like representations appears to change how infants process auditory sequences such as those used in the oddball paradigm (Dehaene-lambertz and Dehaene 1994), allowing infants to more rapidly process auditory information when new syllables are consistent with their linguistic knowledge. In one study, 12 and 24-month-old Finnish-speaking infants recognized an unexpected auditory syllable such as [kɑ] more quickly (as indexed by an earlier differential electrophysiological brain response) when it was in the context of the familiar word *kukka* (flower in Finnish) than as an isolated syllable (Ylinen et al. 2017), suggesting that infants use linguistic context to form expectations about upcoming syllables based on their knowledge about word forms.

Infants also begin to develop the ability to use their word knowledge to make predictions about their visual environment over the course of their second year of life. While there is little direct evidence that infants are able to make visual predictions based on the words they are hearing before around two years of age, there is intriguing evidence that infants' cortical processing shows early distinct ERP signatures in response to unexpected word-object pairings, similar to the N400 response to semantic violations found in adults. By 12 months of age, infants show an early negative event-related potential when viewing images of known objects and listening to familiar words (Friedrich and Friederici 2004, 2005). Differences between familiar words that match versus familiar words that

do not match an image emerge between 100–250 ms post auditory stimulus onset. Given how rapidly these responses to mismatching words unfold, and the evidence from adults that ERP signatures of this kind may stem from errors in prediction, these ERP signatures may reflect violations of infants' expectations about the words they will hear in the context of a known object. Similar ERP effects emerge when exposing infants as young as 3–6 months to violations of newly learned word-object associations (Friedrich and Friederici 2011, 2017), suggesting that infants are rapidly forming expectations about how a novel word relates to the visual world.

Infants and children form linguistic expectations that they can use to recognize not only words presented in isolation, but also when processing sentences. Many key results come from the looking-while-listening paradigm, in which infants view a set of images (usually two, e.g. an image of a ball and an image of a shoe), one of which is subsequently labeled (*Where is the ball?*). The speed and accuracy with which children fixate the target image is a measure of children's language processing, particularly their ability to recognize the target noun. Using this paradigm, researchers have shown that, around the ages of 2 and 3 children can use verb semantics (Mani and Huettig 2012), grammatical gender (Lew-Williams and Fernald 2007), and even coarticulatory cues (Mahr et al. 2015) to recognize word meanings more quickly. Lew Williams and Fernald (2007) find that 3-year-old Spanish children shift looking towards the target image faster when the grammatical gender of the name of the target image and of the distractor image differ (i.e. the gender of the article disambiguates the two images). Mahr et al. (2015) showed that infants can use coarticulatory information present in the vowel of the word *the* to more efficiently process a subsequent noun. Including coarticulatory information about the upcoming noun leads to faster looking towards the target image as compared to a condition that does not include coarticulatory information.

These contextual effects in language processing are subject to the question raised earlier about whether facilitating effects are due to prediction or to more rapid integration of upcoming information. For some types of linguistic cues, however, in particular semantic cues, the results are more clear-cut that children can predict upcoming language input (Gambi, Pickering and Rabagliati 2016; Gambi et al. 2018; Mani and Huettig 2012). Mani and Huettig (2012) show that 2-year-olds use the meaning of verbs to anticipate which noun they will encounter. When hearing a verb such as *eat* (but not a neutral verb such as *see*), children begin looking toward a picture of a cake (rather than an image of an inedible object) even before the noun *cake* occurs. These predictions do not appear to rely merely on associations: Gambi et al. (2016) find that when hearing the verb *ride*

in a sentence such as *Pingu will ride the horse*, 3–5-year-olds look predictively to an image of a probable patient such as horse. However, children do not look toward a picture of a cowboy, which is also strongly associated with the word *ride*, but unlikely to take the patient role in the sentence. Interestingly, the ability to predict upcoming nouns from verb semantics relates strongly to vocabulary knowledge in 3–10-year-old children (Borovsky, Elman and Fernald 2012; see also Mani and Huettig 2012).[6] Though most studies in language processing in infants show facilitative, rather than truly anticipatory effects of visual and linguistic cues, the general picture that emerges is that infants use an array of cues to form expectations about incoming linguistic input.

4 Prediction and learning

4.1 Learning in "bootstrap heaven"

Infants are actively predicting what will occur in the world around them, in particular how linguistic signals will unfold over time. But what are these predictions for? Is predicting simply a processing strategy adopted "for the moment", with errors in prediction hastily discarded to anticipate the next input? Or is prediction a processing principle that is more deeply connected to how a cognitive system develops? One of the most intriguing possibilities is that generating predictions is integrally connected to learning (Huettig 2015; Rabagliati, Gambi and Pickering 2016; O'Reilly, Wyatte and Rohrlich 2014).

On predictive processing accounts, learning is a natural consequence of the mechanisms by which we perceive the world (Clark 2015). When we experience an unexpected event, the discrepancies between top-down predictions and bottom-up input are fed forward through the processing system, essentially becoming error signals that catalyse learning. Prediction-generating models are revised and adjusted in response to these error signals, which changes the kinds of predictions we will make for future events. In other words, every perceptual event is simultaneously a learning event.

6 Since these data are correlational, there is an interesting question as to the direction of the causal effect here (see also Reuter et al. 2018 for additional evidence with children between 12 and 24 months of age). Are children with larger vocabularies better able to predict upcoming input? Or are children who are better predictors more effective word learners? Or is there some third variable (e.g. some construct such as "general intelligence") that is the source of the relationship?

Viewing prediction error as a learning signal is particularly attractive for developmental theories because it reframes the developmental question of how infants are able to learn so much over the first years of life. Infants, on this view, are "learning in bootstrap heaven" (Clark 2015: 17). This picture of early cognitive development stands in clear contrast to a traditional view of infants as passive organisms faced with James's "blooming, buzzing confusion". Instead, by actively attempting to predict what will happen next, infants can exploit the dense information in their world as a rich source of error signal, which in turn catalyses learning. This point is made eloquently by O'Reilly and colleagues (2014):

> [P]redictive forms of learning are particularly compelling because they provide a ubiquitous source of learning signals: if you attempt to predict everything that happens next, then every single moment is a learning opportunity. This kind of pervasive learning can for example explain how an infant seems to magically acquire such a sophisticated understanding of the world, despite their seemingly inert overt behavior … – they are becoming increasingly expert predictors of what they will see next, and as a result, developing increasingly sophisticated internal models of the world. (O'Reilly, Wyatte, and Rohrlich 2014: 3)

In the following sections, I explore the idea that prediction – in particular responding to prediction errors – is crucially involved in the learning process.

4.2 Prediction error and learning

The idea that prediction error is intimately connected with learning has enjoyed broad application in psychology. The key insight that many models grounded in prediction share is that prediction error is not only a *signal* to update expectations, it is also a *guide* as to how to update expectations. Prediction error communicates information about which expectations to adjust: for instance, by tracing an error backwards through a generative model, a model can adjust the specific expectations that contributed to the error. This is a key idea behind the training of neural networks, discussed below (Rumelhart, Hinton and Williams 1986). Prediction error also contains information about how strongly to adjust expectations, a key idea behind the Rescorla-Wagner model of association learning.

The Rescorla-Wagner model is one of the most productive applications in psychology of the idea that prediction error drives learning. In its basic form, the Rescorla-Wagner model provides a rule according to which a learner should adjust an association between two stimuli. Crucially, the model updates associations according to the difference between actual and expected outcomes, the prediction error. Originally proposed as a descriptive model of conditioning in animals, the model has seen broad application across psychology (Miller, Barnet

and Grahame 1995), and has been successfully applied in explaining diverse phenomena in language development, such as how infants learn word meanings (Baayen et al. 2016) and inflectional morphology (Ramscar, Dye and McCauley 2013). Modern reinforcement learning models built on this basic prediction-based learning mechanism have been expanded to learning complex structured representations, that allow a system to e.g. map a spatial environment or make context-sensitive decisions in a multidimensional task (Niv et al. 2015; Gershman 2017; Daw 2012).

In a different modeling tradition, the notion of prediction error as a driver of learning has been extensively mined in research on neural networks (McClelland et al. 2010; Rumelhart and Todd 1993; McClelland and Rumelhart 1981; Rumelhart and McClelland 1986), particularly in models implementing backpropagation of error (Rumelhart et al. 1986). In the backpropagation algorithm, errors between model output and target are fed backward through the neural network model, with the weights between individual nodes in the network continuously adjusted (or "penalized") according to how much they contributed to the error (Hinton 2014; Rumelhart et al. 1986). Greater error means greater adjustment of the weights responsible for error, in the service of reducing future error. In other words, greater prediction error leads to larger revisions of the underlying model governing the system's predictions. Error in a model's output is both the signal to learn and the guide as to how to update the system.

While these modeling traditions show the power of prediction-error driven learning, to what extent is there support that our brains function in this manner? A long line of evidence has documented that prediction errors are encoded in the brain (e.g. Schultz and Dickinson 2000; Schultz, Dayan and Montague 1997; O'Doherty et al. 2004) and influence reward-seeking behavior (e.g. Pessiglione et al. 2006). Recent evidence suggests that prediction error plays a more general role in the neural implementation of learning. In one study, participants performed a visual-detection task in the presence of auditory distractors (den Ouden et al. 2009). Unbeknownst to the subjects, some auditory distractors were predictive of the presence or absence of the visual stimulus. Across the course of the experiment, the visual primary cortex (V1) showed progressively greater activation to unpredicted and progressively less activation to predicted visual stimuli, demonstrating learning of the dependency between predictive auditory distractors and visual targets. Moreover, participants showed greater response in V1 for unexpected stimuli even in the absence of a visual stimulus, indicating that the activations being measured were truly prediction error responses and not simply (more or less attenuated) bottom-up visual inputs. Most interestingly, the magnitude of prediction error predicted changes in visual-auditory connectivity. This

indicates that prediction error not only encodes violations of expectation (i.e. surprise), but also plays a functional role in learning, reshaping connections to adapt to ongoing tasks.

4.3 Patterns from predictions: What recurrent neural networks can learn

An early illustration of the power of learning driven by prediction is Elman's (1990, 1991) recurrent neural network model of sentence processing. Elman set out to adapt neural network models to predict outcomes as they unfold over time. Elman constructed a simple three-layer neural network model with a deceptively simple tweak: he introduced a context layer that copies hidden unit activations from the previous learning event. The context layer subsequently provides inputs to the hidden units in the next learning event (see figure 10.1). This creates a recurrent processing loop that allows previous representations to influence current activations. The model was given a very simple task: given the current word, predict what word will come next. For this task, the model was fed the model a corpus of two- and three-word sentences with simple subject-verb and subject-verb-object structure. Although the model was not tasked with discovering syntactic structure or semantic relationships between words, the model's hidden units developed latent structure that represented complex grammatical and semantic relationships between words – since these prove helpful to the task of predicting what word will come next. For instance, the hidden units represented nouns differently from verbs, even though words were not tagged with this information. The model also represented semantically similar words as more similar to each other: for example, inanimate nouns were represented as more similar to each other than animate nouns. This latent structure in the hidden units of the model emerged simply as a consequence of the model attempting to minimize prediction error on the next word it encountered. The lesson from Elman's model is that relatively complex representations of the kind needed in language processing can emerge from a simple mechanism – predicting what word will come next (Elman 1990, 1991, 2004, 2009).

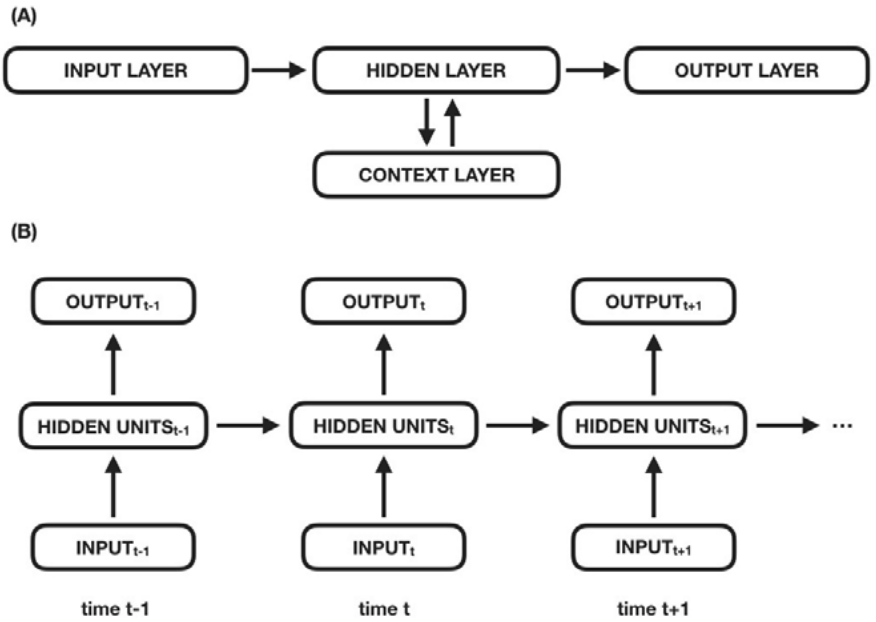

Fig. 10.1: Schematic representation of a simple recurrent network
(A) A simple representation of a three-layer recurrent neural network (Elman 1990). (B) The same recurrent network "unfolded" in time to illustrate its recurrent structure. At time t, the hidden layer receives input both from the input layer at time t and crucially from the hidden layer of the previous time step t-1.

The same logic has been fruitfully extended to show that recurrent neural networks can learn non-adjacent dependencies (Cleeremans and McClelland 1991; Willits 2013), sequence and event structure (Botvinick and Plaut 2004, 2006), abstract rule-like structure (Willits 2013), semantic categories from child-directed speech corpora (Huebner and Willits 2018), as well as perform more complex language comprehension and production tasks (Chang, Dell and Bock 2006; Chang 2002). For instance, Chang and colleagues (2006) developed a model of language processing that learned from a far greater set of training sentences than Elman's original model and was subsequently tested on both comprehension and production. The model succeeded at developing relatively complex abstract syntactic representations. While the architectural constraints underlying the model were far more complex than Elman's original model, the fundamental task of the model and the mechanism by which the model learned remained the same. The model incrementally predicted upcoming words, and when a prediction deviated

from the target word, the weights of the model were updated according to the source of the prediction error.

Recent advances in the architecture of recurrent neural networks have placed these models at the forefront of the state-of-the-art in natural language processing (LeCun, Bengio and Hinton 2015). Modern recurrent networks are not only excellent at learning to predict the next word in a sequence (Mikolov et al. 2013), but the underlying representations yield structure that can be used to solve surprisingly complex tasks. For instance, in machine translation, the hidden units learned by a model trained to probabilistically predict upcoming English words can subsequently be used to generate a (probabilistic) French translation of the English sentence (Cho et al. 2014). Recurrent neural networks can be used in a similar fashion to generate image captions by "translating" high-level image representations generated by neural networks into phrases (Vinyals et al. 2015). Recurrent neural networks are also at the forefront of speech recognition, with modern networks converting audio into text with surprising accuracy (Graves and Jaitly 2014; Graves, Mohamed and Hinton 2013).

Some caution is warranted in drawing strong conclusions about predictive mechanisms from these successes, since many of these breakthroughs depend on specific modeling techniques, e.g. adjustments to the memory structure of the model that allow it to learn otherwise difficult long-term dependencies.[7] While the key idea of predicting an upcoming word in a sequence is preserved, the architecture and training methods are much more complex than in Elman's (1990) original simple architecture, and it is still unclear how these architectures relate to the cognitive architecture of the mind. More generally, how recurrent neural networks actually succeed at diverse tasks once trained – the computations they perform – is still something of a black box. However, recent research is beginning to investigate the underlying computations performed in recurrent neural networks (Sussillo and Barak 2013) and to demonstrate analogs to neural dynamics, e.g. in the prefrontal cortex (Mante et al. 2013). Recurrent predictive processing is rapidly being recognized not just as a framework for creating surprisingly powerful models capable of discovering complex patterns in visual and linguistic data, but a promising framework for understanding the architecture of the mind (Hunt and Hayden 2017).

7 See http://colah.github.io/posts/2015-08-Understanding-LSTMs/ (last accessed June 23, 2018) for an accessible explanation – along with excellent visualizations – of some of the key features of these architectures.

5 Predicting patterns: Prediction and statistical learning

One of the most fruitful discoveries of the past twenty years has been uncovering the powerful statistical learning mechanisms that support pattern-learning from early infancy (see Aslin 2017, and Saffran and Kirkham 2018 for two recent reviews; see Romberg and Saffran 2010 for a review focusing on the role of statistical learning in language development). A seminal finding in statistical learning is that infants can use transitional probabilities to learn word boundaries in a continuous sequence of spoken syllables (Saffran, Aslin and Newport 1996; Aslin, Saffran and Newport 1998). In Saffran et al. (1996), 8-month-old infants heard a spoken sequence constructed from four nonsense words presented in random order, resulting in a continuous auditory stream, e.g. *pabikugolatudaropipabikudaropi*... Crucially, the auditory stream contained no acoustic or prosodic cues to word boundaries such as pauses or differences in stress. The only cues to word boundaries were the transitional probabilities between syllables (see figure 10.2 below): syllables within a word (e.g. *pabi*) had higher transitional probabilities (1.0, i.e. *pa* was always followed by *bi*) than syllables between words (0.33, i.e. *ku* was equally likely to be followed by the three beginning syllables *go*, *ti* or *da*, the first syllables in the three other words). After a brief exposure to the auditory stream, infants discriminated "part-words" (constructed from syllables that crossed word boundaries, e.g. *kugola*) from words (e.g. *pabiku*), showing that infants had learned to identify words within the sequence.

This study opened the door to a host of other findings showing that statistical learning mechanisms operate across a variety of domains, including learning visual regularities (Fiser and Aslin 2002; Kirkham, Slemmer and Johnson 2002), predicting actions and events (Baldwin et al. 2008; Endress and Wood 2011; Stahl et al. 2014), and learning in social contexts (Tummeltshammer et al. 2014; Wu et al. 2011). Statistical learning has also often come to be construed in a broad sense to describe learners' prodigious ability to extract statistical patterns from the input (e.g. Romberg and Saffran 2010). In this more general sense of sensitivity to statistical structure in the environment, statistical learning has been proposed as a method by which infants can learn many aspects of their language environment, including phonological categories (Maye, Werker and Gerken 2002) and learning to map words to their referents (Smith and Yu 2008). Moreover, statistical learning has been argued to aid in uncovering more complex relations such as dependencies between non-adjacent linguistic elements (Gómez 2002; Newport and Aslin 2004) and learning more abstract rule-like patterns (Marcus et al. 1999).

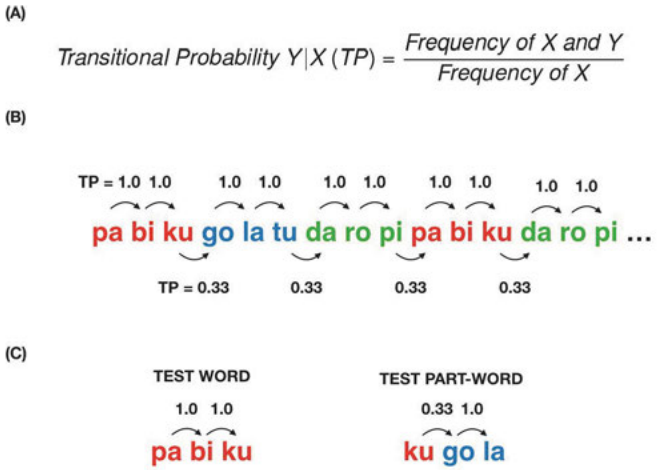

Fig. 10.2: Illustration of the design of Saffran et al. (1996)
(A) The formula for computing transitional probabilities. (B) An example of an auditory sequence of syllables from the experiment, including the transitional probability between syllables. The colours are used here only to illustrate the underlying structure of the auditory stimulus stream and do not reflect a difference in auditory cues. (C) Examples of the words and part-words used at test, along with their statistical structure.

One question that remains controversial is how statistical learning mechanisms operate. In the case of word segmentation, the initial proposal was that participants compute transitional probabilities between items in a sentence (Aslin 2017; Aslin, Saffran and Newport 1998; Saffran, Aslin and Newport 1996). This fits well with a prediction-based account of statistical learning, by which learners are developing probabilistic expectations about upcoming units. These expectations track the transitional evidence in the data over the course of exposure to a continuous stream of syllables. Parsimonious models of sequential pattern learning that are grounded in computing transitional probabilities can account for a diverse pattern of both behavioural and neuropsychological results (Meyniel, Maheu and Dehaene 2016). Other proposals suggest that learners instead extract larger chunks of syllables (French, Addyman and Mareschal 2011; Perruchet and Vinter 1998; see Frank, Goldwater, Griffiths and Tenenbaum 2010 for a comparison of different models) and focus on the role of memory structure in statistical learning (Thiessen 2017).

Regardless of the outcome of these specific debates, there is diverse evidence that statistical learning in general may be grounded in our ability to generate

probabilistic expectations. For instance, adults use statistical learning to anticipate future events and elements in sequences (Misyak, Christiansen and Tomblin 2010; Dale, Duran and Morehead 2012). Moreover, recent evidence from intracranial recordings of neural assemblies suggests that adults generate forward-looking probabilistic predictions about likely upcoming syllables when processing known words (Leonard et al. 2015). Another intriguing line of evidence suggests that statistical learning helps to sharpen our predictions about expected inputs (see also Saffran and Kirkham 2018), with predictability helping to enhance the representation of items involved in a sequence (see e.g. Otsuka and Saiki 2016). For example, more predictable items in a visual sequence become easier to visually detect, suggesting that forming expectations about upcoming elements in a pattern has beneficial consequences for the representation of predictable elements (Barakat, Seitz and Shams 2013).

Amid these models, an important goal for future research will be to tease out the degree to which infants' statistical learning is grounded in developing probabilistic expectations – do infants anticipate upcoming units during statistical learning, and how does this relate to learning? One way to approach this question is to test predictions that follow from explanations grounded in prediction-based probability computation. One prediction of such models is that past transitional probabilities should be preserved such that they can influence later learning: for example, if the transitional probabilities between syllables at an early time point T1 change during a later learning experience at time point T2, the transitional probabilities from T1 should influence the degree to which infant learners adapt to the transitional probabilities at T2.[8]

A second prediction is that the global context within which a pattern is embedded can differentially constrain expectations about upcoming elements in a sequence. For instance, do infants develop higher-level expectations about the predictability of different contexts? In the statistical word segmentation task from Saffran et al. (1996), infants could encounter words with high within-word transitional probabilities in two different contexts: a context with little regularity based on transitional probabilities (i.e. a highly noisy syllable transition context) or a context with a more regular pattern for transitional probabilities (i.e. a more predictable syllable transition context). The degree to which infants make predictions that use transitional probabilities may depend on whether transitional probabilities yield useful predictions in the larger context, not just their

8 One caveat here is that learners are exquisitely attuned to changes in context and are able to rapidly update their expectations to contextual changes (Qian, Jaeger, and Aslin 2016). Thus, care would need to be taken to maintain the continuity of the learning context from T1 to T2.

informativity within a particular word. This fits well with the notion that the particular information that enters into a learner's prediction is highly context-sensitive and may crucially depend on the expected utility of that information for making accurate predictions in the future (see Kuperberg and Jaeger 2016).

6 Language development theory in light of prediction

While prediction offers a unified perspective from which to view language development, it is important to recognize that it does not adjudicate many of the central theoretical debates in the field, in particular the classic dialogue between nativist and empiricist/constructivist theories regarding the origin of linguistic knowledge and the role of experience in language development (see e.g. Ambridge and Lieven 2011 for an overview over theoretical debates in different areas of language). Are the foundations of linguistic knowledge present from birth, or does linguistic knowledge emerge from language experience over the course of development? The prediction-based approaches sketched here appear to be largely agnostic about this question. Crucially, prediction-based theories may vary in how they explain the source of initial expectations that constrain prediction, the types of linguistic elements over which probabilistic expectations are computed and how expectations are updated, leaving room to interpret these mechanisms in terms of domain-general or domain-specific constraints. However, prediction offers a domain-general computational principle operating across language learning mechanisms.

The prediction-based framework may advance theoretical discussion by focusing on learning and inference over statistical patterns. Prediction-based approaches establish a deep continuity between language processing and learning (Chang, Dell and Bock 2006; Kuperberg and Jaeger 2016), helping to connect our understanding of how learners accumulate and exploit statistical knowledge about linguistic patterns (see e.g. MacWhinney and Bates 1987; Seidenberg and MacDonald 1999). A key aspect of prediction-based theories is their emphasis on the ubiquity of learning. Every moment of processing linguistic input is simultaneously providing information updating infants' probabilistic expectations about future language patterns they may encounter. One consequence of this view is that it reframes debates about the "impoverished" nature of leaners' language input (Laurence and Margolis 2001; Chomsky 1965) by demonstrating the vast amounts of probabilistic inferences that can be made from the language

input a child experiences (e.g. Huebner and Willits 2018). The focus of the debate can thus be moved to the problem of constraining possible probabilistic inferences from the rich language data available to a prediction-driven learner (see also, e.g. Clark and Lappin 2011).

Prediction may also help to address specific problems in the language development literature, such as the *no-negative-evidence* problem (Bowerman 1988). Children sometimes overgeneralize in their use of lexical items, e.g. using an intransitive verb transitively in phrases such as *don't giggle me*. Children rarely – if ever – receive direct feedback that these usages are ungrammatical (negative evidence), presenting a puzzle as to how children successfully "unlearn" these ungrammatical forms. Prediction suggests that children might in some sense create negative evidence themselves during learning. If children are creating probabilistic expectations about linguistic structures (e.g. that *giggle* can be used transitively), but these expectations are violated (*giggle* is only used intransitively, and *tickle* is encountered in transitive situations), then children could update their expectations based on the internally generated prediction error (negative evidence). Chomsky himself recognized the potential importance of what he described as "indirect negative evidence" (see also Rabagliati et al. 2016 for discussion):

> [A] not unreasonable acquisition system can be devised with the operative principle that if certain structures or rules fail to be exemplified in relatively simple expressions, *where they would expect to be found*, then a (possibly marked) option is selected excluding them in the grammar, so that a kind of "negative evidence" can be available even without corrections, adverse reactions, etc. (Chomsky 1981: 9; emphasis mine)

Since we are constantly making predictions about upcoming input, we are generating, in some sense, our own evidence as we develop more refined linguistic expectations.

7 Conclusion

The task faced by young learners of language is daunting. Syllable after syllable unfolds at a rapid pace, with ambiguity at virtually all levels of processing. Prediction offers a framework for understanding how infants succeed at this task by exploiting patterns in their language environment to develop expectations about upcoming auditory signals and the meanings they communicate. There are many questions left unanswered in prediction-based theories in their current form – simply recognizing the predictive nature of infants' early language learning cannot explain language development in all of its complexity. However, the

prediction framework offers a fruitful way of unifying many of the central insights in the field and opens up new avenues for exploring how infants come to uncover the patterns in language.

References

Ambridge, Ben & Elena V. M. Lieven. 2011. *Child language acquisition: Contrasting theoretical approaches*. Cambridge: Cambridge University Press.

Andics, Attila, Viktor Gal, Klara Vicsi, Gabor Rudas & Zoltan Vidnyanszky. 2013. FMRI repetition suppression for voices is modulated by stimulus expectations. *NeuroImage* 69. 277–283.

Aslin, Richard N. 2014. Infant learning: Historical, conceptual, and methodological challenges. *Infancy* 19 (1). 2–27.

Aslin, Richard N. 2017. Statistical learning: A powerful mechanism that operates by mere exposure. *Wiley Interdisciplinary Reviews: Cognitive Science* 8 (1–2). 1–7.

Aslin, Richard N., Jenny R. Saffran & Elissa L. Newport. 1998. Computation of conditional probability statistics by 8-month-old infants. *Psychological Science* 9 (4). 321–324.

Aslin, Richard N. 2007. What's in a look? *Developmental Science* 10 (1). 48–53.

Baayen, R. Harald, Cyrus Shaoul, Jon Willits & Michael Ramscar. 2016. Comprehension without segmentation: A proof of concept with naive discriminative learning. *Language, Cognition and Neuroscience* 31 (1). 106–128.

Baillargeon, Renée & Susan Carey. 2012. Core cognition and beyond: The acquisition of physical and numerical knowledge. In Sabina Pauen (ed.), *Early childhood development and later achievement*, 33–65. Cambridge: Cambridge University Press.

Baillargeon, Renée, Elizabeth S. Spelke & Stanley Wasserman. 1985. Object permanence in five-month-old infants. *Cognition* 20. 191–208.

Baldwin, Dare, Annika Andersson, Jenny Saffran & Meredith Meyer. 2008. Segmenting dynamic human action via statistical structure. *Cognition* 106 (3). 1382–1407.

Bar, Moshe. 2009. The proactive brain: Memory for predictions. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 364 (1521). 1235–1243.

Barakat, Brandon K., Aaron R. Seitz & Ladan Shams. 2013. The effect of statistical learning on internal stimulus representations: Predictable items are enhanced even when not predicted. *Cognition* 129 (2). 205–211.

Basirat, Anahita, Stanislas Dehaene & Ghislaine Dehaene-Lambertz. 2014. A hierarchy of cortical responses to sequence violations in three-month-old infants. *Cognition* 132 (2). 137–150.

Bendixen, Alexandra, Mathias Scharinger, Antje Strauss & Jonas Obleser. 2014. Prediction in the service of comprehension: Modulated early brain responses to omitted speech segments. *Cortex* 53 (1). 9–26.

Bergelson, Elika & Daniel Swingley. 2012. At 6–9 months, human infants know the meanings of many common nouns. *Proceedings of the National Academy of Sciences of the United States of America* 109 (9). 3253–3258.

Berkum, Jos J. A. van, Colin M. Brown, Pienie Zwitserlood, Valesca Kooijman & Peter Hagoort. 2005. Anticipating upcoming words in discourse: Evidence from ERPs and reading times. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 31 (3). 443–467.

Borovsky, Arielle, Jeffrey L. Elman & Anne Fernald. 2012. Knowing a lot for one's age: Vocabulary skill and not age is associated with anticipatory incremental sentence interpretation in children and adults. *Journal of Experimental Child Psychology* 112 (4). 417–436.

Botvinick, Matthew M & David C Plaut. 2004. Doing without schema hierarchies: A recurrent connectionist approach to normal and impaired routine sequential action. *Psychological Review* 111 (2). 395–429.

Botvinick, Matthew M & David C Plaut. 2006. Short-term memory for serial order: A recurrent neural network model. *Psychological Review* 113 (2). 201–33.

Bowerman, M. 1988. The "no negative evidence" problem: How do children avoid constructing an overly general grammar? In John A. Hawkins (ed.), *Explaining language universals*, 73–101. Oxford: Blackwell.

Brothers, Trevor, Tamara Y. Swaab & Matthew J. Traxler. 2015. Effects of prediction and contextual support on lexical processing: Prediction takes precedence. *Cognition* 136. 135–149.

Canfield, Richard L & Marshall M Haith. 1991. Young infants' visual expectations for symmetric and asymmetric stimulus sequences. *Developmental Psychology* 27 (2). 198–208.

Canfield, Richard L, Elliott G Smith, Michael P Brezsnyak, Kyle L Snow, Richard N Aslin, Marshall M Haith, Tara S Wass & Scott A Adler. 1997. Information processing through the first year of life: A longitudinal study using the visual expectation paradigm. *Monographs of the Society for Research in Child Development* 62 (2). 1–160.

Chang, Franklin. 2002. Symbolically speaking: A connectionist model of sentence production. *Cognitive Science* 26 (5). 609–651.

Chang, Franklin, Gary S Dell & Kathryn Bock. 2006. Becoming syntactic. *Psychological Review* 113 (2). 234–272.

Chennu, Srivas, Valdas Noreika, David Gueorguiev, Alejandro Blenkmann, Silvia Kochen, Agustín Ibáñez, Adrian M Owen & Tristan A Bekinschtein. 2013. Expectation and attention in hierarchical auditory prediction. *The Journal of Neuroscience* 33 (27). 11194–11205.

Cho, Kyunghyun, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk & Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In Alessandro Moschitti, Bo Pang & Walter Daelemans (eds.), *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, 1724–1734. Stroudsburg, PA: Association for Computational Linguistics.

Chomsky, Noam. 1959. A review of B.F. Skinner's Verbal Behavior. *Language* 35 (1). 26–58.

Chomsky, Noam. 1965. *Aspects of a theory of syntax*. Cambridge, MA: MIT Press.

Chomsky, Noam. 1980. On cognitive structures and their development: A reply to Piaget. In Massimo Piatelli-Palmarini (ed.), *Language and learnability: The debate between Jean Piaget and Noam Chomsky*, 35–52. Cambridge, MA: Harvard University Press.

Chomsky, Noam. 1981. *Lectures on government and binding*. Dordrecht: Foris.

Clark, Alexander & Shalom Lappin. 2011. *Linguistic nativism and the poverty of the stimulus*. Chichester: Wiley-Blackwell.

Clark, Andy. 2013. Whatever next? Predictive brains, situated agents, and the future of cognitive science. *The Behavioral and Brain Sciences* 36 (3). 181–204.

Clark, Andy. 2015. *Surfing uncertainty: Prediction, action, and the embodied mind*. Oxford: Oxford University Press.

Cleeremans, Axel & JL McClelland. 1991. Learning the structure of event sequences. *Journal of Experimental Psychology* 120 (3). 235–253.
Dale, Rick, Nicholas D Duran & J Ryan Morehead. 2012. Prediction during statistical learning, and implications for the implicit/explicit divide. *Advances in Cognitive Psychology* 8 (2). 196–209.
Daw, Nathaniel D. 2012. Model-based reinforcement learning as cognitive search: Neurocomputational theories. In Peter M. Todd, Thomas T. Hills & Trevor W. Robbins (eds.), *Cognitive search: Evolution, algorithms, and the brain*, 195–208. Cambridge, MA: MIT Press.
Dehaene-Lambertz, Ghislaine & Elizabeth S. Spelke. 2015. The infancy of the human brain. *Neuron* 88 (1). 93–109.
Dehaene-Lambertz, Ghislaine & Stanislas Dehaene. 1994. Speed and cerebral correlates of syllable discrimination in infants. *Nature* 370 (6487). 292–295.
DeLong, Katherine A., Thomas P. Urbach & Marta Kutas. 2005. Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nature Neuroscience* 8 (8). 1117–1121.
DeLong, Katherine A., Thomas P. Urbach & Marta Kutas. 2017a. Is there *a* replication crisis? Perhaps. Is this *an* example? No: a commentary on Ito, Martin, and Nieuwland (2016). *Language, Cognition and Neuroscience* 32 (8). 966–973.
DeLong, Katherine A., Thomas P. Urbach & Marta Kutas. 2017b. Concerns with Nieuwland et al. multi-lab replication attempt (2017). (http://kutaslab.ucsd.edu/pdfs/FinalDUK17 Comment9LabStudy.pdf; accessed 24 June 2018)
Dikker, Suzanne, Hugh Rabagliati, Thomas A. Farmer & Liina Pylkkänen. 2010. Early occipital sensitivity to syntactic category is based on form typicality. *Psychological Science* 21 (5). 629–634.
Elman, Jeffrey L. 1990. Finding structure in time. *Cognitive Science* 14 (2). 179–211.
Elman, Jeffrey L. 1991. Distributed representation, simple recurrent networks, and grammatical structure. *Machine Learning* 7. 195–225.
Elman, Jeffrey L. 2004. An alternative view of the mental lexicon. *Trends in Cognitive Sciences* 8 (7). 301–306.
Elman, Jeffrey L. 2009. On the meaning of words and dinosaur bones: Lexical knowledge without a lexicon. *Cognitive Science* 33. 547–582.
Elman, Jeffrey L., Elizabeth A Bates, Mark H Johnson, Annette Karmiloff-Smith, Domenico Parisi & Kim Plunkett. 1996. *Rethinking innateness: A connectionist perspective on development*. Cambridge, MA: MIT Press.
Emberson, Lauren L., John E. Richards & Richard N. Aslin. 2015. Top-down modulation in the infant brain: Learning-induced expectations rapidly affect the sensory cortex at 6 months. *Proceedings of the National Academy of Sciences* 112 (31). 9585–9590.
Endress, Ansgar D. & Justin N. Wood. 2011. From movements to actions: Two mechanisms for learning action sequences. *Cognitive Psychology* 63 (3). 141–171.
Fantz, Robert L. 1961. The origin of form perception. *Scientific American* 204 (5). 66–72.
Fantz, Robert L. 1963. Pattern vision in newborn infants. *Science* 140 (3564). 296–297.
Ferreira, Fernanda & Charles Clifton, Jr. 1986. The independence of syntactic processing. *Journal of Memory and Language* 25. 348–368.
Ferreira, Fernanda & Nikole D. Patson. 2007. The 'good enough' approach to language comprehension. *Language and Linguistics Compass* 1 (1–2). 71–83.

Fiser, József & Richard N. Aslin. 2002. Statistical learning of higher-order temporal structure from visual shape sequences. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 28 (3). 458–467.
Frank, Michael C., Sharon Goldwater, Thomas L. Griffiths & Joshua B. Tenenbaum. 2010. Modeling human performance in statistical word segmentation. *Cognition* 117 (2). 107–125.
French, Robert M., Caspar Addyman & Denis Mareschal. 2011. TRACX: A recognition-based connectionist framework for sequence segmentation and chunk extraction. *Psychological Review* 118 (4). 614–636.
Friedrich, Manuela & Angela D. Friederici. 2017. The origins of word learning: Brain responses of 3-month-olds indicate their rapid association of objects and words. *Developmental Science* 20 (2). e12357.
Friedrich, Manuela & Angela D. Friederici. 2004. N400-like semantic incongruity effect in 19-month-olds: Processing known words in picture contexts. *Journal of Cognitive Neuroscience* 16 (8). 1465–1477.
Friedrich, Manuela & Angela D. Friederici. 2005. Phonotactic knowledge and lexical-semantic processing in one-year-olds: Brain responses to words and nonsense words in picture contexts. *Journal of Cognitive Neuroscience* 17 (11). 1785–1802.
Friedrich, Manuela & Angela D. Friederici. 2011. Word learning in 6-month-olds: Fast encoding-weak retention. *Journal of Cognitive Neuroscience* 23 (11). 3228–3240.
Friston, Karl. 2010. The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience* 11 (2). 127–138.
Gambi, Chiara, Fiona Gorrie, Martin J. Pickering & Hugh Rabagliati. 2018. The development of linguistic prediction: Predictions of sound and meaning in 2-to-5 year olds. *Journal of Experimental Child Psychology* 173. 351–370.
Gambi, Chiara, Martin J. Pickering & Hugh Rabagliati. 2016. Beyond associations: Sensitivity to structure in pre-schoolers' linguistic predictions. *Cognition* 157. 340–351.
Gergely, György, Zoltán Nádasdy, Gergely Csibra & Szilvia Bíró. 1995. Taking the intentional stance at 12 months of age. *Cognition* 56 (2). 165–193.
Gershman, Samuel J. 2017. Predicting the past, remembering the future. *Current Opinion in Behavioral Sciences* 17. 7–13.
Gómez, Rebecca L. 2002. Variability and detection of invariant structure. *Psychological Science* 13 (5). 431–436.
Graves, Alex & Navdeep Jaitly. 2014. Towards end-to-end speech recognition with recurrent neural networks. *JMLR Workshop and Conference Proceedings* 32 (1). 1764–1772.
Graves, Alex, Abdel-rahman Mohamed & Geoffrey Hinton. 2013. Speech recognition with deep recurrent neural networks. *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, 6645–6649.
Grill-Spector, Kalanit, Richard Henson & Alex Martin. 2006. Repetition and the brain: Neural models of stimulus-specific effects. *Trends in Cognitive Sciences* 10 (1). 14–23.
Haith, MM, C Hazan & GS Goodman. 1988. Expectation and anticipation of dynamic visual events by 3.5-month-old babies. *Child Development* 59 (2). 467–479.
Hawkins, Jeff & Sandra Blakeslee. 2004. *On intelligence*. New York: Henry Holt.
Henderson, John M. 2017. Gaze control as prediction. *Trends in Cognitive Sciences* 21 (1). 15–23.
Hespos, Susan J. & Elizabeth S. Spelke. 2004. Conceptual precursors to language. *Nature* 430 (6998). 453–456.
Hinton, Geoffrey. 2014. Where do features come from? *Cognitive Science* 38 (6). 1078–1101.
Hohwy, Jakob. 2014. *The predictive mind*. Oxford: Oxford University Press.

Huebner, Philip A. & Jon A. Willits. 2018. Structured semantic knowledge can emerge automatically from predicting word sequences in child-directed speech. *Frontiers in Psychology* 9. 1–18.

Huettig, Falk. 2015. Four central questions about prediction in language processing. *Brain Research* 1626. 118–135.

Hunt, Laurence T. & Benjamin Y. Hayden. 2017. A distributed, hierarchical and recurrent framework for reward-based choice. *Nature Reviews Neuroscience* 18 (3). 172–182.

Ito, Aine, Andrea E. Martin & Mante S. Nieuwland. 2017. How robust are prediction effects in language comprehension? Failure to replicate article-elicited N400 effects. *Language, Cognition and Neuroscience* 32 (8). 954–965.

Itti, Laurent & Pierre Baldi. 2009. Bayesian surprise attracts human attention. *Vision Research* 49 (10). 1295–1306.

James, William. 1890. *The principles of psychology*. Vol. 1. New York: Holt, Rinehart, and Winston.

Jusczyk, Peter W. & Richard N. Aslin. 1995. Infants' detection of the sound patterns of words in fluent speech. *Cognitive Psychology* 29. 1–23.

Kidd, Celeste & Benjamin Y Hayden. 2015. The psychology and neuroscience of curiosity. *Neuron* 88. 449–460.

Kidd, Celeste, Steven T. Piantadosi & Richard N. Aslin. 2014. The Goldilocks effect in infant auditory attention. *Child Development* 85 (5). 1795–1804.

Kidd, Celeste, Steven T. Piantadosi & Richard N. Aslin. 2012. The Goldilocks effect: Human infants allocate attention to visual sequences that are neither too simple nor too complex. *PLoS One* 7 (5). e36399.

Kirkham, Natasha Z., Jonathan A. Slemmer & Scott P. Johnson. 2002. Visual statistical learning in infancy: Evidence for a domain general learning mechanism. *Cognition* 83 (2). B35–B42.

Kok, Peter, Gijs J. Brouwer, Marcel A. J. van Gerven & Floris P. de Lange. 2013. Prior expectations bias sensory representations in visual cortex. *Journal of Neuroscience* 33 (41). 16275–16284.

Kok, Peter, Michel Failing & Floris P. de Lange. 2014. Prior expectations evoke stimulus templates in the primary visual cortex. *Journal of Cognitive Neuroscience* 26 (7). 1546–1554.

Kok, Peter, Janneke F. M. Jehee & Floris P. de Lange. 2012. Less Is more: Expectation sharpens representations in the primary visual cortex. *Neuron* 75 (2). 265–270.

Kouider, Sid, Bria Long, Lorna Le Stanc, Sylvain Charron, Anne-Caroline Fievet, Leonardo S. Barbosa & Sofie V. Gelskov. 2015. Neural dynamics of prediction and surprise in infants. *Nature Communications* 6. 8537.

Kovács, Agnes Melinda & Jacques Mehler. 2009a. Cognitive gains in 7-month-old bilingual infants. *Proceedings of the National Academy of Sciences* 106 (16). 6556–6560.

Kovács, Agnes Melinda & Jacques Mehler. 2009b. Flexible learning of multiple speech structures in bilingual infants. *Science* 325v(5940). 611–612.

Kuperberg, Gina R. & T. Florian Jaeger. 2016. What do we mean by prediction in language comprehension? *Language Cognition & Neuroscience* 31 (1). 32–59.

Kutas, Marta, Kara D. Federmaier & Thomas P. Urbach. 2014. The "negatives" and "positives" of prediction in language. In Michael Saunders Gazzaniga (ed.), *The cognitive neurosciences*, 649–656. 5th edn. Cambridge: MIT Press.

Kutas, Marta & Kara D. Federmeier. 2011. Thirty years and counting: Finding meaning in the N400 component of the event-related brain potential (ERP). *Annual Review of Psychology* 62. 621–647.

Kutas, Marta & Steven Hillyard. 1980. Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science* 207 (4427). 203–205.
Kutas, Marta & Steven Hillyard. 1984. Brain potentials during reading reflect word expectancy and semantic association. *Nature* 307 (5947). 161–163.
Laurence, Stephen & Eric Margolis. 2001. The poverty of the stimulus argument. *The British Journal for the Philosophy of Science* 52. 217–276.
LeCun, Yann, Yoshua Bengio & Geoffrey Hinton. 2015. Deep learning. *Nature* 521. 436–444.
Leonard, Matthew K., Kristofer E. Bouchard, Claire Tang & Edward F. Chang. 2015. Dynamic encoding of speech sequence probability in human temporal cortex. *Journal of Neuroscience* 35 (18). 7203–7214.
Lew-Williams, Casey & Anne Fernald. 2007. Young children learning Spanish make rapid use of grammatical gender in spoken word recognition. *Psychological Science* 18 (3). 193–198.
MacDonald, Maryellen C., Marcel Adam Just & Patricia A. Carpenter. 1992. Working memory constraints on the processing of syntactic ambiguity. *Cognitive Psychology* 24 (1). 56–98.
MacWhinney, Brian & Elizabeth A. Bates. 1987. Competition, variation, and language learning. In Brian MacWhinney (ed.), *Mechanisms of language acquisition*, 157–194. Hillsdale: Lawrence Erlbaum.
Mahr, Tristan, Brianna T.M. McMillan, Jenny R. Saffran, Susan Ellis Weismer & Jan Edwards. 2015. Anticipatory coarticulation facilitates word recognition in toddlers. *Cognition* 142. 345–350.
Mani, Nivedita & Falk Huettig. 2012. Prediction during language processing is a piece of cake – but only for skilled producers. *Journal of Experimental Psychology: Human Perception and Performance* 38 (4). 843–847.
Mante, Valerio, David Sussillo, Krishna V. Shenoy & William T. Newsome. 2013. Context-dependent computation by recurrent dynamics in prefrontal cortex. *Nature* 503 (7474). 78–84.
Marcus, Gary F., Subramanijan Vijayan, S. Bandi Rao & Peter M. Vishton. 1999. Rule learning by seven-month-old Infants. *Science* 283 (5398). 77–80.
Maye, Jessica, Janet F. Werker & Lou Ann Gerken. 2002. Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition* 82 (3). B101–111.
McClelland, James L. & J. Kevin O'Regan. 1981. Expectations increase the benefit derived from parafoveal visual information in reading words aloud. *Journal of Experimental Psychology: Human Perception and Performance* 7 (3). 634–644.
McClelland, James L. & David E. Rumelhart. 1981. An interactive activation model of context effects in letter perception: Part 1. An account of basic findings. *Psychological Review* 88 (5). 375–407.
McClelland, James L., Matthew M. Botvinick, David C. Noelle, David C. Plaut, Timothy T. Rogers, Mark S. Seidenberg & Linda B. Smith. 2010. Letting structure emerge: Connectionist and dynamical systems approaches to cognition. *Trends in Cognitive Sciences* 14 (8). 348–356.
McMurray, Bob & Richard N. Aslin. 2004. Anticipatory eye movements reveal infants' auditory and visual categories. *Infancy* 6 (2). 203–229.
Meyniel, Florent, Maxime Maheu & Stanislas Dehaene. 2016. Human inferences about sequences: A minimal transition probability model. *PLoS Computational Biology* 12 (12). 1–27.
Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg Corrado & Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. *Proceedings of Advances in Neural Information Processing Systems* 26. 3111–3119.
Miller, Ralph R., Robert C. Barnet & Nicholas J. Grahame. 1995. Assessment of the Rescorla-Wagner model. *Psychological Bulletin* 117 (3). 363–386.

Misyak, Jennifer B., Morten H. Christiansen & J. Bruce Tomblin. 2010. Sequential expectations: The role of prediction-based learning in language. *Topics in Cognitive Science* 2 (1). 138–153.
Moon, Christine, Robin Panneton Cooper & William P. Fifer. 1993. Two-day-olds prefer their native language. *Infant Behavior and Development* 16 (4). 495–500.
Newport, Elissa L. & Richard N. Aslin. 2004. Learning at a distance I. Statistical learning of non-adjacent dependencies. *Cognitive Psychology* 48 (2). 127–162.
Nieuwland, Mante S., Stephen Politzer-Ahles, Evelien Heyselaar, Katrien Segaert, Emily Darley, Nina Kazanina, Sarah von Grebmer zu Wolfsthurn et al. 2018. Large-scale replication study reveals a limit on probabilistic prediction in language comprehension. *eLife* 7. e33468.
Niv, Yael, Reka Daniel, Andra Geana, Samuel J. Gershman, Yuang Chang Leong, Angela Radulescu & Robert C. Wilson. 2015. Reinforcement learning in multidimensional environments relies on attention mechanisms. *Journal of Neuroscience* 35 (21). 8145–8157.
O'Doherty, John, Peter Dayan, Johannes Schutlz, Ralf Deichmann, Karl J. Friston & Raymond J. Dolan. 2004. Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *Science* 304 (5669). 452–454.
O'Reilly, Randall C., Dean Wyatte & John Rohrlich. 2014. Learning through time in the thalamocortical loops. *arXiv* 1407.3432. (https://arxiv.org/abs/ 1407.3432; accessed 24 June 2018)
Otsuka, Sachio & Jun Saiki. 2016. Gift from statistical learning: Visual statistical learning enhances memory for sequence elements and impairs memory for items that disrupt regularities. *Cognition* 147. 113–126.
Ouden, Hanneke E. M. den, Karl Friston, Nathaniel D. Daw, Anthony R. McIntosh & Klaas E. Stephan. 2009. A dual role for prediction error in associative learning. *Cerebral Cortex* 19 (5). 1175–1185.
Perruchet, Pierre & Annie Vinter. 1998. PARSER: A model for word segmentation. *Journal of Memory and Language* 39 (2). 246–263.
Pessiglione, Mathias, Ben Seymour, Guillaume Flandin, Raymond J. Dolan & Chris D. Frith. 2006. Dopamine-dependent prediction errors underpin reward-seeking behaviour in humans. *Nature* 442 (7106). 1042–1045.
Piaget, Jean. 1971. *Biology and knowledge*. Chicago: Chicago University Press.
Piantadosi, Steven T., Celeste Kidd & Richard Aslin. 2014. Rich analysis and rational models: Inferring individual behavior from infant looking data. *Developmental Science* 17 (3). 321–337.
Pickering, Martin J. & Andy Clark. 2014. Getting ahead: Forward models and their place in cognitive architecture. *Trends in Cognitive Sciences* 18 (9). 451–456.
Pickering, Martin J. & Simon Garrod. 2007. Do people use language production to make predictions during comprehension? *Trends in Cognitive Sciences* 11 (3). 105–110.
Qian, Ting, T. Florian Jaeger & Richard N. Aslin. 2016. Incremental implicit learning of bundles of statistical patterns. *Cognition* 157. 156–173.
Rabagliati, Hugh, Chiara Gambi & Martin J. Pickering. 2016. Learning to predict or predicting to learn? *Language, Cognition and Neuroscience* 31 (1). 94–105.
Ramscar, Michael, Melody Dye & Stewart M. McCauley. 2013. Error and expectation in language learning: The curious absence of mouses in adult speech. *Language* 89 (4). 760–793.
Rao, Rajesh P. N. & Dana H. Ballard. 1999. Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience* 2 (1). 79–87.
Reuter, Tracy, Lauren Emberson, Alexa R. Romberg & Casey Lew-Williams. 2018. Individual differences in nonverbal prediction and vocabulary size in infancy. *Cognition* 176. 215–219.
Romberg, Alexa R. & Jenny R. Saffran. 2010. Statistical learning and language acquisition. *Wiley Interdisciplinary Reviews: Cognitive Science* 1 (6). 906–914.

Romberg, Alexa R. & Jenny R. Saffran. 2013. Expectancy learning from probabilistic input by infants. *Frontiers in Psychology* 3. 610.
Rumelhart, David E., Geoffrey E. Hinton & Ronald J. Williams. 1986. Learning representations by back-propagating errors. *Nature* 323 (6088). 533–536.
Rumelhart, David E. & James L. McClelland. 1986. On learning the past tense of English verbs. In David E. Rumelhart & James L. McClelland (eds.), *Parallel distributed processing*. Vol. 2: *Psychological and biological models*, 216–271. Cambridge: MA: MIT Press.
Rumelhart, David E. & Peter M. Todd. 1993. Learning and connectionist representations. In David E. Meyer & Sylvan Kornblum (eds.), *Attention and performance XIV: Synergies in experimental psychology, artificial intelligence, and cognitive neuroscience* 2, 3–30. Cambridge, MA: MIT Press.
Rumelhart, David E., Geoffrey E. Hinton & Ronald J. Williams. 1986. Learning internal representations by error propagation. In David E. Rumelhart, James L. McClelland & PDP Research Group (eds.), *Parallel distributed processing*. Vol. 1: *Foundations: explorations in the microstructure of cognition*, 318–362. Cambrigde, MA: MIT Press.
Saffran, Jenny R., Richard N. Aslin & Elissa L. Newport. 1996. Statistical learning by 8-month-old infants. *Science* 274 (5294). 1926–1928.
Saffran, Jenny R. & Natasha Z. Kirkham. 2018. Infant statistical learning. *Annual Review of Psychology* 69 (2). 1–23.
Schultz, Wolfram, Peter Dayan & P. Read Montague. 1997. A neural substrate of prediction and reward. *Science* 275 (5306). 1593–1599.
Schultz, Wolfram & Anthony Dickinson. 2000. Neuronal coding of prediction errors. *Annual Review of Neuroscience* 23. 473–500.
Seidenberg, Mark S. & Maryellen C. MacDonald. 1999. A probabilistic constraints approach to language acquisition and processing. *Cognitive Science* 23 (4). 569–588.
Smith, Linda B. & Esther Thelen. 2003. Development as a dynamic system. *Trends in Cognitive Sciences* 7 (8). 343–348.
Smith, Linda B. & Chen Yu. 2008. Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition* 106 (3). 1558–1568.
Spelke, Elizabeth S. & Katherine D. Kinzler. 2007. Core knowledge. *Developmental Science* 10 (1). 89–96.
Stahl, Aimee E., Alexa R. Romberg, Sarah Roseberry, Roberta Michnick Golinkoff & Kathryn Hirsh-Pasek. 2014. Infants segment continuous events using transitional probabilities. *Child Development* 85 (5). 1821–1826.
Stanovich, Keith E. & Richard F. West. 1979. Mechanisms of sentence context effects in reading: Automatic activation and conscious attention. *Memory & Cognition* 7 (2). 77–85.
Staub, Adrian. 2015. The effect of lexical predictability on eye movements in reading: Critical review and theoretical interpretation. *Language and Linguistics Compass* 9 (8). 311–327.
Summerfield, Christopher, Emily H. Trittschuh, Jim M. Monti, M. Marsel Mesulam & Tobias Egner. 2008. Neural repetition suppression reflects fulfilled perceptual expectations. *Nature Neuroscience* 11 (9). 1004–1006.
Summerfield, Christopher, Valentin Wyart, Vanessa Mareike Johnen & Vincent de Gardelle. 2011. Human scalp electroencephalography reveals that repetition suppression varies with expectation. *Frontiers in Human Neuroscience* 5 (July). 67.
Sussillo, David & Omri Barak. 2013. Opening the black box: Low-dimensional dynamics in high-dimensional recurrent neural networks. *Neural Computation* 25 (3). 626–649.

Szewczyk, Jakub M. & Herbert Schriefers. 2013. Prediction in language comprehension beyond specific words: An ERP study on sentence comprehension in Polish. *Journal of Memory and Language* 68 (4). 297–314.

Thiessen, Erik D. 2017. What's statistical about learning? Insights from modeling statistical learning as a set of memory processes. *Philosophical Transactions of the Royal Society B: Biological Sciences* 372 (1711). 20160056.

Todorovic, Ana, Freek van Ede, Eric Maris & Floris P. de Lange. 2011. Prior expectation mediates neural adaptation to repeated sounds in the auditory cortex: an MEG study. *The Journal of Neuroscience* 31 (25). 9118–9123.

Todorovic, Ana & Floris P. de Lange. 2012. Repetition suppression and expectation suppression are dissociable in time in early auditory evoked fields. *The Journal of Neuroscience* 32 (39). 13389–13395.

Tummeltshammer, Kristen Swan, Rachel Wu, David M. Sobel & Natasha Z. Kirkham. 2014. Infants track the reliability of potential informants. *Psychological Science* 25 (9). 1730–1738.

Vinyals, Oriol, Alexander Toshev, Samy Bengio & Dumitru Erhan. 2015. Show and tell: A neural image caption generator. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 3156–3164.

Vouloumanos, Athena, Marc D. Hauser, Janet F. Werker & Alia Martin. 2010. The tuning of human neonates' preference for speech. *Child Development* 81 (2). 517–527.

Vouloumanos, Athena & Janet F. Werker. 2004. Tuned to the signal: the privileged status of speech for young infants. *Developmental Science* 7 (3). 270–276.

Wacongne, Catherine, Jean-Pierre Changeux & Stanislas Dehaene. 2012. A neuronal model of predictive coding accounting for the mismatch negativity. *Journal of Neuroscience* 32 (11). 3665–3678.

Wacongne, Catherine, Etienne Labyt, Virginie van Wassenhove, Tristan Bekinschtein, Lionel Naccache & Stanislas Dehaene. 2011. Evidence for a hierarchy of predictions and prediction errors in human cortex. *Proceedings of the National Academy of Sciences* 108 (51). 20754–20759.

Wicha, Nicole Y. Y., Elizabeth A. Bates, Eva M. Moreno & Marta Kutas. 2003. Potato not Pope: Human brain potentials to gender expectation and agreement in Spanish spoken sentences. *Neuroscience Letters* 346 (3). 165–168.

Wicha, Nicole Y. Y., Eva M. Moreno & Marta Kutas. 2003. Expecting gender: An event-related brain potential study on the role of grammatical gender in comprehending a line drawing within a written sentence in spanish. *Cortex* 39 (3). 483–508.

Wicha, Nicole Y. Y., Eva M. Moreno & Marta Kutas. 2004. Anticipating words and their gender: An event-related brain potential study of semantic integration, gender expectancy, and gender agreement in Spanish sentence reading. *Cognitive Neuroscience* 16 (7). 1272–1288.

Willits, Jon A. 2013. Learning nonadjacent dependencies in thought, language, and action: Not so hard after all … *Proceedings of the 31st Annual Conference of the Cognitive Science Society*, 2570–2575. Austin, TX: Cognitive Science Society.

Wu, Rachel, Alison Gopnik, Daniel C Richardson & Natasha Z Kirkham. 2011. Infants learn about objects from statistics and people. *Developmental Psychology* 47 (5). 1220–1229.

Wynn, Karen. 1992. Addition and subtraction by human infants. *Nature* 358 (6389). 749–750.

Ylinen, Sari, Alexis Bosseler, Katja Junttila & Minna Huotilainen. 2017. Predictive coding accelerates word recognition and learning in the early stages of language development. *Developmental Science* 20 (6). e12472.