What do we mean by prediction in language comprehension?

Gina R. Kuperberg & T. Florian Jaeger

To cite this article: Gina R. Kuperberg & T. Florian Jaeger (2015): What do we mean by prediction in language comprehension?, Language, Cognition and Neuroscience, DOI: 10.1080/23273798.2015.1102299

To link to this article: http://dx.doi.org/10.1080/23273798.2015.1102299

Published online: 13 Nov 2015.

Article views: 91

View related articles

View Crossmark data
What do we mean by prediction in language comprehension?

Gina R. Kuperberga,b and T. Florian Jaegerc,d,e

aDepartment of Psychology and Center for Cognitive Science, Tufts University, Medford, MA, USA; bDepartment of Psychiatry and the AthinoulA. Martinos Center for Biomedical Imaging, Massachusetts General Hospital, Harvard Medical School, Charlestown, MA, USA; cDepartment of Brain and Cognitive Sciences, University of Rochester, Rochester, MA, USA; dDepartment of Computer Science, University of Rochester, Rochester, MA, USA; eDepartment of Linguistics, University of Rochester, Rochester, MA, USA

ABSTRACT
We consider several key aspects of prediction in language comprehension: its computational nature, the representational level(s) at which we predict, whether we use higher-level representations to predictively pre-activate lower level representations, and whether we “commit” in any way to our predictions, beyond pre-activation. We argue that the bulk of behaviour and neural evidence suggests that we predict probabilistically and at multiple levels and grains of representation. We also argue that we can, in principle, use higher-level inferences to predictively pre-activate information at multiple lower representational levels. We suggest that the degree and level of predictive pre-activation might be a function of its expected utility, which, in turn, may depend on comprehenders’ goals and their estimates of the relative reliability of their prior knowledge and the bottom-up input. Finally, we argue that all these properties of language understanding can be naturally explained and productively explored within a multi-representational hierarchical actively generative architecture whose goal is to infer the message intended by the producer, and in which predictions play a crucial role in explaining the bottom-up input.

Language comprehension is predictive. To some, this is a controversial statement. However, under some interpretations, this is something that the field has known for several decades. To consider a well-known and broadly accepted piece of evidence, consider the phenomenon of garden-pathing during sentence comprehension. In sentences like (1a), the comprehender encounters a temporarily ambiguous sequence of words – a context. Upon encountering new bottom-up input (e.g. “conducted” … in (1b)), this ambiguity is resolved to the a priori less frequent syntactic interpretation (or parse), leading to processing difficulty. This increase in processing difficulty is known as the garden path effect, and it manifests both as relatively slower per-word reading times (Ferreira & Clifton, 1986; Garney, Pearlmutter, Myers, & Lotocky, 1997; MacDonald, Just, & Carpenter, 1992; Spivey-Knowlton, Trueswell, & Tanenhaus, 1993) and poorer comprehension accuracy (Ferreira, Christianson, & Hollingworth, 2001; Ferreira & Patson, 2007). If, however, the comprehender had encountered another context such as (1c), which avoided the temporary ambiguity, she would not have experienced a garden path effect. Importantly, as we will discuss further in the next section, the magnitude of the garden path effect is graded and highly dependent on the predictability of the intended parse given the preceding context.

(1a) The experienced soldiers warned about the dangers …
(1b) … conducted the midnight raid.
(1c) The experienced soldiers who were warned about the dangers …

Similar effects of contextual predictability are known to influence lexico-semantic processing. Reaction times are faster to predictable versus unpredictable words in a variety of behavioural tasks, ranging from lexical or phrasal decision (Arnon & Snider, 2010; Fischler & Bloom, 1979; Forster, 1981; Schwanenflugel & Lacount, 1988; Schwanenflugel & Shoben, 1985; Stanovich & West, 1983), naming (Forster, 1981; McClelland & O’Regan, 1981; Stanovich & West, 1979, 1981, 1983; Traxler & Foss, 2000), gating (Grosjean, 1980), and speech monitoring (Cole & Perfetti, 1980; Marslen-Wilson, Brown, & Tyler, 1988). Moreover, eye-tracking studies show that readers fixate less on predictable than unpredictable words (Balota, Pollatsek, & Rayner,
1985; Ehrlich & Rayner, 1981; Rayner, Binder, Ashby, & Pollatsek, 2001; Rayner & Well, 1996; see also Boston, Hale, Kliegl, Patil, & Vasisht, 2008; Demberg & Keller, 2008; Demberg, Keller, & Koller, 2013; Frank & Bod, 2011; McDonald & Shilcock, 2003; Smith & Levy, 2013; see Staub, 2015 for a recent review). And, as early as 1980, Kutas and Hillyard reported evidence for a reduced neural signal – the N400 event-related potential (ERP) – to semantically predictable versus unpredictable words in sentence contexts (Kutas & Hillyard, 1980; see also DeLong, Urbach, & Kutas, 2005; Kutas & Federmeier, 2011; Kutas & Hillyard, 1984).

The simple point we wish to make at this stage is that it is logically impossible to explain these effects without assuming that the context influences the state of the language processing system before the bottom-up input is observed. This is the minimal sense in which language comprehension must be predictive. And, indeed, as we will discuss in Section 1, almost all models of syntactic parsing and lexico-semantic processing posit that the comprehender has anticipated some structure or some semantic information prior to encountering new bottom-up information.

Given this logic, the role of prediction in language processing should not be so controversial. Yet, debates about its contributions have been central to psycholinguistic theory for decades, with researchers taking strong positions on both sides. Some, for example, have argued that, given the large number of possible continuations of any given context, predicting such information ahead of time would be an unnecessary waste of processing resources (e.g. Forster, 1981; see Van Petten & Luka, 2012 for discussion). Others have argued that, given the noisiness, ambiguity and speed of our linguistic input, prediction is the most efficient solution for fast, efficient, and accurate comprehension (e.g. Kleinschmidt & Jaeger, 2015).

These debates can be quite nuanced, with researchers focusing on different aspects of prediction. Some have distinguished expectation or anticipation from prediction (e.g. Van Petten & Luka, 2012); some have distinguished predictive pre-activation from predictive commitment (e.g. Lau, Holcomb, & Kuperberg, 2013). Finally, within the computational psycholinguistics literature, the term prediction has been used in yet other ways, in relation to a growing number of probabilistic models of language processing (e.g. Beijanki, Clayards, Knill, & Aslin, 2011; Demberg et al., 2013; Feldman, Griffiths, & Morgan, 2009; Hale, 2011; Jurafsky, 1996; Keller, 2003; Kleinschmidt & Jaeger, 2015; Norris & McQueen, 2008; Smith & Levy, 2013).

The end result is that prediction has come to mean quite different things to different people. Indeed, our review of the literature led us to the conclusion that different subfields and different researchers have critically different conceptions of what it means to predict during language comprehension. This has led to much confusion with researchers sometimes arguing at cross-purposes. The term prediction has become so loaded that some are hesitant to use it at all, while others seem to underestimate (Huettig & Mani, 2015) or even reject its role in language processing, despite growing evidence that in real-world communicative situations, the use of prediction to comprehend language is the norm. It has long been noted that, during natural conversation, we often seem to know when to take our turn, with virtually no gap or overlap between exchanges (Sacks, Schegloff, & Jefferson, 1974; Stivers et al., 2009). There is now compelling evidence that these fast exchanges arise because listeners are able to predict when a speaker’s conversational turn is about to end, and that such predictions are based on the lexical and syntactic content of what they have just heard (Magyari & de Ruiter, 2012; de Ruiter, Mitterer, & Enfield, 2006; see Garrod & Pickering, 2015 for recent discussion).

This review aims to help clarify some sources of confusion around the role of prediction in language comprehension. Our first goal is to lay out several orthogonal senses in which term prediction has been used in the psycholinguistic and cognitive neuroscience literatures, surveying the main debates and pointing to some relevant papers (although, because of space limitations, we do not aim to comprehensively review these literatures). Our second goal is to describe, in qualitative terms, how some of the different psycholinguistic views of prediction can be understood within a probabilistic (Bayesian) computational framework. We are not committed to the idea that language processing is strictly Bayesian. Indeed, many of the ideas that we discuss could be instantiated in many different ways at Marr’s (1982) algorithmic and implementational levels of analysis. However, we find this framework helpful in articulating, at Marr’s computational level, some potential links between psycholinguistic constructs that have been used to understand different aspects of prediction, and this growing computational literature. Our third aim is to summarise some of these insights by sketching out a multi-representational hierarchical actively generative architecture of language comprehension that can potentially explain and link several of the phenomena we discuss.

In Section 1, we consider what is meant by prediction in the minimal sense of the word, asking whether it is an all-or-nothing phenomenon, a graded phenomenon in which one upcoming possibility is considered at a time,
or a parallel graded phenomenon in which multiple upcoming possibilities are considered in parallel. In Section 2, we survey a large body of work suggesting that, at any given time, we can use multiple different types of information in a context to facilitate the processing of new inputs at multiple other levels of representation, ranging from syntactic, semantic, to phonological, orthographic and perceptual. In Section 3, we address the debates about whether such facilitation actually reflects the use of higher-level information that we have extracted from the context to predictively pre-activate information at lower levels of representation, before new bottom-up information becomes available to these lower levels. In Section 4, we consider the debates about whether we go beyond pre-activation by pre-updating information at higher levels of representation, incurring additional processing consequences when this pre-updated information is violated by new bottom-up input. Finally, in Section 5, we summarise the main computational insights gleaned from each section, and we return to the role of prediction in relation to the multi-representational hierarchical actively generative architecture of comprehension that we propose.

1. The probabilistic nature of contextual prediction

1.1. The data and the debates

As noted above, the minimal sense in which the term prediction has been used is to simply imply that context changes the state of the language processing system before new input becomes available, thereby facilitating processing of this new input. Throughout this review, we will broadly refer to the internal state that the comprehenser has inferred from the context, just ahead of encountering a new bottom-up input, as the internal representation of context. We postpone the question of whether the comprehenser can use high-level information within her internal representation of context to predictively pre-activate upcoming information at lower level(s) of representation until Section 3. Rather, at this stage, we focus on the nature of prediction itself and discuss the ways in which it has been conceptualised in the literature.

Some older views of prediction conceptualised it as a deterministic, all-or-nothing phenomenon. For example, the original explanations of the garden path phenomenon held that the parser predicted just one possible structure of the sentence – usually the “simplest” structure (which, interestingly, was often the most frequent and therefore the most likely structure, see Ferreira & Clifton, 1986; Frazier, 1978; with aspects of this idea going back to Bever, 1970). If the bottom-up input disconfirmed this predicted structure, the parser needed to back off and fully reanalyse the context in order to come up with the correct interpretation. Similar all-or-nothing assumptions were implicit in early views of lexico-semantic prediction, where prediction also entailed additional assumptions such as necessarily being strategic and attention-demanding (Becker, 1980, 1985; Forster, 1981; Neely, Keefe, & Ross, 1989; Posner & Snyder, 1975; see Kutas, DeLong, & Smith, 2011 for discussion), and they provided plenty of ammunition for arguments against prediction playing any major role in language comprehension: given the huge number of possible continuations of any given context, it seemed, why bother predicting just one candidate, only to be proved wrong? (see Jackendoff, 2002; Van Petten & Luka, 2012 for discussion).

More recent accounts view prediction as a graded and probabilistic phenomenon. This view is based on strong evidence of graded effects of context on processing. For example, the magnitude of the garden path effect depends on how much a particular verb (Garnsey et al., 1997; Hare, Tanenhaus, & McRae, 2007; Trueswell, Tanenhaus, & Kello, 1993; Wilson & Garnsey, 2009), thematic structure (MacDonald, Pearlmutter, & Seidenberg, 1994; Trueswell, Tanenhaus, & Garnsey, 1994) and/or wider discourse context (Spivey-Knowlton et al., 1993) biases against the intended syntactic parse. Similarly, it is well established that the magnitude of the N400 effect evoked by an incoming word is inversely correlated with that word’s probability in relation to its preceding context, as operationalised by its cloze probability1 (e.g. DeLong et al., 2005; Wlotko & Federmeier, 2012).

Further evidence for probabilistic prediction comes from a series of recent studies reporting a correlation between the surprisal of words and (a) their processing times (Hale, 2001; Levy, 2008) and (b) neural activity associated with processing them (Frank, Otten, Galli, & Vigliocco, 2015). Surprisal is an information theoretic measure that indexes the new Shannon information gained after encountering new input (MacKay, 2003; Shannon, 1948). It is quantified as the logarithm of the inverse of the probability of this input with respect to its context. There is now evidence that processing difficulty, as indexed by reading times, is linearly correlated with surprisal due to more (versus less) predictable parses (Boston et al., 2008; Dembreg & Keller, 2008; Frank & Bod, 2011; Hale, 2001; Levy, 2008; Linzen & Jaeger, 2015) or words (Boston et al., 2008; Dembreg & Keller, 2008; Dembreg et al., 2013; Frank & Bod, 2011; McDonald & Shillcock, 2003; Smith & Levy, 2013; see also Arnon & Snider, 2010).2 There is also recent evidence
sugest that surprisal correlates with the amplitude of the N400 to words within sentences (Frank et al., 2015; see also Rabovsky & McRae, 2014, for discussion of relationships between surprisal and the N400 to words outside sentence contexts).

The studies described above provide strong evidence that prediction is graded in nature. However, there remains some debate about whether it proceeds in a serial or parallel fashion. This debate has been most clearly articulated in the parsing literature. Serial models of parsing hold that just one upcoming structure of a sentence is predicted with a certain strength at any particular time. If the bottom-up input alters this structure, the parser reanalyses and goes on to the next possibility (van Gompel, Pickering, Pearson, & Liversedge, 2005; van Gompel, Pickering, & Traxler, 2001; Traxler, Pickering, & Clifton, 1998). In contrast, parallel models assume that the parser computes multiple syntactic parses in parallel, each with some degree of probabilistic support. This does not necessarily imply that all possible parses are searched exhaustively, but rather that multiple sufficiently probable parses are considered in parallel (cf. Crocker & Brants, 2000; Jurafsky, 1996; Lewis, 2000; see also Levy, Bicknell, Slattery, & Rayner, 2009; Traxler, 2014 for discussions of this issue). If the bottom-up input is inconsistent with these predicted parses, they are then shifted or reweighted (Crocker & Brants, 2000; Gorrell, 1987, 1989; Jurafsky, 1996; Levy, 2008; Narayanan & Jurafsky, 2002).

A similar debate has ensued in relation to semantic prediction. Some have suggested that because cloze probabilities are derived by averaging across participants and trials (see note 1), they are not reflective of what an individual comprehender predicts on any given trial. These researchers assume that the comprehender first predicts the word with the highest cloze probability (the strength of the prediction being related to this probability), and if this is disconfirmed by the bottom-up input, she turns to the word with the next highest cloze probability (Van Petten & Luka, 2012). Others, however, interpret the cloze profile as reflecting the strength/probability of parallel expectations that an individual’s brain computes on any given trial. So, for example, if a context has a cloze profile of 55% probability for word X, 25% for word Y and 20% for word Z, then all three possibilities are computed and represented with degrees of belief that correspond to these probabilities; if the bottom-up input turns out to be word Z, then there is a shifting or reweighting of these relative beliefs such that the comprehender now believes continuation Z with nearly 100% probability (DeLong et al., 2005; Wlotko & Federmeier, 2012; see also Staub, Grant, Astheimer, & Cohen, 2015).

In practice, it can often be difficult to experimentally distinguish between serial and parallel probabilistic prediction (for discussion in relation to syntactic prediction, see Gibson & Pearlmutter, 2000; Lewis, 2000; and in relation to lexico-semantic prediction, see Van Petten & Luka, 2012). However, as we discuss below, under certain assumptions, there is a mathematical relationship between surprisal and Bayesian belief updating, which is consistent with the idea that we can predictively compute multiple candidates in parallel, each with different strengths or degrees of belief.

1.2. Computational insights

In his now highly influential work, Anderson (1990) proposed a rational approach to cognition (for discussion, see Simon, 1990). The “ideal observer” and related models that have grown out of this work have had a tremendous influence on many disciplines in the cognitive sciences (see Chater & Manning, 2006; Clark, 2013; Griffiths, Chater, Kemp, Perfors, & Tenenbaum, 2010; Knill & Pouget, 2004 for reviews, and see Perfors, Tenenbaum, Griffiths, & Xu, 2011, for an excellent introductory overview). This is also true of language processing (e.g. Biejanki et al., 2011; Chater, Crocker, & Pickering, 1998; Clayards, Tanenhaus, Aslin, & Jacobs, 2008; Feldman et al., 2009; Kleinschmidt & Jaeger, 2015; Levy, 2008; Norris, 2006; Norris & McQueen, 2008; see also Crocker & Brants, 2000; Hale, 2001; Jurafsky, 1996; Narayanan & Jurafsky, 2002, for important antecedents of this work in the parsing literature).

Within this framework, the way that a rational comprehender can maximise the probability of accurately recognising new linguistic input is to use all her stored probabilistic knowledge, in combination with the preceding context, to process this input. The reason for this is that we communicate in noisy and uncertain environments – there is always uncertainty about the bottom-up input, and neural processing itself is noisy (for reviews and references, see Feldman et al., 2009; Norris, 2006; Shadlen & Newsome, 1994). However, so long as our probabilistic knowledge closely resembles the actual statistics of the linguistic input, then we should be able to use this knowledge to maximise the average probability of correct recognition (see e.g. Bicknell, Tanenhaus, & Jaeger, 2015; Kleinschmidt & Jaeger, 2015; Norris & McQueen, 2008, for discussion). Similar arguments hold for the speed of processing new inputs, although here more complex considerations hold (for relevant discussion, see Lewis, Shvartsman, & Singh, 2013; Smith & Levy, 2013), and, indeed, as noted above, there is strong evidence that the speed of processing new input depends on the probability of this input.
To illustrate the principles of how a probabilistic framework can be used to understand the incremental process of sentence comprehension, we describe a model of parsing by Levy (2008; see also Hale, 2003; Jurafsky, 1996; Linzen & Jaeger, 2015; Narayanan & Jurafsky, 2002). As in many probabilistic frameworks of cognition, a basic assumption of this model is that, at any given time, the agent's knowledge is encoded by multiple hypotheses. In this case, the parser's probabilistic hypotheses are about the syntactic structure of the sentence. These hypotheses are each held with different strengths or degrees and, in Bayesian terms, are known as beliefs. Together, these beliefs can be described as a probability distribution. The comprehender's goal is to infer the underlying latent or "hidden" higher-level cause of the observed data -- the underlying syntactic structure -- with as much certainty as possible. To achieve this goal, the parser draws upon a probabilistic grammar (in the broadest sense). Importantly, because the input unfolds linearly, word by word, this goal must be achieved in an incremental fashion. Using Bayes' rule, which acts to shift an original prior probability distribution to a new posterior probability distribution. This posterior distribution then becomes the new prior distribution for a new cycle of belief updating when the following word is encountered. In this way, the parser "homes in on" or discovers the underlying structure of the observed word sequences.

The process of shifting from a prior to a posterior probability distribution on any given cycle is called belief updating, and the degree of belief updating as the comprehender shifts from a prior to a posterior distribution is known as Bayesian surprise (Doya, Ishii, Pouget, & Rao, 2007), which is quantified as the Kullback-Leibler divergence between these two probability distributions. Bayesian surprise is therefore one way of computationally formalising prediction error -- the difference between the comprehender's predictions at a given level of representation before and after encountering new input at that level of representation. Unless the parser abandons the process, this cycle of belief updating will continue until it is fairly certain of the structure of the sentence being conveyed. Certainty is represented by the spread or entropy of the probability distribution. Thus, the parser may start out relatively uncertain of the structure of the sentence (described as a relatively flat probability distribution, with small probabilities of belief distributed over multiple possible structures). By the end of the sentence, however, the parser will tend to be more certain of the structure of a sentence (described as a more peaked probability distribution, with high probability beliefs that over this particular structure).

Conceptualising comprehension as an incremental process of belief updating (and thus probabilistic inference) helps address a potential criticism that is sometimes levied against prediction -- even graded forms of prediction: the idea that it might entail costs of suppressing predicted candidates that do not match the bottom-up input. Because all beliefs/hypotheses within a probability distribution must add up to 1, increasing belief about new bottom-up information will necessarily entail decreasing belief over any "erroneous" predictions. While this will entail Bayesian surprise (the shift in belief entailed in transitioning from the prior to the posterior distribution), so will not predicting at all (shifting from a flat high uncertainty prior distribution to a higher certainty posterior distribution).

An important contribution of Levy (2008; see also Levy, 2005) is that he showed that, under certain assumptions, there is a mathematical equivalence between Bayesian surprise and the information theoretic construct of surprisal, which, as noted above, is correlated with the processing times and neural activity to words during sentence comprehension. Given that the Bayesian formalisation assumes that we hold multiple beliefs in parallel, this equivalence therefore can also be taken to provide indirect support for parallel probabilistic prediction. It also helps explain some phenomena in the ERP literature, for example, why the amplitude of the N400 is large, not only to low probability words that violate highly constraining/predictable sentence contexts, such as "plane" following context (2), but also to low probability words that follow non-constraining contexts, such as "plane" following context (3) (Federmeier, Wlotko, De Ochoa-Dewald, & Kutas, 2007), and indeed to words encountered in isolation of any context (see Kutas & Federmeier, 2011 for a comprehensive review). In all of these cases, the probability of the incoming word is small, and there is a large shift from a prior to a posterior distribution (Bayesian surprise is large; see also Rabovsky & McRae, 2014, for related discussion).

2) The day was breezy so the boy went outside to fly a …
3) It was an ordinary day and the boy went outside and saw a …

Levy's (2008) model, and other probabilistic models of syntactic parsing, are inherently predictive because, over each cycle of belief updating, the newly computed posterior probability distribution (the new set of inferred hypotheses) becomes the prior distribution for the next cycle, just before new input is encountered. This new prior probability distribution thus corresponds to probabilistic predictions for a new sentence structure at the
beginning of the next cycle. These parsing models are also generative in the sense that an underlying syntactic structure can be conceptualised as generating words (Levy, 2008) or word sequences (Bicknell & Levy, 2010; Bicknell, Levy, & Demberg, 2009; Fine, Qian, Jaeger, & Jacobs, 2010; Kleinschmidt, Fine, & Jaeger, 2012), and the comprehender must infer this underlying structure from these observed data. On the other hand, none of these frameworks are actively generative: none of them assume that the comprehender’s hypotheses about syntactic structure are used to predictively pre-activate information at lower levels of representation – that is, change the prior distribution of belief at these lower levels, prior to encountering bottom-up input. We will consider what an actively generative computational framework of language comprehension might look like when we consider predictive pre-activation in Section 3.

2. Using different types of information within a context to facilitate processing of new inputs at multiple levels of representation

2.1. The data and the debates

As noted in Section 1, we assume that, just before encountering any new piece of bottom-up information, the comprehender has built an internal representation of context from the linguistic and non-linguistic information in the context that she has encountered thus far. We assume that this internal representation of context includes partial representations inferred from previously processed contextual input, ranging from subphonemic representations (e.g. Bicknell et al., 2015; Connine, Blasko, & Hall, 1991; Szostak & Pitt, 2013) all the way up to higher-level representations. Such higher-level representations may include partial representations of specific events, event structures,6 event sequences, general schemas (see Altmann & Mirkovic, 2009; Kuperberg, 2013; McRae & Matsuki, 2009, for reviews and discussion), as well as partial message-level representations (in the sense of Bock & Levelt, 1994; Dell & Brown, 1991).

In Section 1, we discussed the idea that the comprehender can use her representation of context to facilitate syntactic and lexical processing. Syntactic and lexical information, however, are not the only types of information that can be facilitated by context during processing. In this section, we survey the evidence that a comprehender can use information in a context to facilitate the processing of new information at multiple levels of representation, and that she can draw upon multiple different types of information within her internal representation of context to facilitate such processing. At this point, we continue to remain agnostic about whether the comprehender is actually able to use information within her internal representation of context to predictively pre-activate upcoming information at lower level(s) of representation prior to bottom-up input reaching these lower levels. We will consider this question in Section 3.

There is evidence that a comprehender can use her internal representation of context to facilitate the processing of coarse-grained semantic categories (Altmann & Kamide, 1999; Kamide, Altmann, & Haywood, 2003; Paczynski & Kuperberg, 2011, 2012) as well as finer-grained semantic properties (Altmann & Kamide, 2007; Chambers, Tanenhaus, Eberhard, Filip, & Carlson, 2002; Federmeier & Kutas, 1999; Kamide et al., 2003; Kuperberg, Paczynski, & Ditman, 2011; Matsuki et al., 2011; Metusalem et al., 2012; Paczynski & Kuperberg, 2012; Xiang & Kuperberg, 2015) of incoming words. These and other findings can be taken as evidence that we are able to predict (in the minimal sense, as defined in Section 1) the most likely structure of an upcoming event (a representation of “who does what to whom”: e.g. Altmann & Kamide, 1999; Garnsey et al., 1997; Hare, McRae, & Elman, 2003; Kamide et al., 2003; Paczynski & Kuperberg, 2011, 2012; Wilson & Garnsey, 2009), quite specific information about an upcoming event (e.g. Chambers et al., 2002; Kaiser & Trueswell, 2004; Kamide et al., 2003; Matsuki et al., 2011; Metusalem et al., 2012; Paczynski & Kuperberg, 2012), information about past or future events and states (e.g. Altmann & Kamide, 2007; Hare et al., 2003; Kuperberg et al., 2011; Hartshorne, O’Donnell & Tenenbaum, 2015; Pykkonen & Järviä, 2010; Rohde & Horton, 2014; Xiang & Kuperberg, 2015), as well as more general schema information (e.g. Paczynski & Kuperberg, 2012).

In addition, there is a large body of evidence that a comprehender can use her internal representation of context to facilitate the processing of incoming information at multiple other levels of representation. For example, contextual information can lead to facilitated processing of incoming information at the level of syntactic structure (see Section 1, and Arai & Keller, 2013; Farmer, Christiansen, & Monaghan, 2006; Garnsey et al., 1997; Gibson & Wu, 2013; Hare et al., 2003; Rohde, Levy, & Kehler, 2011; Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995; Wilson & Garnsey, 2009), phonological information (Allopenna, Magnuson, & Tanenhaus, 1998; DeLong et al., 2005), and orthographic information (DeLong et al., 2005; Dikker, Rabagliati, Farmer, & Pykkänen, 2010).

Moreover, facilitation of new incoming information can stem from multiple types of information within a given context. For example, to facilitate semantic
processing of new words, comprehenders are able to use information within a linguistic context about specific discourse connectives (Rohde & Horton, 2014; Xiang & Kuperberg, 2015), inferential causal relationships (Kuperberg et al., 2011), the selection restrictions of a verb (Altmann & Kamide, 1999; Paczynski & Kuperberg, 2012), the tense of a preceding verb (Altmann & Kamide, 2007), the combination of a specific verb and argument (Kamide et al., 2003; Matsuki et al., 2011; Metusalem et al., 2012; Paczynski & Kuperberg, 2012), pre-verbal arguments (Bornkessel-Schlesewsky & Schlesewsky, 2009; Kamide et al., 2003), specific prepositions (Chambers et al., 2002), and prosody (Kurumada, Brown, Bibyk, Pontillo, & Tanenhaus, 2014; Snedeker & Yuan, 2008). Similarly, to facilitate the processing of incoming information at the level of syntactic structure, comprehenders can use information within a verbal context about its referential discourse structure (Gibson & Wu, 2013), discourse coherence relationships (Rohde et al., 2011), thematic relationships between verbs and arguments (Garnsey et al., 1997; Wilson & Garnsey, 2009), the specific sense of a verb (Hare et al., 2003), or even their knowledge about a verb’s phonological typicality (Farmer et al., 2006). There is also evidence that syntactic information within a context can facilitate the processing of orthographic information (Dikker et al., 2010) or even low-level perceptual features (Dikker, Rabagliati, & Pylkkänen, 2009). Finally, comprehenders can pick up on non-verbal information in the context to influence the processing of a referent (e.g., Knoeferle, Crocker, Scheepers, & Pickering, 2005; Sedivy, Tanenhaus, Chambers, & Carlson, 1999; Tanenhaus et al., 1995).

Taken together, this literature supports the idea that, at any given time, a comprehender’s internal representations of context encode multiple different types of information, at different grains of representation (see also Jackendoff, 1987, pp. 112–115 for theoretical discussion). How much information is maintained at each of these different levels, and for how long, remains an open question (see e.g. Bicknell et al., 2015; Dahan, 2010), but it seems fair to assume that the maintenance of lower level information within the internal representation of context is shorter-lived than higher-level information.

This literature also highlights the fact that because language processing is highly interactive, with extensive communication across representational levels during processing (Elman, Hare, & McRae, 2004; McClelland & Rumelhart, 1981; Rumelhart & McClelland, 1982), many of these different types of information, encoded within the internal representation of context, can be used to facilitate the processing of incoming information at almost any other level of representation (see Altmann & Steedman, 1988; Crain & Steedman, 1985; Tanenhaus & Trueswell, 1995, for reviews and discussion). We next consider the computational implications of this type of interactivity for understanding the role of prediction in language comprehension.

2.2. Computational insights

In the probabilistic models of parsing we considered in Section 1, the aim of the parser was to infer the structure of the sentence that was being communicated. This structure was conceptualised as generating words or word sequences. Several other generative probabilistic models of language have attempted to model inference at different levels and types of representation. For example, phonetic categories can be understood as generating phonetic cues (Clayards et al., 2008; Feldman et al., 2009; Kleinschmidt & Jaeger, 2015; Kleinschmidt, Raizada, & Jaeger, 2015; Sondergiger & Yu, 2010; Toscano & McMuray, 2010), while semantic categories (Kemp & Tenenbaum, 2008) or topics (Griffiths, Steyvers, & Tenenbaum, 2007; Qian & Jaeger, 2011) can be understood as generating words.

One simplifying feature of all these models is that they each generate just one type of input (although see Brandl, Wrede, Joublin, & Goerick, 2008; Feldman, Griffiths, Goldwater, & Morgan, 2013; Kwiatkowski, Goldwater, Zettelmoer, & Steedman, 2012, for exceptions in the developmental literature). The ultimate goal of comprehension, however, is not to infer a syntactic structure, a phonemic category, a semantic category or a topic. Rather, it is to infer the full meaning of the input – the message (Bock, 1987; Bock & Levelt, 1994; Dell & Brown, 1991) or situation model (Johnson-Laird, 1983; Van Dijk & Kintsch, 1983; Zwaan & Radavsky, 1998) that the speaker or writer intends to communicate (Altmann & Mirkovic, 2009; Jaeger & Ferreira, 2013; Kuperberg, 2013; McClelland, St. John, & Taraban, 1989). For a comprehender to infer this message, she must draw upon multiple different types of stored information. Given this logic, any complete generative model of language comprehension (the process of language understanding) must consider message-level representations as probabilistically generating information at these multiple types and levels of representation. One way of modelling this type of architecture might be within a multi-representational hierarchical generative framework – the type of framework that been proposed as explaining other aspects of complex cognition (Clark, 2013; Friston, 2005; Hinton, 2007; see Brown & Kuperberg, in press; Farmer, Brown, & Tanenhaus, 2013; Pickering & Garrod, 2007, for perspectives on language processing).
Within such a framework, the comprehender would achieve her goal of inferring the producer’s message by incrementally updating her hypotheses about this message on the basis of each new piece of information as it becomes available. Such inference and belief updating, which we described for syntactic parsing in Section 1, would proceed at all levels of the hierarchy of linguistic representation. As discussed in Section 1, so long as the comprehender’s probabilistic knowledge at these levels of the hierarchy closely resembles the actual statistics of the linguistic input, then she should be able to use it to maximise the average probability of correctly (and perhaps more quickly) recognising incoming information at these levels of representation. This, in turn, should enable information to pass more efficiently up the hierarchy so that she can update her message-level representation of context (indeed, within some frameworks, such as predictive coding, it is only the information that is unpredicted – or “unexplained” – that is passed up from lower to higher levels of the hierarchy, see Clark, 2013; Friston, 2005). In the next section, we will extend this idea by arguing that, under some circumstances, information does not just flow up the hierarchy in a bottom-up fashion, but that it can also flow down the hierarchy, with information at higher levels being used to predictively pre-activate information at lower levels.

3. Predictive pre-activation

3.1. The data and the debates

In Section 2 we presented evidence that we can use multiple types of information in the context to facilitate processing of new inputs at multiple different representational levels. Facilitation, however, does not necessarily imply predictive pre-activation. To give a concrete example, imagine reading the context in (2) and finding that it can be used to facilitate processing at the phonological level (e.g. the consonant /k/ or the phonemes /k/, /aɪ/, and /t/). Just before encountering the incoming word, “kite”, our internal representation of context is likely to include a hypothesis, held with a high degree of belief, at an event level of representation, that the event being conveyed is <boy flies kite>. In theory, there are two possibilities for how this high-level inference/hypothesis might facilitate phonological processing of the incoming word, “kite”. The first is that we wait for the bottom-up input, “kite”, to activate its phonological representation (and its neighbours), and we then use our high-level event hypothesis to select the correct phonological representation. The second possibility is that we use our high-level event hypothesis to predictively pre-activate the phonological representation of “kite” prior to the bottom-up input reaching this lower phonological level of representation.

In this section, we discuss the debate about whether or not we can actually predictively pre-activate information at lower representational levels on the basis of information at higher levels within our internal representations of context, ahead of the bottom-up input reaching these lower levels. This debate has a long history in the language processing literature, and has been discussed with respect to the relationships between several different levels and types of representation.

In the speech recognition literature, many researchers would acknowledge that higher-level lexical information that has been activated by prior bottom-up phonetic input can be used to predictively pre-activate upcoming potential phonemes, prior to new bottom-up acoustic information arriving at the phonemic level of representation (Dahan & Magnuson, 2006; McClelland & Elman, 1986). In this literature, the main debate has been whether feedback connections from the lexical level to the phonological level can continue to influence the activation of the phonetic/phonological input that is currently being processed, such as lexical activity to fish leading to further enhancement of activity to /fl:/ (see Norris, 1994; Norris & McQueen, 2008; Norris, McQueen, & Cutler, 2000 for discussion).

In the sentence and discourse processing literatures, there has been more controversy about whether higher-level information within our internal representations of context can be used to predictively pre-activate upcoming information at lower levels of representation (see Federmeier, 2007; Kutas et al., 2011 for discussion). Early models argued for predictive pre-activation of lexical items (Morton, 1969). Later models, however, argued that a message-level representation of context influenced processing of new inputs only after lexical (Forster, 1981; Marslen-Wilson, 1987; Swinney, 1979) or more distributed (Gaskell & Marslen-Wilson, 1997, 1999) representations had been initially activated from the bottom-up input (see Fraunfelder & Tyler, 1987, for discussion). Only at this stage could this message-level representation exert its effect, acting to select the most appropriate candidates. This slightly later effect of context was said to lead to facilitated integration of the incoming word, and it distinguished these frameworks from the more fully interactive activation models from which they were originally inspired (Elman & McClelland, 1984; McClelland & Rumelhart, 1981). While constraint-based models of sentence processing generally remained agnostic with respect to the role of pre-activation in processing (MacDonald et al., 1994), there was sometimes an implicit assumption that high-level contextual influences like plausibility and coherence...
act primarily to select syntactic frames that had already been activated by the bottom-up lexical input (see Ferreira, 2003; Kuperberg, 2007 for discussion).

3.2. Predictive pre-activation versus pre-activation through priming

One theme that emerged from the lexical, sentence, and discourse processing literatures, was a distinction between pre-activation through top-down prediction, and pre-activation through priming.8 Some researchers distinguished between these processes, allowing pre-activation through priming, but not predictive pre-activation, to influence processing of new bottom-up input. Unlike predictive pre-activation, which entails the use of high-level information within the internal representation of context to pre-activate upcoming information at lower level(s) of representation, priming was assumed to stem from lingering activation due to previously processed material at lower levels of representation. The assumption was that this lingering activation would facilitate processing of upcoming information at this same lower level, through mechanisms such as spreading activation (e.g. Forster, 1981; see also Fodor, 1983).9 Priming was therefore often viewed as non-targeted (in that activation was taken to spread indiscriminately to related nodes at a single level of representation), and short-term (in that any lingering activation of material at lower levels of representation in the context was assumed to decay rapidly).

Some researchers also assumed other differences between priming and predictive pre-activation. For example, priming was often taken to be non-strategic (in that it served no purpose), automatic (in that it occurred without conscious control), and sometimes even involuntary (in that it could not be suppressed). This was again taken to be different from predictive pre-activation, which, as noted in Section 1, was originally believed to be strategic and sometimes targeted in that only one or a few highly probable candidates were predicted (Becker, 1980, 1985; Forster, 1981; Neely et al., 1989; Posner & Snyder, 1975).

A problem with interpreting this literature, however, is that not every account that appealed to priming subscribed to all of these assumptions, and exactly what distinguished pre-activation through priming from predictive pre-activation was not always made explicit. Moreover, there has sometimes been a tendency to hold on to some older assumptions about both priming and predictive pre-activation. For example, as discussed in Section 1, prediction is no longer assumed to be strategic or all-or-nothing, but rather implicit and probabilistic in nature (e.g. DeLong et al., 2005; Federmeier & Kutas, 1999), and there is also evidence that even “automatic” priming can sometimes be subject to some strategic control (e.g. Hutchison, 2007).

3.3. Arguments against predictive pre-activation

By the late 1990s, many psycholinguists were somewhat dubious that predictive pre-activation played much of a role in normal language comprehension (but see Altmann, 1999; Federmeier & Kutas, 1999; Federmeier et al., 2007; and also Tanenhaus et al., 1995, for early discussions of predictive pre-activation in the behavioural and ERP literatures). There was certainly widespread acknowledgment that high-level information within the comprehender’s internal representation of context could influence comprehension quickly and incrementally. However, most sentence processing frameworks assumed (either implicitly or explicitly) that such high-level information facilitated the processing of new lower level information only after this new lower level information had initially been activated by the bottom-up input.

There were several reasons for this scepticism. The first was an intuition that allowing predictive pre-activation to influence processing might afford our prior beliefs too much power, leading to distortions of perceptual or interpretational reality (e.g. Massaro, 1989). These initial concerns, however, may have been overblown. Within the speech recognition literature, there remain some legitimate concerns that feedback loops between lexical and phonemic representations might lead to auditory hallucinations (see Norris et al., 2000, p. 302 for discussion). However, under the current proposal, lexical inferences based on prior bottom-up input would be used to pre-activate upcoming phonemic information. Moreover, we argue that any predictive pre-activation would primarily influence perception in cases when there is relatively high uncertainty about the bottom-up input, as in, for example, the phonemic restoration effect (Warren, 1970), or, more generally, processing in the presence of high degrees of environmental noise (McGowan, 2015; Miller, Heise, & Lichten, 1951; Stilp & Kluender, 2010; Woods, Yund, Herron, & Ua Cruadhthaoich, 2010; reviewed by Davis & Johnsrude, 2007).10 Similarly, in the sentence processing literature, strong predictions, based on real-world knowledge or frequent structures, can sometimes lead to misinterpretations, particularly if they are compatible with certain features within bottom-up input – so-called good enough processing (Ferreira, 2003; for related discussion, see Kuperberg, 2007). The key point is that these phenomena are, in effect, examples of perceptual hallucinations (in the case of speech perception) or
“cognitive” hallucinations (in the case of good enough processing), and the way that they can be explained is precisely through the combination of strong predictive pre-activation and (relative) uncertainty about the bottom-up input.

A second concern that was sometimes raised about predictive pre-activation is similar to that discussed in Section 1: that it may entail costs of inhibiting or suppressing predicted candidates that are not supported by the bottom-up input. As we argued in Section 1, however, so long as prediction is based on our prior beliefs and the statistics of the input, then, within a purely rational framework of comprehension, the benefits of facilitation should, on average, outweigh the costs.

A third argument against using higher-level information in our internal representation of context to predictively pre-activate upcoming information is that doing so might be metabolically costly. Proponents of predictive pre-activation have sometimes ignored this issue, focusing on the idea that, under cost-free assumptions, it is computationally the most efficient way for the comprehender to keep up with the rapidly unfolding bottom-up input. In fact, both sides of the argument are likely to be valid, and when we turn next to computational insights, we will see how it may be possible to formalise the trade-off between the costs of predictively pre-activating lower level representation(s), and the benefits of facilitated bottom-up processing at multiple levels of representation.

A final reason why many psycholinguists in the late 1990s were reluctant to endorse predictive pre-activation was that, at the time, there was little direct evidence for it. As discussed in Section 2, behavioural and ERP studies provided evidence that higher-level information in the internal representation of context could facilitate processing of incoming information at multiple lower representational levels. However, as also noted above, it was often possible to argue that such facilitation was not actually due to predictive pre-activation at lower representational levels, but rather due to reduced integration at higher representational levels (see Federmeier, 2007; Kutas et al., 2011). This changed with a series of studies showing that, at least under some circumstances, it is possible to detect behavioural or neural activity to predicted versus unpredicted inputs before the onset of these inputs.

First, the visual world paradigm allowed for the measurement of eye movements while participants listened to (and sometimes acted upon) spoken language while viewing an array of images (for an in-depth review of these paradigms and their experimental logic, see Tanenhaus & Trueswell, 2006). If a linguistic context constrains towards the semantic, syntactic or phonological properties of an upcoming word, our eyes tend to move towards images that are related (versus unrelated), along this representational dimension, to the predicted word or referent. Importantly, these eye movements are sometimes anticipatory – detectable before the target word is spoken. There have now been numerous studies using the visual world paradigm, and together they provide strong evidence that, under certain circumstances, we are able to predictively pre-activate upcoming information at multiple representational levels, including syntactic (Arai & Keller, 2013; Kamide, 2012; Tanenhaus et al., 1995), semantic (Altmann & Kamide, 1999; Altmann & Mirkovic, 2009) and phonological (Allopenna, et al., 1998).

A second line of direct evidence for predictive pre-activation came from a series of ERP studies that reported differential modulation of neural activity prior to the onset of predicted versus unpredicted words. These studies used clever designs in which ERPs were measured to function elements that were dependent on a subsequent predicted content word (DeLong et al., 2005; Van Berkum, Brown, Zwitserlood, Kooijman, & Hagoort, 2005; Wicha, Moreno, & Kutas, 2004). For example, DeLong et al. (2005) showed that, in written contexts like (2), a smaller negativity was evoked by the article “a”, relative to the article “an”. “An” can only precede words starting with a vowel, and so it is inconsistent with the predicted noun, “kite”. This finding therefore provides strong evidence for predictive pre-activation – not only for upcoming semantic, but also for upcoming phonological and orthographic information. Other studies using similar types of designs in other languages have shown evidence for predictive pre-activation of syntactic gender (Van Berkum et al., 2005; Wicha et al., 2004), not only during reading but also in spoken language comprehension (Van Berkum et al., 2005). In addition, a recent study using magnetoencephalography (MEG) reported increased evoked activity, localising to the left middle temporal gyrus, in response to the presentation of highly predictive (versus less predictive) adjectives, which was taken to reflect lexical pre-activation (Fruchter, Linzen, Westerlund, & Marantz, 2015).

Finally, a few MEG studies have reported differential low-frequency oscillatory neural activity to contexts that are more versus less predictive for upcoming perceptual features. Unlike evoked ERP or MEG responses, which index neural activity that is time-locked to specific events (Luck, 2014), and which are therefore best suited to detecting facilitation when a new incoming stimulus appears, low-frequency oscillatory activity may be better suited for capturing top-down predictive neural
activity, prior to the onset of an incoming stimulus (for general discussion, see Arnal & Giraud, 2012; Engel & Fries, 2010; Weiss & Mueller, 2012, and for recent discussion in relation to language comprehension, see Lewis & Bastiaansen, 2015). These studies generally used simple contexts that constrained strongly (versus weakly) for the perceptual features of new inputs. They report differential oscillatory activity prior to the appearance of such inputs that localised to early visual (Dikker & Pylkkänen, 2013) and auditory (Sohoglu, Peelle, Carlyon, & Davis, 2012) cortices. They therefore provide some suggestive evidence that it is possible to predictively pre-activate incoming information, even at these low-level perceptual representations.

Together, these studies provide strong evidence that, at least under some circumstances, higher-level information within our internal representations of context can lead to the pre-activation of incoming information at multiple lower level representations. This is important because it implies that there are no hard architectural or neuroanatomical constraints on the flow of information activated by our internal representation of context on the processing of new bottom-up inputs. However, it is important to recognise that, just because we can use information in a context to pre-activate multiple types of information, this does not necessarily mean that we will do so in every situation. Indeed, as we discuss below, several factors have been shown to influence both the degree and the representational level at which incoming information is predictively pre-activated.

### 3.4. Factors influencing predictive pre-activation

The first important factor known to influence predictive pre-activation is the constraint of the context. As discussed above, DeLong et al. (2005) provided evidence that, following highly lexically constraining contexts like (2), predictive pre-activation of the semantic, phonological, and orthographic features of “kite” could modulate the ERP waveform, both before and as the critical word, “kite”, was actually presented. Importantly, these ERP effects were inversely proportional to the lexical constraint of the context, providing strong evidence that the lexical constraint of a context can influence the degree of pre-activation.

In addition to influencing the degree of pre-activation, there is also evidence that contextual constraint can influence the representational level of predictive pre-activation. Highly lexically constraining contexts can influence the very early stages of processing incoming words, suggesting that they can be used to pre-activate information at sublexical levels of representation, with evidence from ERP and MEG studies for facilitation on early ERP components (prior to the N400) that reflect phonological (Brothers, Swaab, & Traxler, 2015; Connolly & Phillips, 1994; Groppe et al., 2010), orthographic (Federmeier, Mai, & Kutas, 2005; Kim & Lai, 2012; Lau et al., 2013), or even early perceptual (Dikker & Pylkkänen, 2011) processing (see also Staub, 2015, for a recent review of early effects of lexically constraining contexts on eye movements during reading). Contexts that are less lexically constraining, however, do not appear to modulate these early ERP components, even when they facilitate semantic processing, as reflected by modulation within the N400 time window (e.g. Dikker & Pylkkänen, 2011; Paczynski & Kuperberg, 2012; see also Lau et al., 2013).

Most empirical work has focused on the effects of lexical constraint, as operationalised using cloze procedures (see note 1 in Section 1). Contexts that are lexically constraining, by definition, constrain strongly for multiple types of representation (semantic, phonological, and syntactic). It is important to recognise, however, that a context can constrain strongly for just one type of upcoming representation, leading just to facilitation of incoming information at this representational level, independently of any other. For example, a discourse context can constrain strongly for a general semantic schema (e.g. a restaurant schema), but not for a specific event or specific lexical item, in which case it can lead to facilitated semantic processing of words whose semantic features are related to this schema, as reflected by an attenuation of the N400 ERP component, even when this incoming word is lexically highly unexpected or even anomalous (e.g. Kolk, Chwilla, van Herten, & Oor, 2003; Kuperberg, 2007; Kuperberg, Sitnikova, Caplan, & Holcomb, 2003; Metusalem et al., 2012; Paczynski & Kuperberg, 2012).

A second important factor that can influence predictive pre-activation is the comprehender’s current goal. One way of experimentally examining the effect of goal is to manipulate task instructions or demands, and there is indeed evidence that task can influence whether neural (ERP) facilitation is seen to incoming words (e.g. see Chwilla, Brown, & Hagoort, 1995; Kuperberg, 2007; Paczynski & Kuperberg, 2012; Xiang & Kuperberg, 2015; see also McCarthy & Nobre, 1993). For example, in a recent ERP study, Xiang and Kuperberg (2015) showed that, with a requirement to explicitly judge discourse coherence, comprehenders were able to construct a deep situation-level representation of context and use it to access their stored knowledge of real-world event relationships to predict upcoming events, thereby facilitating semantic processing of incoming coherent words. With no such requirement,
however, no such semantic facilitation was seen, at least for some types of sentences. There is less work using the visual world paradigm that explicitly contrasts patterns of eye movements with different task instructions. However, there is at least some evidence that task demands can influence the degree to which anticipatory eye movements are seen towards a particular referent (Altmann & Kamide, 1999; Ferreira, Foucart, & Engelhardt, 2013; Sussman, 2006; see Salverda, Brown, & Tanenhaus, 2011 for discussion in relation to the visual world paradigm, and see Hayhoe & Ballard, 2005 for more general discussion).

Goals, of course, are not only influenced by the types of explicit tasks given to participants in psycholinguistic experiments; they play a critical role in everyday language comprehension (see Clark, 1992; Kuperberg, 2007; Tanenhaus & Brown-Schmidt, 2008 for discussion). As noted above, one can understand the broad goal of comprehension as being to infer the message communicated by the speaker or writer. However, a comprehender’s specific goal will depend on the particular communicative situation in which she finds herself. During everyday conversation, this will often be to discern the producer’s underlying intention as conveyed by speech acts (see Brown-Schmidt, Yoon, & Ryskin, 2015; Levinson, 2013; Yoon, Koh, & Brown-Schmidt, 2012 for discussion), and there are now several studies using the visual real-world paradigm showing that the presence or absence of anticipatory eye movements can be influenced by multiple different types of information in both the discourse and non-verbal context, which can cue comprehenders towards carrying out the particular action intended by the producer (see Salverda et al., 2011; Tanenhaus, Chambers, & Hanna, 2004; Tanenhaus & Trueswell, 2006 for discussion and reviews). For example, Chambers, Tanenhaus, and Magnuson (2004) asked participants to act on spoken instructions like “Pour the egg in the bowl over the flour”, and showed that anticipatory eye movements, which reflected participants syntactic parse of the sentence, were influenced by whether or not there were pourable liquid eggs in a bowl (versus solid, non-pourable eggs in a bowl). There is also evidence that our goals can influence comprehension during reading. For example, both the mechanisms we engage during processing, as well as our future recall, are influenced by whether we read to prepare for a test or whether we read for entertainment (van den Broek, Lorch, Linderholm, & Gustafson, 2001).

Finally, whether or not we see pre-activation at any particular representational level will likely depend on the speed at which the bottom-up input unfolds; contextual facilitation is greater when linguistic input is presented at slower than faster rates (e.g. Camblin, Ledoux, Boudewyn, Gordon, & Swaab, 2007; Wlotko & Federmeier, 2015). Moreover, the degree to which predictive pre-activation (versus bottom-up input) drives button presses during self-paced reading or eye movements during reading is known to be sensitive to the relative importance of comprehension speed versus accuracy (see Norris, 2006 for discussion), which can, in turn, be affected by external reward structures (cf. Bicknell, 2011; Bicknell & Levy, 2010; Lewis et al., 2013; see also Lewis, Howes, & Singh, 2014).

Taken together, all these factors suggest that the question we should be asking is not whether we can use higher-level information in our representation of context to predictively pre-activate upcoming information at lower levels of representation, but rather when we do so. We now consider the computational issues that may shed light on the question of when, and to what degree, we use higher-level information within our internal representation of context to pre-activate upcoming information at lower representational level(s).  

3.5. Computational insights

In computational terms, predictive pre-activation can be understood as the use of beliefs at a higher level of representation (level k) to change the prior distribution at a lower level of representation (k − 1), ahead of new bottom-up input reaching this lower level representation. So long as such predictive pre-activation is based on the comprehender’s stored probabilistic knowledge, then, on average, it will serve to reduce the degree of shift that the comprehender expects when she encounters new input at this lower level of representation: it will reduce her expected surprise at k − 1. In other words, by shifting her prior beliefs at k − 1 prior to encountering new information at k − 1, when such new information does reach k − 1, any further shift in belief (Bayesian surprise) will, on average, be less than if she had not pre-activated (shifted the prior at k − 1) at all. Information that has been pre-activated at k − 1 should therefore, on average, be supported by the new bottom-up input to k − 1, and its processing should therefore be relatively facilitated.

Note that an architecture in which inferences at higher levels of representation lead to the generation of predictions at lower level(s) by changing the prior probability belief distributions at these lower levels, is not only generative in the theoretical sense described in Sections 1 and 2; it is actively generative in the sense that, during real-time processing, information is passed down to lower levels of representation (i.e. higher-level
information is used to predictively pre-activate lower level information). This propagation of probabilistic beliefs from higher to lower level representations is said to be subserved by internal generative models (Friston, 2005; Hinton, 2007; cf. forward models in the motor literature).11

Faster recognition at lower levels of representation should enable information to pass more efficiently up the hierarchy to the highest message-level representation. Therefore, if we assume a completely rational framework, predictive pre-activation should, on average, lead to more efficient comprehension. There is, however, an important caveat to this claim: our brains do not have unbounded metabolic resources, and therefore, if we assume a completely rational framework, predictive pre-activation should, on average, serve her ultimate goal. Intuitively, it seems wasteful to predictively pre-activate information when it is not necessary to do so. For example, if the comprehender’s goal is to deeply comprehend a sentence, then she will likely use high-level representations (events and event structures) to predictively pre-activate information at the lower levels of representation (e.g. semantic and syntactic) that will enable her to achieve this goal. If, however, the comprehender’s goal is to monitor for the word “reviewer”, then she may be more likely to pre-activate information at the lower levels of representation (e.g. phonological) that will enable her to most efficiently perform this task.

One way of understanding the role of goal in relation to the type of architecture outlined above, is to conceptualise it as defining the generative model that the agent is employing at any given time, such that the goal is achieved by minimising Bayesian surprise across the whole model (see Friston et al., 2015, for a more general discussion of the relationships between utility and generative models). Extrapolating to language comprehension, achieving the goal of inferring the producer’s underlying message would entail minimising Bayesian surprise at the message-level representation, as well as at all levels of representation below it that allow the comprehender to achieve this goal.

Understanding the role of goal within this type of framework can also help explain how task can influence how much the comprehender values, for instance, speed or accuracy of recognition (for applications of this idea to reading, see Bicknell et al., 2012; Howes et al., 2009). Finally, this framework extends nicely to understanding decisions about behaviours that are predictively triggered as a function their utility. For example, it might potentially explain when anticipatory eye movements are seen based on the expected gain or utility of such eye movements (for related discussion, see Hayhoe & Ballard, 2005; for applications to reading, see Bicknell & Levy, 2012; Howes et al., 2009). More generally, this perspective suggests that a failure to observe behavioural evidence of predictive pre-activation at a particular representational level does not necessarily imply that we are not able to predictively pre-activate information at this level of representation (even when this information is, in principle, available within the preceding context). Since the utility of predictive behaviours depends on task, goal, and the statistical contingencies between stimuli, it is necessary to consider their contributions before concluding that predictive pre-activation at any given representational is not possible. Critically, evidence that we predict during naturalistic language processing tasks (Brown-Schmidt & Tanenhaus, 2008) and in everyday conversation (de Ruiter et al., 2006), suggests that the utility of predictive pre-activation is relatively high during everyday language processing.

The second (and related) way in which the resource-bound comprehender might be able to maximise the utility of her predictions and rationally allocate resources, is to estimate the reliability of her prior at any given level of representation within her generative model, and use this estimate to modulate the degree to which she updates her beliefs (for a given prior distribution and
likelihood function) at this level of representation (i.e. “weight” prediction error, for related discussion, see Feldman & Friston, 2010; Friston, 2010). Such estimations of reliability may play an important role in allowing us to flexibly adapt comprehension to the demands of a given situation. For example, during speech perception, it may allow us to quickly recognise familiar individual speakers, generalise our mechanism of processing to similar groups of speakers, accents and dialect, and adapt to novel speakers (see Kleinschmidt & Jaeger, 2015 for discussion), and, as discussed in Section 4, it may allow us to comprehend words that violate contexts that are highly lexically constraining.

Finally, this broad utility-based framework could, in theory, accommodate the metabolic costs of predictive pre-activation itself (as well as any metabolic costs of bottom-up message-passing). Such metabolic costs might, for example, be influenced by the speed at which the bottom-up linguistic input unfolds. This is because it presumably takes more energy to pre-activate upcoming information at a given level of representation before new input arrives at this level of representation, and so we are most likely to predictively pre-activate upcoming lower level information when the input unfolds at a slower rather than a faster rate. The costs of predictive pre-activation are also likely to be influenced by the speed of neural information flow, which is likely to differ between individuals, within individuals across the lifespan (e.g. Federmeier, 2007; Federmeier, Kutas, & Schul, 2010), and which is likely to be affected by different psychopathologies (see Brown & Kuperberg, in press; Kuperberg, 2007 for discussion).

In sum, by considering our predictions as having a utility, which is influenced by Bayesian surprise, our goals, as well as the metabolic costs of predictive pre-activation, it may be possible to understand when, to what degree, and at what level(s) of representation we pre-activate upcoming information at any given time, and to what degree we weight these predictions against new incoming evidence.

4. Predictive pre-updating and the consequences of prediction violation

4.1. The data and the debates

Within the psycholinguistics literature, some have argued that, even if we do use higher-level information within our internal representation of context to predictively pre-activate information at lower representational level(s), this still does not constitute true prediction; “true” prediction, these researchers might argue, goes beyond predictive pre-activation by entailing some kind of “commitment” to these pre-activated candidates, ahead of encountering or combining new bottom-up input.

Different researchers have discussed the idea of commitment in different ways. Some have distinguished between a graded pre-activation of multiple candidates, and a predictive commitment to one specific pre-activated candidate such as a single lexical item (Van Petten & Luka, 2012). Others have distinguished between a graded pre-activation of multiple candidates within long-term memory (which we have referred to here as predictive pre-activation), and some kind of commitment to using one (or more) of these candidate(s) to pre-update the internal representation of context (e.g. Kamide, 2008; Lau et al., 2013). For example, Lau et al. (2013) suggested that, after reading context (4a), just before encountering the incoming word (“kite”), the comprehender builds a partial representation of the event (<boy flies>) within working memory, which she uses to predictively pre-activate lower level representation(s) of <kite> (e.g. its semantic features and its phonological properties) within long-term memory. Pre-updating would refer to the additional step of updating her internal representation of context, within working memory, such that it now contains the pre-activated lower level information in addition to the partial event representation.

One notion that seems to be common to these views is the idea that, if such predictive commitments are violated by the bottom-up input (e.g. the word “plane” is encountered instead of “kite”), this would lead to a further increase in reaction times or additional neural activity that goes beyond what would ensue if the comprehender had not committed in this fashion. These increases in reaction time or prolonged neural activity have sometimes been conceived of as reflecting the costs or consequences of violating a strong prediction (see DeLong, Troyer, & Kutas, 2014; Federmeier, 2007; Kutas et al., 2011 for discussion).

(4a) The day was breezy so the boy went outside to fly a …
(4b) … kite
(4c) … plane
(5a) It was an ordinary day and the boy went outside and saw a …
(5b) … plane

Experimentally, the way researchers have sought evidence for additional neural or behavioural processing associated with violating strong, high-certainty predictions is to compare behavioural responses or neural activity to incoming words like (4c), which violate contexts like (4a) that constrain very strongly for a different specific lexical item (<kite>) or event (<boy flies kite>),
and incoming words like (5b) that follow non-constraining (non-predictable) contexts like (5a). Any differences in processing time or neural activity between the critical incoming words in (4c) and (5b) are taken to reflect the additional processing engaged as a result of violating a strong prediction. This difference is compared with another contrast – between (5b) and (4b). In (4b), the critical word is fully supported by the highly constraining context. Any differences in processing time or neural activity between (5b) and (4b) are taken to reflect reduced facilitation (due either to reduced pre-activation at lower level(s) of representation, or reduced integration at the higher event level of representation).

Behavioural studies using this type of logic have found mixed evidence that prediction violations (4c vs. 5b) lead to increased processing, over and above reduced predictive facilitation (5b vs. 4b) (Forster, 1981; Frisson, Rayner, & Pickering, 2005; Schwanenflugel & Lacount, 1988; Schwanenflugel & Shoben, 1985; Stanovich & West, 1981, 1983; Traxler & Foss, 2000). One reason for these mixed findings may be that not all of these studies matched the probability of critical words in (4c) and (5b).

Some evidence for additional neural processing that is specifically associated with violating highly lexically constraining contexts has, however emerged from the ERP literature. While a full analysis of this literature is outside the scope of this article (see Kuperberg, 2013; Van Petten & Luka, 2012, for reviews), we note that critical words like (4c), which violate highly lexically constraining contexts like (4a), evoke a larger anteriorly distributed late positivity than critical words like (5b). This is the case even when the critical words in these two conditions are matched on their cloze probabilities, and even when they evoke N400s of the same magnitudes (e.g. Federmeier et al., 2007). There is also evidence for additional prolonged neural processing, beyond that reflected by the N400, in association with words that violate contexts that constrain very strongly for a specific event structure - a specific interpretation of ‘who does what to whom’, without necessarily constraining for a specific lexical item or event. This additional prolonged processing manifests as another late positivity ERP component with a more posterior scalp distribution, known as the P600 (see Kuperberg, 2007, 2013 for reviews). Together, these late positivity effects provide some evidence that the brain can incur additional neural consequences when it encounters words that violate highly constraining contexts, over and above those reflected by the N400.

4.2. Computational insights

The psycholinguistic construct of pre-updating is compatible with the hierarchical, actively generative architecture discussed in the previous section. Within this architecture, pre-updating corresponds to the completion of an inference at a particular level of representation, in which the shift from prior to posterior gives rise to a very high-certainty posterior distribution with belief centred over only very few (and possibly one) high probability hypotheses. This pre-updating, in turn, leads to strong predictive pre-activation at lower levels of representation. Note that this view is somewhat different from the account of predictive pre-updating described above (e.g. Lau et al., 2013), which assumed that pre-activation preceded pre-updating (e.g. first using a partial representation of an event, <boy flies>, to predictively pre-activate lower level semantic, syntactic and/or phonological information, and only then pre-updating the internal representation of context with this pre-activated information). Within a hierarchical actively generative architecture, these stages are reversed: the comprehender is assumed to have already pre-updated her belief about the entire event that the producer is attempting to convey (<boy flies kite>) – a hypothesis that she holds with a high degree of belief (with a low degree of belief over hypotheses about other possible events). This high certainty inference or prediction is what leads her to predictively pre-activate information at lower levels of representation. (Note also that, given that the comprehender’s internal representation of context is multi-representational, as discussed in Sections 2 and 3, pre-updating is assumed not only to occur at high levels of representation, such as events or event structures, but also at other representational levels. For example, inferring a particular lexical item with a high degree of probability might correspond to pre-updating of beliefs at the lexical level of representation, leading to predictive pre-activation of upcoming phonemes).

One remaining question concerns the neural signatures associated with violations of highly constraining contexts, that is, the late positivities described above. One possibility is that these late positivities reflect computational mechanisms that go beyond simple belief updating (Bayesian surprise) at any single level of representation. They might, for example, reflect a process of adaptation (or learning), in which the comprehender updates her entire internal generative model to better reflect the broader statistical structure of the current environment (see Kuperberg, 2015, for further discussion; see also Kuperberg, 2013). On this account, after encountering “plane” (instead of “kite”) following context (4a), the comprehender might update her beliefs about the statistical contingencies between her semantic, syntactic, and phonological knowledge (for computational extensions of this type of generative
framework to adaptation during language processing, see Fine et al., 2010; Kleinschmidt et al., 2012; Kleinschmidt & Jaeger, 2015).

A second possibility, which is slightly different although related to the first, is that the late positivities reflect a type of “model switching”. For example, the comprehender might have previously learned (and stored) different generative models that correspond different statistical environments (Kleinschmidt & Jaeger, 2015, pp. 180–181; for related models beyond language processing, see also Gershman & Niv, 2012; Qian, Jaeger, & Aslin, 2012). For example, comprehenders might have learned generative models for particular genres (Fine, Jaeger, Farmer, & Qian, 2013; Kuperberg, 2013), dialects (Fraundorf & Jaeger, 2015; Niedzielski, 2019), or accents (Hanulikova, van Alphen, van Goch, & Weber, 2012). They might even have learned generative models for particular languages (cf. Jaeger, 2010, p. 53) or when talking to someone believed to have a language deficit (Arnold, Kam, & Tanenhaus, 2007). The late positivities might then reflect a re-allocation of resources associated with inferring (or switching to) these new generative models (for further discussion, see Kuperberg, 2015). Exploring these possibilities will be an important step in fleshing out the generative architecture described here.

5. Towards a hierarchical multi-representational generative framework of language comprehension

In this review, we considered several ways in which prediction has been discussed in relation to language comprehension. In Section 1, we argued that, in its minimal sense, 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. In Section 2, we surveyed a large body of work suggesting that we can use multiple types of information within our representation of context to facilitate the processing of new bottom-up inputs at multiple other levels of representation, ranging from syntactic, semantic, to phonological, orthographic, and perceptual. In Section 3, we discussed evidence that, at least under some circumstances, facilitation at lower level representations results from the use of higher-level inferences to predictively pre-activate information at these lower level(s), ahead of new bottom-up information reaching these levels. We also discussed several factors known to influence the degree and representational level(s) to which we predictively pre-activate, suggesting that these factors might act by influencing the utility of predictive pre-activation. Finally, in Section 4, we suggested that, when our inferences at high-level representations are particularly certain (corresponding to the psycholinguistic construct of pre-updating), and the bottom-up turns out to be incompatible with this high-certainty inference, this will lead to additional neural processing, which might reflect adaptation.

In the psycholinguistics literature, the constructs we considered in this review have sometimes been discussed as being qualitatively different from one another. For example, using context to facilitate the processing of upcoming information has sometimes been viewed as distinct from using context to pre-activate upcoming information, and predictive pre-activation has sometimes been viewed as being distinct from pre-updating. Here, however, we have argued that these constructs may be linked by appealing to a hierarchical, dynamic, and actively generative framework of language comprehension, in which the comprehender’s goal is to infer, with as much certainty as possible, the message-level interpretation or situation model that the producer intends to communicate, at a rate that allows her to keep up with the speed at which the linguistic signal unfolds.

Within this framework, this goal is achieved through incremental cycles of belief updating (Bayesian inference) at multiple levels of representation – the highest message-level representation, as well as at all the levels below that allow the comprehender to achieve her specific goal. We have also suggested that the comprehender actively propagates beliefs/predictions down to successively lower levels of representation (corresponding to predictive pre-activation). In this way, when new bottom-up input is encountered at each of these levels of representation, any Bayesian surprise will, on average, be less than if the comprehender had not predictively pre-activated at all. Finally, we have suggested that, by weighting the degree of updating by her estimates of relative reliabilities of her priors and likelihoods at any given level of representation, a comprehender with bounded resources can achieve this goal more efficiently, quickly, and flexibly. Thus, within this type of actively generative framework, prediction is not simply an “add-on” that aids the recognition of bottom-up input; it plays a pivotal role in driving higher-level inference: the goal of comprehension itself.

Of course, there is much work to be done in formalising and implementing this framework. By adopting a probabilistic approach in discussing the role of prediction in language comprehension at Marr’s computational level analysis, we are not claiming that the brain literally computes probabilities, but rather that it may be possible to describe what it is computing in probabilistic terms. In
addition, as has sometimes been pointed out, we are consciously aware of only one experience (or, in the case of language, one interpretation) at any one time (see Jackendoff, 1987, pp. 115–119, for discussion). It will therefore be important to understand how such probabilistic inference drives our (conscious) comprehension of language (for one theory in the perceptual domain, see Hohwy, Roepstorff, & Friston, 2008, and discussion by Clark, 2013, pp. 184–185).

It is also important to note that constructs such as Bayesian surprise can be instantiated in many different ways at the algorithmic and neural levels. For example, key components of incremental belief updating have been implemented within recurrent connectionist networks (e.g. Chang et al., 2006; Dell & Chang, 2014; Elman, 1990; Gaskell, 2003), where there are close links between formalisations of prediction error and Bayesian surprise (see Jaeger & Snider, 2013; McClelland, 1998, 2013 for discussion). Actively generative models have also been instantiated in some neural networks (e.g. Dayan & Hinton, 1996; Dayan, Hinton, Neal, & Zemel, 1995; Hinton, 2007, see also forward models in the motor literature, e.g. Jordan & Rumelhart, 1992). Finally, it has been proposed that this type of hierarchical actively generative architecture is instantiated at the neural level in the form of predictive coding (Friston, 2005, 2008; see Kuperberg, 2015; Lewis & Bastiaansen, 2015, for discussion in relation to the neural basis of language comprehension), although it is important to recognise that the most direct evidence for predictive coding in the brain comes from Rao and Ballard’s (1999) initial descriptions within the visual system. Given these considerations, we believe that this type of multi-representational hierarchical actively generative architecture can potentially provide a powerful bridge across the fields of computational linguistics, psycholinguistics and the neurobiology of language, and we hope that, by sketching out its principles, this will stimulate cross-disciplinary collaboration across these areas.

We conclude by taking up one more important point. In this review, we have mainly focused on the role and value of probabilistic prediction in language comprehension, generally assuming that our probabilistic predictions mirror the statistics of our linguistic and non-linguistic environments. In reality, however, during everyday communication these statistics are constantly changing: every person we converse with will have their own unique style, accent and sets of syntactic and lexical preferences. And every time we read a scientific manuscript, a sci-fi chapter, or a novel by Jane Austen, we will be exposed to quite different statistical structures in our linguistic inputs. As alluded to in Sections 3 and 4, the type of actively generative framework that we have sketched out here is, in fact, well suited for dealing with such variability in our environments. In particular, our ability to weight Bayesian surprise by our estimations of the reliability of our prior beliefs may play a more general role in allowing us to rationally allocate resources, allowing us to switch to and/or learn new generative models that allow us to optimally achieve our goals in multiple different communicative environments (for discussion in relation to phonological and speakerspecific adaptation, see Kleinschmidt & Jaeger, 2015; for discussion of other aspects of syntactic, semantic variability, and adaptation, see Fine et al., 2013, and for discussion of neural adaptation, in relation to the P600 and other late positivities in language comprehension, see Kuperberg, 2013; Kuperberg, 2015). A key goal for future research will be to understand whether the multi-representational hierarchical actively generative architecture that we have sketched out here can bridge our understanding of the relationships between language processing, adaptation, and learning (e.g. Brown-Schmidt et al., 2015; Chang et al., 2006; Dell & Chang, 2014; Jaeger & Snider, 2013).

Notes

1. To derive cloze probabilities, a group of participants are presented with a series of sentence contexts and asked to produce the most likely next word for each context. The cloze probability of a given word in a given sentence context is estimated as the proportion of times that particular word is produced over all productions (Taylor, 1953). In addition, the constraint of a context can be calculated by taking the most common completion produced by participants who saw this context, regardless of whether or not this completion matches the word that was actually presented, and tallying the number of participants who provided this completion.

2. For an alternative conceptualisation of the linking function between probabilistic belief updating and reading times, see Hale (2003, 2011). For empirical evaluation and further discussion, see Frank (2013), Linzen and Jaeger (2015), Roark, Bachrach, Cardenas, and Pallier (2009), and Wu, Bachrach, Cardenas, and Schuler (2010).

3. There are, of course, other ways of formalising prediction error, dating back to Bush and Mosteller (1951) and Rescorla and Wagner (1972). One difference between these formalisations and a Bayesian formalisation (Bayesian surprise) is that the former do not take into account uncertainty during inference or prediction (see Kruschke, 2008 for an excellent discussion). Regardless of how it is formalised, however, prediction and prediction error play a central role in both learning and processing, providing a powerful way of bridging literatures and of potentially linking across computational and algorithmic levels of analysis (see Jaeger & Snider, 2013; Kuperberg, 2015).

4. As we will discuss in Section 4, however, very low probability incoming words that mismatch the most likely continuation in a highly constraining context can evoke a
5. In this sense, the meaning of the word *generative* has some similarities with Chomsky’s original conception of a *generative* syntax, in which a grammar generated multiple possible structures (Chomsky, 1965). There is, however, an important difference: whereas generative grammars in the Chomskyan tradition served to test whether a sentence could be generated from a grammar (in which case it was accepted by that grammar), the generative computational models referred to here represent *distributions* of outputs (e.g., sentences). That is, rather than to stop at the question of whether a sentence can be generated, these models aim to capture how *likely* a sentence is to be generated (although it is worth noting that a generative syntax was formalised in probabilistic terms as early as Booth, 1969, and that probabilistic treatments of grammars have long been acknowledged in the field of sociolinguistics, see Cedergren & Sankoff, 1974; Labov, 1969 for early discussion).

6. Here, we refer to knowledge, stored at multiple grains within memory about the conceptual features that are necessary (Chomsky, 1965; Dowty, 1979; Katz & Fodor, 1963), as well as those that are most likely (McRae, Ferretti, & Amyote, 1997) to be associated with a particular semantic-thematic role of an individual event or state. This knowledge might also include the necessary and likely temporal, spatial, and causal relationships that link multiple events and states together to form sequences of events. The latter are sometimes referred to as scripts, frames, or narrative schemas (Fillmore, 2006; Schank & Abelson, 1977; Sitnikova, Holcomb, & Kuperberg, 2008; Wood & Grafman, 2003; Zwaan & Radvansky, 1998).

7. Note, however, that the term *integration* has been used in different ways in the literature. The usage described here contrasts integration with pre-activation (Federmeier, 2007; see also Van Petten & Luka, 2012, for discussion). Others, however, have used the term *integration* to refer more specifically to the process by which a word is combined or unified with its context to come up with a propositional meaning (e.g. Hagoort, Baggio, & Willems, 2009; Jackendoff, 2002; Lau, Phillips, & Poeppel, 2008).

8. The term, *priming*, is sometimes used simply to describe the phenomenon of facilitated processing of a target that is preceded by a prime, with which it shares one or more representation(s), regardless of mechanism. Pre-activation is just one of these mechanisms. For example, multiple different mechanisms have been proposed to account for the phenomena of both semantic priming (see Neely, 1991 for a review) and syntactic priming (e.g. Chang, Dell, & Bock, 2006; Jaeger & Snider, 2013; Tooley & Traxler, 2010).

9. For example, memory-based models of text processing assumed that simple lexico-semantic relationships within the internal representation of context, approximating to a “bag of words” (quantified using measures like latent semantic analysis, Kintsch, 2001; Landauer & Dumais, 1997; Landauer, Foltz, & Laham, 1998), could interact with lexico-semantic relationships stored within long-term memory, and prime upcoming lexico-semantic information through spreading activation (Kintsch, 1988; McKoon & Ratcliff, 1992; Myers & O’Brien, 1998; Sanford, 1990; Sanford & Garrod, 1998). This was known as resonance, and it can be distinguished from the use of high-level representations of events or event structures (that include information about “who does what to whom”) to predictively pre-activate upcoming semantic features or categories (see Kuperberg et al., 2011; Lau et al., 2013; Otten & Van Berkum, 2007; Paczynski & Kuperberg, 2012 for discussion).

10. There is, however, also evidence that top-down influences on the perception of lower level information is not the exception, but rather the norm, at least at the lowest levels of speech perception. For example, the internal distributional structure of phonological categories is known to affect the perception of subphonemic acoustic similarity (known as the perceptual magnet effect, Feldman et al., 2009; Kuhl, 1991). This effect has been shown to be a rational consequence of the fact that there is always uncertainty about the perceptual input (due to noise in the neural systems underlying perception). In inferring the percept, comprehenders thus rely on what they know about the statistical structure underlying the speech signal (Feldman et al., 2009; see also Haefner, Berkes, & Fiser, 2014, for a discussion of how sampling-based top-down pre-activation can explain otherwise surprising correlations in firing rates in neural populations).


12. Hierarchical predictive coding in the brain takes the principles of the hierarchical generative framework to an extreme by proposing that the flow of bottom-up information from primary sensory cortices to higher level association cortices constitutes only the prediction error, that is, only information that has not already been “explained away” by predictions that have propagated down from higher level cortices (see Clark, 2013; Friston, 2005, 2008; Wacongne et al., 2011).

**Acknowledgements**

We thank Meredith Brown, Ralf Haefner, David Kleinschmidt, Rajeev Raizada, Michael Tanenhaus, and Eddie Wlotko for extended and very helpful discussions reflected in this paper. We also thank Meredith Brown, Vera Demberg, JP de Ruiter, Kara Federmeier, Karl Friston, Ray Jackendoff, Tal Linzen, and our two anonymous reviewers for their excellent feedback on the manuscript. All errors remain the authors. We are also very grateful to Arim Choi Perrachione for all her help with manuscript preparation.


**Disclosure statement**

No potential conflict of interest was reported by the authors.
Funding
This work was partially funded by NIMH [R01 MH071635] and NICHD [R01 HD082527] grants to G. R. K., as well as by NICHD [R01 HD075797] and an NSF CAREER grant [IIS 1150028] to T. F. J.

References


Federmeier, K. D., Kutas, M., & Schul, R. (2010). Age-related and...

Federmeier, K. D., & Kutas, M. (1999). A rose by any other name:...

Feldman, N. H., Grif

Federmeier, K. D. (2007). Thinking ahead: The role and roots of...


Elman, J. L., & McClelland, J. L. (1984). Speech perception as a...

Subsequent studies revealed that...


G. R. KUPERBERG AND T. F. JAEGE


