The N400 in silico: A Review of Computational Models

Samer Nour Eddine[1], Trevor Brothers[1], Gina R. Kuperberg [1,2]

[1]Department of Psychology and Center for Cognitive Science, Tufts University,

[2]Department of Psychiatry and the Athinoula A. Martinos Center for Biomedical Imaging,

Massachusetts General Hospital, Harvard Medical School

Word count: 34,093

Abstract

The N400 event-related brain potential is elicited by each word in a sentence and offers an important window into the mechanisms of real-time language comprehension. Since the 1980s, studies investigating the N400 have expanded our understanding of how bottom-up linguistic inputs interact with top-down contextual constraints. More recently, a growing body of computational modeling research has aimed to formalize theoretical accounts of the N400 to better understand the neural and functional basis of this component. Here, we provide a comprehensive review of this literature. We discuss "word-level" models that focus on the N400's sensitivity to lexical factors and simple priming manipulations, as well as more recent sentence-level models that explain its sensitivity to broader context. We discuss each model's insights and limitations in relation to a set of cognitive and biological constraints that have informed our understanding of language comprehension and the N400 over the past few decades. We then review a novel computational model of the N400 that is based on the principles of *predictive coding*, which can accurately simulate both word-level and sentence-level phenomena. In this predictive coding account, the N400 is conceptualized as the magnitude of lexico-semantic prediction error produced by incoming words during the process of inferring their meaning. Finally, we highlight important directions for future research, including a discussion of how these computational models can be expanded to explain language-related ERP effects outside the N400 time window, and variation in N400 modulation across different populations.

Keywords: predictive coding; neural network; language comprehension; semantic; event

Section 1: General Introduction

Whether we are listening to a podcast, reading a novel, or skimming a scientific article, each word that we encounter carries a unique meaning that needs to be activated in semantic memory. We also know from a large body of research using event-related potentials (ERPs) that each of these incoming words triggers a characteristic neural response – the N400, which has provided language researchers with an important window into how the brain processes language. In this review, we explore several computational models of the N400, and their contributions in characterizing this important neural phenomenon.

During language comprehension, the N400 can be detected at the scalp surface as a negative-going waveform with a central-posterior scalp distribution, observed between 300-500ms following word onset(Kutas & Hillyard, 1980; Kutas & Hillyard, 1984). This component is often interpreted as a basic component of semantic processing, as it is known to be elicited by any meaningful stimulus, including not only words, but also images, videos, and environmental sounds (see Kutas & Federmeier, 2011 for a review).

Historically, the N400 was first described by Kutas and Hillyard (1980), who showed that semantically anomalous sentence completions produce a larger N400 than congruous completions (He spread the warm bread with *<u>*socks*</u> vs. <u>*butter*</u>). However, later studies of sentence comprehension showed that large N400s are also evoked by plausible words, as long as they are unexpected in relation to their preceding contexts. Indeed, one of the strongest predictors of the N400 is a word's contextual predictability, with increasing levels of predictability leading to graded reductions in N400 amplitude (Kutas & Hillyard, 1984; DeLong, Urbach & Kutas, 1995).

Although this component was first described in sentence contexts, it soon became clear that a robust N400 is also evoked by words with no prior context, and that the magnitude of the N400 produced by words in isolation varies systematically as a function of various lexical characteristics. For example, words trigger larger N400 responses when they are less frequent (*crypt* > *house*; Rugg, 1990; Van Petten & Kutas, 1990), when they carry a larger number of semantic features (*house* > *thing*; e.g. Kounios & Holcomb, 1994; Holcomb, Kounios, Anderson & West, 1999; Amsel, 2011; Rabovsky, Sommer & Abdel Rahman, 2012), and when they have more overlapping orthographic neighbors (*core* > *kiwi*; Holcomb, Grainger & O'Rouke, 2002; Laszlo & Federmeier, 2007; Laszlo & Federmeier, 2011; Laszlo & Federmeier, 2014). In addition, the N400 is also modulated in simple priming tasks: Repeated target words (e.g., *nurse-nurse;* Rugg, 1985; Misra & Holcomb, 2003) and semantically related target words (e.g., *doctor – nurse,* Bentin, McCarthy &Wood, 1985; Rugg, 1985; Holcomb, 1988; Holcomb & Neville, 1990) produce smaller N400 responses than target words that are unrelated to their preceding primes (e.g., *taco – nurse*).

There has been much discussion about the functional role of the N400. Some researchers have focused on its sensitivity to lexical factors, arguing that the N400 amplitude reflects the ease of "accessing" (or "retrieving") a word's lexico-semantic features. Specifically, these researchers have argued that this access or retrieval process should be "easier" for words with higher frequency (e.g., Rugg, 1990; Van Petten & Kutas, 1990), and when fewer semantic features need to be retrieved (e.g., Lee and Federmeier, 2008). Other researchers have focused on the sensitivity of the N400 to sentence-level context, arguing that this component reflects a process of "integrating" or "unifying" an incoming word into its preceding context (e.g. Hagoort,

Baggio, & Willems, 2009), which occurs only after an initial state of lexical access/retrieval is complete.

Over time, however, the field of psycholinguistics has moved away from this strict, serialized conception of lexical access *versus* integration, and towards a more interactive account, in which these two processes proceed in parallel, with continuous interactions between lower-level lexico-semantic and higher-level event representations (e.g. MacDonald, Pearlmutter & Seidenberg, 1994; Kuperberg & Jaeger, 2016). Similarly, our understanding of the N400 has evolved in tandem. Perhaps the clearest expression of this interactive view comes from Kutas & Federmeier (2011), who argue that incoming words induce changes in the activation state of semantic memory, and that the amplitude of the N400 reflects the magnitude of these changes.

An important strength of Kutas and Federmeier's framework is that it captures a wide variety of empirical findings that initially appear to have little in common (e.g. the effects of orthographic neighborhood size and the effects of contextual predictability). For example, under this formulation, words with a large number of orthographic neighbors (*core*) will automatically activate overlapping items in semantic memory (*bore*, *more*, *tore*), resulting in a larger N400 response. In addition, repeated words (*nurse – nurse*) or words that are highly predictable in context will elicit a small N400 response because their lexico-semantic representations were already active in semantic memory, resulting in a smaller change in semantic activity after word onset. Because many different types of stimuli can influence the state of semantic memory, this framework is also general enough to explain why non-linguistic inputs, such as pictures or environmental sounds, also generate robust N400s.

Despite its successes, this verbal description of the N400 leaves many questions unanswered: What is the nature of the "state" of semantic memory? How do we quantify the

impact of an incoming stimulus on this state? More specifically, how does an incoming stimulus interact with semantic memory? One way of answering these questions is to develop an explicit implementation of these theoretical assumptions within a computational model. Like the human brain, computational models can process linguistic inputs in order to perform certain tasks. By probing the internal states of these models, we can determine which computations or cognitive operations most closely resemble the patterns of N400 activity produced during online language comprehension.

The goal of this review is to describe and discuss a group of computational models that have been developed to simulate the N400.[1] In Section 2, we describe a set of "word-level" models that were trained to transform letter strings (e.g. C-A-T) into word meanings. In Section 3, we describe a set of "sentence-level" models that process sequences of word inputs, and that include representations of whole events. For each of these models, we introduce the authors' underlying theoretical assumptions and the empirical scope of the model. We then describe the model's architecture (including its training procedure), and the set of N400 findings it is able to simulate. We conclude by discussing the important insights and limitations of each modeling approach. In Section 4, we provide four important cognitive and biological constraints to

[1] There has also been work that relates the N400 to other computational models, developed in the field of Natural Language Processing, in which complex neural networks are trained to predict upcoming words in large corpora of spoken or written text (e.g. Radford et al., 2019; Devlin et al., 2018). For example, it has been shown that predictability estimates from these models (Michaelov et al., 2021; Heilbron, Armeni, Schoffelen, Hagoort & de Lange, 2020), or activity within some of their layers (Lindborg & Rabovsky, 2021) can predict N400 amplitudes produced during natural language comprehension. However, the architectures of these models are generally biologically implausible, and the nature of their internal representations is quite opaque and difficult to link specific cognitive operations in the human brain. Therefore, in this review, we focus on the neural network models that were explicitly designed to provide insights into the cognitive and neurobiological mechanisms of the N400.

consider when comparing different modeling approaches, and we provide a general summary of the word-level and sentence-level models in relation to these constraints. In Section 5, we describe a novel Predictive Coding model of the N400, developed in our own lab, which satisfies these constraints, and successfully explains a wide range of N400 phenomena at *both* the word-level and sentence-level. Finally, in Section 6, we consider outstanding questions and future directions for the field.

Section 2: Word-level Models

Traditionally, models of word recognition assumed that incoming word-forms make contact with their unique lexical entries at a discrete "recognition point" (Forster, 1979; Forster, 1981). The N400, however, is a dynamic neural process that unfolds over several hundred milliseconds following word onset (300-500ms). Moreover, as discussed above, this stimulus-driven process is thought to activate semantic memory, which is typically characterized in terms of *distributed* (rather than localist) representations (cf. Hinton, McClelland & Rumelhart, 1986). In this section, we discuss connectionist models that implement these insights. We refer to these as "word-level" models because they focus on simulating the sensitivity of the N400 to various lexical-level variables, as well as its sensitivity to minimal single word contexts in priming paradigms. In this set of models, orthographic features are mapped onto distributed semantic representations. However, as we will see, these models differ in how they operationalize the N400, with one class defining it as the total magnitude of semantic activation induced by the bottom-up input, and the other defining it as the difference between the model's current semantic state and an "ideal" target state (prediction error).

Semantic Activation Model: Laszlo & Plaut, 2012

Introduction

The first computational model to simulate the N400 using a neural network was developed by Laszlo & Plaut (2012). In their formulation, Laszlo and Plaut (2012) hypothesized that the N400 tracks the total magnitude of activity produced by bottom-up orthographic input within a Semantic Output layer. The authors had two goals. The first was to determine whether the activation dynamics induced in the Semantic Output layer would reproduce the time course and morphology of the N400. Like many ERP components, the N400 first rises to a peak (at 400ms) and then falls to baseline. The authors asked whether a set of biologically-motivated architectural constraints would produce the same rise-and-fall pattern of activation within the model's semantic units.

Laszlo & Plaut's second goal was to determine whether their semantic activation metric (mean semantic activation) was able to simulate the N400's sensitivity to differences in orthographic neighborhood size. Empirically, it has been shown that the amplitude of the N400 is larger to words like "*core*", which have many overlapping orthographic neighbors (e.g. *more*, *bore*, *care*) than to words like "*kiwi"* with fewer neighbors (Holcomb, Grainger & O'Rouke, 2002; Laszlo & Federmeier, 2007; Laszlo & Federmeier, 2011; Laszlo & Federmeier, 2014). Critically, the same N400 neighborhood effect is observed when readers process unfamiliar, non-word letter-strings (e.g. *dore* > *diwi*) (Laszlo & Federmeier, 2007). These findings suggests that semantic activation is largely obligatory, and that unfamiliar letter strings can also produce feed-forward semantic activation, even if these strings fail to map onto a pre-stored lexical representation. They can be intuitively explained within a connectionist framework in which there is continuous interaction between orthographic representations and *distributed* semantic features (e.g., Harm & Seidenberg, 2004; Rogers & McClelland, 2008). Because each

orthographic unit is linked to multiple semantic features, more semantic activity should be triggered by more densely-connected orthographic inputs. Moreover, because orthographic-to-lexical weights are shared across strings (e.g., *core*, *dore*), orthographic overlap effects should also be observed on completely novel letter sequences that were never presented during training.

Model characteristics

Laszlo and Plaut's connectionist model architecture included an Orthographic Input layer, a Semantic Output layer, and two intermediate hidden layers (one of which was an autoencoder), see Figure 1A. In the Orthographic Input layer, three-letter orthographic inputs (e.g., C-A-T) were coded as distributed bit patterns over 15 units (5 bits per letter; for example, [1 1 0 0 0] might correspond to the letter "A"). In the Semantic Output layer, distributed semantic representations were coded as sparse random binary patterns across 50 semantic units. The model was trained to activate a unique pattern of semantic units (e.g. <has-whiskers>, <animal>) in response to a given orthographic input (C-A-T) [2].

Because Laszlo & Plaut (2012) were interested in modeling the characteristic *time course* of the N400, they incorporated several neurobiologically motivated constraints on the model's architecture to influence its activation dynamics. Unlike typical connectionist models in which weights can take on any positive or negative value, units in the model were constrained to have *either* excitatory or inhibitory outgoing connections (but not both). In addition, inhibitory neurons were only allowed within-layer connections, and the number of inhibitory units was limited, relative to excitatory units (cf. Crick & Asanuma, 1986).

[2] Throughout this review, we will use italics to refer to a particular lexical item (e.g. *cat*), upper-case letters to refer to its orthographic representation (e.g. C-A-T), and angular brackets to refer to semantic features that are associated with that item (e.g. <has-whiskers>).

Following an orthographic pre-training procedure using the autoencoder, the model was trained to recognize two types of lexical items: 62 "words" with consonant-vowel-consonant sequences (e.g., *dog*), and 15 "acronyms" that always had a consonant in the central position (e.g., *dvd*). In addition, the model was trained to distinguish between these lexical items (i.e. words and acronyms) and the same set of items, but with visual distortions (in each item, a letter was distorted by flipping one bit to the wrong value). Specifically, the model learned to inhibit activation in the semantic layer whenever a visually-distorted lexical item was encountered. The modelers verified that the model was able to distinguish between the lexical items (words and acronyms) and the visually-distorted items by comparing the overall amount of semantic activation produced by these two types of input.

N400 Simulations

In all simulations, a three-letter string was clamped at the input layer, and activity was allowed to propagate forward to the Semantic Output layer. The N400 was operationalized as the mean activation produced across all units within this Semantic Output layer, and this mean value was plotted at each model iteration. The authors showed these simulated time courses shared some important similarities with the neural N400 response. Specifically, a few iterations after stimulus onset, activation within the semantic layer first rose to a peak and then fell to a stable value: a pattern that Cheyette and Plaut, (2017) later referred to as "transient semantic over-activation" (p. 2). The authors also demonstrated that these activation dynamics depended on the specific architectural constraints of the model. For example, when the model was retrained to allow connection weights of any sign (positive or negative), semantic activation no longer exhibited the characteristic rise-and-fall of the N400 response.

The authors also successfully simulated the effects of orthographic neighborhood size on the N400. In the model's trained vocabulary, words had a larger orthographic neighborhood size (6.8 neighbors) than acronyms (0.8 neighbors), and, as predicted, the words produced a much larger N400 response (*dog* > *dvd*). Similar neighborhood effects were also observed on untrained non-word stimuli, with larger simulated N400 responses produced by "pseudowords" (non-words with a word-like consonant-vowel-consonant structure, e.g. *deg*) than to illegal "letter strings" (non-words with an internal consonant, e.g. *xvd*). To explain these effects, the authors argued that both the real words and the pseudowords automatically activated the semantic features of their orthographic neighbors, resulting in greater total semantic activation within the model.

Of note, although the model was able to successfully simulate the effect of orthographic neighborhood size on both words and pseudowords, it did not reproduce another finding that has been reported in the ERP literature: the larger N400s produced by pseudowords, compared to real words (e.g. *deg* > *dog*) (Bentin, 1987), even when when orthographic neighborhood size is controlled (e.g., Holcomb, Grainger & O'Rourke, 2002; Meade, Midgley, Dijkstra & Holcomb, 2018; Meade, Grainger & Holcomb, 2019; but see Laszlo & Federmeier, 2011 who found no such effect of lexical status on the N400). Instead, the model showed increased semantic activity in the *opposite* direction, i.e. words > pseudowords at all levels of orthographic neighborhood size (see Figure 7 in Laszlo & Plaut, 2012). This finding can be explained by the model's training procedure: As noted above, the model was explicitly trained to suppress semantic activity to visually distorted inputs that did not *exactly* match the inputs received during training (words and acronyms). It is therefore not surprising that Semantic Output layer produced less

activation in response to pseudowords and illegal strings than to trained lexical inputs (words and acronyms).

Semantic Activation Model (2): Laszlo & Armstrong, 2014

Introduction

While the model by Laszlo and Plaut (2012) provided an intuitive simulation of orthographic neighborhood effects, its assumptions made it difficult to simulate another important phenomenon — the effects of priming on the N400. It is well known that repeated words (*cat – cat*) elicit a smaller amplitude N400 response than non-repeated words (*sun – cat*) (e.g. Rugg, 1985; Misra & Holcomb, 2003). However, without additional assumptions, presenting the model with the same input twice would be expected to lead to even *greater* levels of semantic activation.

Therefore, to simulate the effects of repetition priming on the N400, Laszlo and Armstrong (2014) extended the original model by introducing a *neural fatigue* mechanism (Grill-Spector, Henson & Martin, 2006). Specifically, they incorporated an exponential decay function (the alpha function), which ensured that any units with activity above a pre-determined threshold would experience time-dependent decay, pushing their activation toward zero. Laszlo and Armstrong (2014) argued that this addition was biologically plausible because the alpha function has previously been used to simulate the rise-and-fall of cortical post-synaptic potentials that produce the ERP signal (Bugmann, 1997).

N400 Simulations

To simulate repetition priming, the extended model was presented with a prime letter string, followed by a single blank "input", followed by a target that was identical or unrelated to the prime (e.g. *cat – cat; sun – cat)*. Mirroring the empirical findings, the authors found that the mean semantic activation produced by the repeated presentation of the input was smaller than that produced by its initial presentation. This was the case, regardless of whether the repeated strings were words or non-words.

Semantic Activation Model (3): Cheyette & Plaut, 2017

Introduction

In a later work, Cheyette and Plaut (2017) further extended the Semantic Activation model to simulate a wider range of lexical and priming effects on the N400 (as well as some additional behavioral effects, which are outside the scope of this review).

First, in addition to replicating the effects of orthographic neighborhood size on the simulated N400 produced by words, the authors aimed to simulate two additional lexical effects. The first was the effect of semantic richness — the larger N400 evoked by words with a larger number of semantic features (Amsel, 2011; Rabovsky, Sommer & Abdel Rahman, 2012; but see Kounios, Green, Payne, Fleck, Grondin & McRae, 2009), with more semantic associates (Laszlo & Federmeier, 2011), and with more concrete meanings (Kounios & Holcomb, 1994; Holcomb, Kounios, Anderson & West, 1999; Lee & Federmeier, 2008). Second, they simulated the effects of word frequency -- the finding that high frequency words elicit a smaller N400 than low frequency words (*nub* > *man*; Rugg, 1990; Van Petten & Kutas, 1990).

In addition to these lexical effects, Cheyette and Plaut also extended Laszlo & Armstrong (2014)'s "neural fatigue" approach (albeit using a slightly different activity-dependent decay

function) to simulate both repetition priming and semantic priming effects on the N400 (e.g. Bentin, McCarthy &Wood, 1985; Rugg, 1985; Holcomb, 1988; Holcomb & Neville, 1990). They distinguished between two types of semantic priming effects: semantic feature priming in which the prime and target share semantic features but are not necessarily associated, and semantic associative priming in which the prime and target co-occur in the same contexts, even when they share few semantic features (see Moss, Ostrin, Tyler & Marslen-Wilson, 1995 for a discussion of this distinction).

N400 Simulations

As expected, the model generally replicated the prior simulations of Laszlo & Plaut (2012) and Laszlo & Armstrong (2014): Again, mean activity at the semantic layer showed the characteristic rise-and-fall of the N400 response, and mean semantic activation showed an effect of orthographic neighborhood size (although see our discussion below regarding some discrepancies across different implementations of the Semantic Activation model).

The authors also successfully simulated two additional lexical effects. To simulate the effect of semantic richness, half the words were assigned three semantic features (low richness), and half were assigned six semantic features (high richness). Mirroring the empirical findings, words with a larger number of semantic features produced greater semantic activation.

To simulate the effects of frequency, during the initial training phase, half of the words were designated as "high frequency" and were presented to the models five times more often than their low frequency counterparts. As predicted, the model N400 was smaller to the "high frequency" words that were presented more often during training. The authors explained this frequency effect by appealing to the notion of orthographic neighborhood. Specifically, they

suggested that low frequency words are more likely to spread activation to their orthographic neighbors, thereby activating more non-target semantic units, resulting in a larger N400 response. In contrast, during training, the model would have learned to more successfully inhibit the competitors of higher frequency words, resulting in less overall semantic activation.

The decay function implemented by Cheyette and Plaut (2017) also allowed them to simulate both repetition and two types of semantic priming effects. To simulate semantic feature priming, they contrasted targets preceded by primes that shared some of the prime's semantic features with primes that shared no features. To simulate semantic associative priming, the authors exploited a feature of their training procedure. Specifically, rather than providing weight updates for each word individually, the model was presented with two words sequentially before the model's weights were updated. By assigning each word in the model's lexicon to another semantic associate, and by presenting these associates together on 30% of training trials, the model learned to activate the semantic features of the associated target while still processing the prime. During the simulation themselves, they contrasted targets preceded by associated and non-associated primes (none of these prime-target pairs shared semantic features). In all three priming simulations, the authors observed greater activity-dependent decay in response to primed versus the non-primed target words.

Finally, the authors were also able to simulate *interactions* between lexical factors and priming effects on the N400. For example, they showed that repetition effects were larger on lower (*versus* higher) frequency words, consistent with the prior literature (see Rugg, 1990; Young & Rugg, 1992), as well as on words with more (*versus* fewer) semantic features (cf. Rabovsky, Sommer, & Abdel Rahman, 2012). According to the authors, because low frequency and semantically rich words elicited more semantic activity when presented as primes, this

resulted in more neural fatigue (i.e., more activity-dependent decay) and a greater attenuation in the N400 response on repeated targets.

Semantic Attractor Model: Rabovsky and McRae, 2014

Introduction

Rabovsky and McRae (2014) took quite a different approach to modeling the N400. Instead of simulating this component as the magnitude of semantic activity induced by an incoming word (as in the three versions of the Semantic Activation model described above), they proposed that it reflected "prediction error"; that is, the *difference* between the internal semantic state of the model and the "true" semantic features associated with the input.

An important inspiration for this model was evidence for close links between language comprehension and language learning (e.g., Elman, 1990; Chang, Dell, & Bock, 2006; Dell & Chang, 2014). In supervised learning, training is accomplished by generating a pattern of neural activity in response to an input, and then calculating the *discrepancy* between this "prediction" and the "target" stimulus that is ultimately encountered. This "prediction error" is then used to modify the network's weights in order to minimize future error. Under this scheme, large prediction errors trigger a greater degree of learning, which allows the model to better predict and represent future inputs. In their simulations, Rabovsky & McRae (2014) asked whether the same prediction errors used for word learning can provide an accurate proxy for the N400 response elicited during word processing.

In contrast to previous work, Rabovsky and McRae (2014) did not aim to tackle neural realism within their model. For example, they did not limit the number or distribution of

inhibitory connections (as in Laszlo & Plaut, 2012). As we discuss later, they also did not attempt to tackle biological realism during learning: prediction errors were calculated outside the model itself. Instead, their aim was to determine whether the magnitude of the externally-computed prediction error (used to train the network) would accurately track changes in the N400 across a range of experimental conditions. Similar to Cheyette and Plaut (2017), the authors also aimed to simulate various lexical and priming effects on the N400, although, as we discuss later, this model was unable to simulate the processing of non-words.

Model characteristics

The model was a simple attractor network, with two levels of linguistic representation and no hidden layers (Cree, McNorgan & McRae, 2006), see Figure 1B. The first layer represented *word-forms* (30 units). Each word-form was represented by activating a unique combination of three units (i.e. loosely analogous to letters, e.g., D-O-G), which allowed the model to simulate different degrees of orthographic overlap. Units in the second layer represented individual semantic features (2526 units), which could be shared across words (e.g. the feature <animal> might be shared across both D-O-G and C-A-T).

The authors trained the model to map from a specific word-form to a sparse set of semantic features. These form-to-meaning weights were learned from input-output training examples (e.g., input: D-O-G; output: <animal> and <barks>) via predictive, error-driven learning (Rogers & McClelland, 2008). During each training trial, a word-form (e.g., D-O-G) was clamped at the model's input layer and the pattern of semantic of activation was allowed to settle into a stable pattern over 20 iterations. Next, the prediction error was calculated as the *cross-entropy* between this pattern of semantic activity and the correct target pattern. This error measure was then used to update the connection weights. This ensured that, when this word was

encountered in the future, the model would settle into a pattern of semantic activity that was closer to the desired target pattern.

N400 Simulations

After training, the authors performed a set of simulations in order to determine whether this same cross entropy error signal would provide an accurate index of the N400 response across different experimental conditions. Each simulation operated similarly to the training trials described above. A word-form input was presented to the model, and activity was allowed to settle in the semantic layer. This model-generated pattern of activation was interpreted as an "implicit prediction" of the correct target pattern (cf. Rogers & McClelland, 2008), and the N400 was operationalized as the cross-entropy between this implicit prediction (i.e., the current semantic activation) and the "correct" pattern of semantic activation associated with the target word. The authors plotted this prediction error signal at each iteration, beginning at stimulus onset.

For each manipulation, in addition to carrying out a simulation using cross-entropy error, the authors also carried out a simulation in which they operationalized the N400 as the total magnitude of semantic activation (across all units in the semantic layer) at each time-step. This was analogous to the approach taken in the Semantic Activation models reviewed above. In general, this semantic activation metric was much less accurate than cross-entropy error in simulating various effects on the N400: it produced the correct N400 pattern for differences in semantic richness, but, for all other simulations, it produced either a null effect or an effect in the opposite direction. We compare these two model metrics in more detail below. At this stage, however, we focus on describing the simulations using cross-entropy error.

Given the model's biologically unrealistic architecture, the time courses of these cross-entropy effects looked nothing like the actual N400 waveform (unlike the simulations reported by Laszlo & Plaut, 2012; Laszlo & Armstrong, 2014; and Cheyette & Plaut, 2017). Nonetheless, the modulation of cross-entropy error across conditions was similar to that of the empirical N400.

To simulate the effects of orthographic neighborhood size, inputs with more orthographic neighbors were compared with those with fewer neighbors (15 *vs.* 7); words from the denser orthographic neighborhoods produced a larger cross-entropy error. The authors explained this effect by pointing out that, because words with larger neighborhoods (e.g. *core*) were more likely to activate semantic features associated with their overlapping neighbors (*care*, *corn*, *cord*), these patterns of semantic activation were less likely to match the target, producing larger errors. In contrast, for low neighborhood words (e.g. *kiwi*), each orthographic unit provided a highly accurate cue for its unique set of semantic features.

Similar to Cheyette and Plaut (2017), in order to simulate the effects of frequency, the authors varied the number of training trials between high and low frequency words. As expected, they found greater cross-entropy errors for low frequency words than high frequency words. Similar to the explanation above, the authors proposed that words with fewer training trials would be less likely to produce the "correct" pattern of semantic activity, thereby producing more error.

Finally, the authors observed a robust effect of semantic richness. Again, similar to Cheyette and Plaut (2017), semantically richer words were assigned a larger number of semantic features (16 *vs.* 9). In Rabovsky and McRae's model, the semantic features were sparsely coded, and the model's default setting was to de-activate most semantic features by placing them in the

"off" setting. Therefore, successfully activating 16 of the model's 2,526 semantic units would always be more difficult and error-prone than activating only 9 semantic features.

In addition to these lexical effects, Rabovsky and McRae (2014) also successfully simulated two different types of N400 priming effects. To simulate the effects of semantic priming, the model was presented with targets that shared semantic features with a preceding prime word (*cat – dog*). The presentation of the prime led the model to "pre-activate" some of the upcoming target's features. Therefore, when the target was presented, the model generated a smaller cross-entropy error than when it was preceded by an unrelated prime.

To simulate repetition priming, the authors used a slightly different procedure, which aimed to provide a more direct link between prediction error and long-term learning. In these simulations, the model continued to update its weights *after* the presentation of each word, and cross-entropy errors were compared before and after this weight update. As expected, this additional training trial resulted in a more accurate pattern of semantic activation and lower cross-entropy error for "repeated" items.

Finally, the authors showed that the repetition priming effect (simulated as described above) interacted with two lexical variables: frequency (cf. Rugg, 1990; Young & Rugg, 1992) and semantic richness (cf. Rabovsky, Sommer, & Abdel Rahman, 2012). For high frequency words, the model already possessed an accurate mapping from form to meaning, and so the additional training trial resulted in minimal updates and a relatively small repetition benefit. In contrast, less familiar words benefited to a much greater extent from the additional training, resulting in a larger reduction in cross entropy error following the repetition. The interaction between repetition and semantic richness can be explained in a similar manner. Recall that semantically rich words activated a larger number of semantic features overall. As a

consequence, with the additional trial of training, the cumulative error improvement across multiple semantic units plausibly exceeded the cumulative error improvement over just a few semantic units. This resulted in larger repetition priming effects for semantically rich words.

Word-level models: Insights and Limitations

To sum up, we have reviewed two types of word-level model, each with distinct approaches for modeling the N400 and its sensitivity to different lexical variables and priming mainpulations. The first approach, taken by Laszlo & Plaut (2012), Laszlo & Armstrong (2014), and Cheyette & Plaut (2017) – henceforth collectively referred to as LPAC – defined the N400 as the total magnitude of semantic activation induced by the bottom-up input, in a connectionist architecture with sparse inhibition. The second approach, taken by Rabovsky & McRae (2014), framed the N400 as an implicit prediction error, where error was defined as the difference between the model's current semantic state and an "ideal" target state. Although both approaches successfully simulated a range of empirical findings, each type of model has its own unique set of strengths and weaknesses, which we discuss below.

The N400 as total semantic activation (Laszlo & Plaut, 2012; Laszlo & Armstrong, 2014, and Cheyette & Plaut, 2017)

An important strength of LPAC's Semantic Activation model is its unique temporal dynamics. By linking the N400 component to the transient over-activation of semantic units, this account successfully simulated the rise and fall of the ERP response observed at the scalp surface. This provides an important empirical benchmark for any computational account of the N400. A second strength is that this semantic activation value was naturally produced by the model itself, analogously to how the N400 is computed in the brain. As we will discuss, this was

not the case for Rabovsky and McRae's model. However, there are three major limitations to LPAC's approach.

First, from the authors' descriptions, it was not always apparent how the model's architecture influenced the pattern of semantic activation in their simulations. As noted above, the authors imposed several architectural constraints on the models, which were critical in producing the rise-and-fall pattern of the simulated N400. Importantly, however, some of these constraints also appear to have played an important role explaining the results of their N400 simulations in ways that were not completely transparent. Moreover, the application of these biological constraints varied across the different versions of the models, and several other constraints were not biologically motivated.

Consider, for example, the effect of orthographic neighborhood size, in which certain inputs (C-A-T) co-activated the semantic features of overlapping neighbors (*cap*; *car*; *hat*), resulting in enhanced semantic activity/N400s. At face value, this effect appears to be explained intuitively in any framework in which: 1) semantic activity is driven by the presence of excitatory feed-forward connections, and 2) these orthographic-to-semantic-links are shared in a distributed fashion across lexical items. However, there were clear differences in the consistency of orthographic neighborhood effects across the different versions of the Semantic Activation model, which likely depended on differences in its architecture. For example, the original model by Laszlo & Plaut (2012) limited the overall number of inhibitory units and fixed inhibitory weights to random values, resulting in a robust orthographic neighborhood effect. However, the updated architecture by Laszlo and Armstrong (2014) showed no clear neighborhood effects, with real words (*dog*) and acronyms (*dvd*) eliciting similar N400 responses, despite large differences in neighborhood size (see Figure 2 in Laszlo & Armstrong, 2014). Finally, in

Cheyette & Plaut (2017)'s model, inhibitory weights were updated during training, and the magnitude of the orthographic neighborhood effect was much smaller than that originally reported by Laszlo & Plaut (2012).

These discrepancies across the models are important because they suggest that feed-forward activation *alone* was insufficient for generating the robust orthographic neighborhood effects, and that additional architectural assumptions played a role. This issue becomes clearest when we consider the additional set of simulations carried out by Rabovsky and McRae (2014). As noted above, in addition to carrying out simulations using cross-entropy error, these authors also performed simulations in which the N400 was modeled as total magnitude of semantic activity. In these simulations, the authors found that the orthographic neighborhood effect was largely absent in their measure of global semantic activation and even *reversed* at some time-points.

Similar issues arise when we consider the lexical frequency effect, simulated by Cheyette and Plaut (2017). To explain this effect, the authors again appealed to the notion of orthographic overlap, arguing that, all else being equal, a word that appears infrequently during training spread activation to a greater number of non-target semantic units, resulting in a larger N400 response. However, this effect again appears to be specific to their particular architecture. In Rabovsky & McRae (2014), total semantic activation did *not* vary as a function of lexical frequency; if anything the words with lower frequency elicited *smaller* semantic activation during a brief processing window (see Figure 4 in Rabovsky & McRae, 2014).

In addition to these issues of transparency, some of the architectural assumptions were not biologically motivated. As one example, Laszlo & Plaut (2012) implemented only a single inhibitory unit at the semantic level, whose weights were randomly initialized and not updated

during training. Because of this constraint, this inhibitory unit was equally likely to suppress the correct set of "target" semantic features as it was to suppress incorrect semantic features (e.g. the features of orthographic competitor). With no source of selective inhibition, the model relied exclusively on excitation in order to select the correct semantic target. As a consequence, words with greater competition (from denser orthographic neighbors) would require greater activation in order to be correctly identified. Although this absence of selective inhibition was not biologically motivated, it was likely an important source of the large orthographic neighborhood effects in this model.

A related issue is that, in these models, the functional role of the semantic "clean-up" layer was not clearly specified. According to previous studies (e.g., Hinton & Shallice, 1991), clean-up layers generally implement a form of lateral competition by learning sets of semantic units that cluster together. For example, if some of the semantic features of an input (*cat*) become active (e.g., <pet> and <meow>), the clean-up layer will learn to activate this word's remaining semantic features (e.g., <soft> and <fur>) during training. Because the clean-up layer's connections are purely excitatory, this layer cannot inhibit competing semantic features that are linked to orthographic neighbors (e.g. <taxi> or <vehicle> -- features associated with the neighbor, *cab*). Instead, it relies exclusively on "driving up" activity in the semantic units of the *cat* cluster. Therefore, it is possible that this clean-up layer also contributed to the transient over-activation of semantic units, especially in the presence of co-activated neighbors.

The second major limitation of these Semantic Activation models concerns the implementation of priming effects (repetition, semantic, and associative priming). As noted earlier, modeling the N400 as total semantic activity does not intuitively explain why one would expect to see a *reduction* in total semantic activity to primed *versus* unrelated targets. Instead,

primed targets should produce *more* semantic activation with repeated presentation, and this was exactly the pattern observed by Rabovsky & McRae (2014) in their N400 simulations using global semantic activity.

In order to address this problem, Laszlo and Armstong (2014) and Cheyette and Plaut (2017) introduced their activity-dependent decay mechanism, which they motivated by appealing to "neural fatigue". However, while neural fatigue provides a biologically plausible account of *neural* suppression, it is at odds with most *cognitive* theories of priming on the N400 in the prior literature. According to these cognitive theories, reductions in the N400 response reflect *facilitated* access to a target's semantic features, which have been pre-activated during the processing of the prime; that is, it should be *easier* to retrieve the semantic features of primed than unprimed targets. However, following neural fatigue, the same semantic features that were activated by the prime would have rapidly decayed, and they would therefore be *less available* and *more difficult* to access when the target was presented.

This inherent paradox becomes most apparent when we consider how Cheyette & Plaut (2017) attempted to simulate the *behavioral* priming effect. To simulate behavioral facilitation, in addition to using the *decayed* semantic activation that was computed at each time point, and that gave rise to the simulated N400, the modelers also used (a running average) of the *undecayed* semantic activity. This, however, raises two new concerns. First, it seems implausible that, at the same time as producing neural activity that is subject to decay, neurons simultaneously track the levels of activity that they *would* have produced if they had *not* been subject to fatigue. Second, if undecayed and decayed semantic activation values were both maintained and updated at the same time – a central assumption of the model – then why would only the *undecayed* semantic

activation values contribute to the summed activation (N400) measure, as measured on the scalp surface?

The N400 as prediction error: Rabovsky & McRae (2014)

The model described by Rabovsky & McRae (2014) had its own set of strengths and weaknesses. According to this model, the N400 reflected a distinct construct – semantic prediction error – which can be dissociated from the total magnitude of semantic activation. This overcomes some inherent difficulties in reconciling overall semantic activity with lexico-semantic facilitation. For example, unlike Laszlo and Armstrong (2014) and Cheyette and Plaut (2017), there was no need to introduce a separate decay function to simulate N400 priming effects because, in this prediction error framework, there was no inherent paradox between the reduction in the simulated and actual N400 response to primed words. In addition, this model provides important insights linking the N400 response and error-driven learning. However, there were several features of the model that severely limited the plausibility of this account.

First, as the authors acknowledge, their model was unable to simulate the characteristic rise-and-fall of the N400 response after word onset. Instead, their measure of cross-entropy error was largest *prior* to word onset, and the errors monotonically *decreased* during word processing as the network settled into a pattern of activation that was closer to the correct target. In addition, while empirical N400 effects show a remarkable temporal consistency (appearing 300-500ms after word onset, see Federmeier & Laszlo, 2009), the latency of this model's lexical and priming effects was highly variable across simulations, with different effects appearing in multiple, non-overlapping time windows.

Perhaps more importantly, the proposed mechanism for calculating prediction error was neither cognitively nor biologically plausible. In this model, the current state of semantic activation was framed as an "implicit prediction", and this prediction was compared, at each time point, with an "ideal" target stimulus. However, during natural language comprehension, the brain does not have access to an ideal semantic "template" that corresponds to each incoming word. Instead, we must *infer* the correct set of semantic features *de novo* based on the bottom-up input. This point becomes especially clear when considering the processing of *non-words* (e.g. *deg*). Non-words, by definition, do not correspond to a specific semantic target, and so it is unclear, even in principle, how the brain would generate an N400 response to these inputs through this type of mechanism.

Moreover, even if the brain *did* have access to an ideal target pattern, this model does not provide a mechanism to link the prediction error either to cognitive processing or to evoked neural activity. This is because the target was not presented to the network itself, and the computation of cross-entropy error was carried out by a separate algorithm that lay outside the model. Of course, these issues of biological plausibility are common to all computational models that rely on supervised learning (see Whittington & Bogacz, 2019 for discussion). However, we emphasize them here because prediction error, and its role in learning, played a central role in this model of the N400.

Finally, we note that, although Rabovsky & McRae (2014) highlight the link between the cross-entropy error and long-term learning, the model actually used a different error measure during learning: the partial derivative of the cross-entropy error with respect to each weight. This value is, in principle, dissociable from cross-entropy error. For example, the learning signal

could be close to zero while the cross-entropy error itself is still large (e.g., in local minima or saddle points; Goodfellow, Bengio & Courville, 2016).

Despite these limitations, the success of Rabovsky & McRae (2014)'s model inspired a body of follow-up work that began to address some of these concerns, which we describe in the next section.

Section 3: Sentence-level Models

In Section 2, we reviewed two classes of computational model that simulated the N400's sensitivity to various lexical and semantic factors, including the effects of minimal contexts (priming). However, during online language comprehension, the brain must also build a higher-level *event* interpretation that is incrementally updated as each incoming word becomes available in real time. This abstract event state is thought not only to represent the *past* (based on the full sequence of words encountered thus far), but also to implicitly predict the *future* (Altmann & Kamide, 1999; Knoeferle & Crocker, 2006; Kuperberg, 2013; McRae & Matsuki, 2009).

As discussed by Elman (1990), dynamic states of this kind can be implemented in connectionist networks by incorporating recurrent elements. Recurrencies allow a model to retain a memory trace of prior inputs by providing the model's prior state as a new input on the next iteration of the model. These recurrent connections also allow the model to implicitly predict its own future state when encouraged to do so (either implicitly or explicitly) during training. In this section, we describe three "sentence-level" models that aimed to simulate the effects of broader context on the N400, all which included a recurrent element of this kind.

We also note two additional points at this stage. First, all these models conceptualized "time" somewhat differently from the word-level models described in Section 2. In all three

cases, the model received two inputs at each time point: a new incoming word and the previous state of the model. Therefore, each word input was only processed for a single model iteration. As a consequence, these models were unable to simulate the rise and fall of the N400 response when processing a single word. Second, all three models used unstructured (localist) lexical representations, which prevented them from simulating a range of lexical effects (e.g. orthographic neighborhood, words *versus* pseudoword), or the interactions between these lexical factors and higher-level context.

Retrieval-Integration Model: Brouwer, Crocker, Venhuizen & Hoeks, 2017

Introduction

Brouwer, Crocker, Venhuizen & Hoeks (2017) provided the first computational model that attempted to simulate sentence-level effects on the N400. Their recurrent model was trained to map a sequence of lexical inputs (e.g. *The meal was prepared by the cook*) on to a higher-level event representation that linked the semantics of each word to their appropriate thematic roles (e.g. Agent: <cook>, Action: <prepared>, Patient: <meal>). Unlike the word-level models described thus far, these authors conceptualized the magnitude of the N400, induced by each word, as the amount of *change* induced by the input in one of the model's internal hidden layers. A central goal of this account was to distinguish changes induced at the *lexico-semantic* level, which they linked to the N400, from changes induced at a higher event-level representation, which authors linked to a later, positive-going ERP response known as the P600.

The primary empirical benchmark for this model was the pattern of N400 and P600 modulation reported in an experiment carried out in Dutch by Hoeks, Stowe & Doedens (2004). In this study, Hoeks and colleagues manipulated the congruity between the prior context and the

sentence-final verb, as well as the sentence structure by presenting sentences in either active or passive voice. This resulted in one expected control condition (Condition 1) and three semantically anomalous conditions (see below). In two of the anomalous conditions (Conditions 2 and 4), the verb (*sung*) was not associated with words in the prior context. However, in Condition 3, despite being globally anomalous, the verb (*prepared*) was semantically associated with words in the prior context. These sentences were referred to as *thematic role reversal anomalies*.

1) *The meal was by the cook prepared*. (Literal translation: The meal was prepared by the cook.)
2) *The meal was by the cook sung*. (Literal translation: The meal was sung by the cook.)
3) *The meal has the cook prepared*. (Literal translation: The meal has prepared the cook.)
4) *The meal has the cook sung*. (Literal translation: The meal has sung the cook.)

At the sentence-final verb, Hoeks et al., (2004) observed a standard contextual congruity effect when semantically anomalous verbs shared no association with words in the prior context i.e., a larger N400 response in Conditions 2 and 4 than that evoked by the verbs in the expected condition (Condition 1), i.e. *sung* > *prepared*. However, when semantically associated role-reversal anomalies (Condition 3) were compared to the expected verbs (Condition 1), there was no modulation on the N400, i.e. *prepared* = *prepared*. Instead, these role-reversal anomalies elicited a late posteriorly-distributed positive-going component known as the P600 (for similar findings, see Kuperberg, Sitnikova, Caplan & Holcomb, 2003; Kolk, Chwilla, van Herten, & Oor, 2003; Kim & Osterhout, 2005, Experiment 1; Chow, Smith, Lau & Phillips, 2016). To explain this reduced N400 on the role reversal anomalies, Brouwer et al. (2017) argued that the anomalous verbs (*prepared*) were primed by the set of semantically associated words within the

prior context, leading to facilitated lexico-semantic access. They suggested that the role reversal anomalies instead caused difficulty at a later, post-lexical "integration" stage in which the semantic features of the verb were used to update the higher-level event state, and that this resulted in the larger P600.

Model Characteristics

The authors implemented a recurrent connectionist model to simulate these findings, see Figure 2A. They separated the model into two hierarchically-organized modules. The lower "Retrieval Module" received localist lexical inputs (e.g. *cook*), and activated their appropriate distributed semantic representations in a Semantic Output layer[3] via a hidden Semantic Retrieval layer. The Semantic Output layer encoded the 35 words in the model's lexicon as a distributed pattern of activation across 100 semantic units. According to the authors, the intermediate hidden layer – the Semantic Retrieval layer – represented the current state of semantic memory.

The higher "Integration Module" mapped from a sequence of semantic inputs (presented to the Semantic Output layer) to an Event Output layer via a hidden Integration layer. The Integration layer had recurrent connections, both to itself and to the Semantic Retrieval layer below, and carried an implicit representation of the unfolding event. The Event Output layer contained 300 units, which were divided into three "slots", with the first 100 units corresponding to the semantic features of the Agent, the second 100 to those of Action and the third 100 to those of the Patient. Thus, the event corresponding to the "The cook prepared a meal" would be

[3] We use the term "Semantic Output layer" to be consistent with the use of this term in the single word models reviewed above. Brouwer and colleagues referred to this layer as the Retrieval_Output layer. Similarly, we will use the term "Event Output layer" instead of Integration_Output layer to be consistent with the nomenclature of models discussed later in this review.

encoded by activating <cook> in the Agent slot, <prepared> in the Action slot, and <meal> in the Patient slot.[4]

To implement the conceptual distinction between the Retrieval Module and the Integration Module, the two modules were trained in separate stages. To train the Integration Module, each sentence was presented as a sequence of distributed semantic vectors (e.g. <the> <cook> <prepared> <the> <meal>), and the model learned mappings between each sequence and the appropriate event output representation (e.g. Agent: <cook>, Action: <prepared>, Patient: <meal>). Backpropagation occurred after each word. Therefore, the model learned to represent the full event and to implicitly predict upcoming semantic-thematic roles before the sentence was complete.

Next, the Retrieval Module was trained to map from a sequence of localist lexical representations (e.g. *the cook prepared the meal)* to the appropriate target event within the Event Output layer, as soon as possible in the sequence. During this second stage of training, the weights that were learned in the first stage were fixed. Therefore, learning to map lexical representations to the correct target event implicitly required that each lexical input (e.g. *cook*) mapped to the appropriate semantics (e.g. <cook>). In addition, because there were feedback connections from the hidden Integration layer to the hidden Semantic Retrieval layer, which were not fixed, the network was trained to minimize the error across *both* the "top-down" event-level information and the bottom-up lexical information simultaneously (these two "inputs" were combined into a single vector). It therefore also learned associative relationships between the

[4] Here, we use angular brackets to refer to full set of distributed semantic features associated with a particular lexical item.

semantic features of sequences of lexical inputs that constituted stereotypical events (which were presented most frequently as targets during training, as described below).

In both stages of training, the model was presented with a set of active and passive sentences with similar structures to those used by Hoeks, Stowe & Doedens, 2004. Stereotypical Agent-Action-Patient combinations were presented 50% of the time (e.g. *The meal was prepared by the cook*). In addition, the model was explicitly trained on "anomalous" Agent-Action-Patient combinations (e.g. *The meal has prepared the cook*), each of which was presented less frequently (the full set of anomalous combinations constituted the other 50% of the training set). Because the training set included equal proportions of active and passive sentences, the model could not use word order alone to assign Agent and Patient roles. In addition, each noun was equally likely to serve as an Agent or Patient.

N400 Simulations

In their simulations, the authors presented sentences to the model, word by word, similar to the procedure used during training. The authors operationalized the N400 as the cosine dissimilarity at the hidden Semantic Retrieval layer across successive time-points/word presentations (from t to t+1), with a larger cosine dissimilarity reflecting a larger state transition. This dissimilarity measure was computed externally by the modelers, and was not used for any internal computations in the model itself. Note that the activation in the Semantic Retrieval layer was computed based on both the new bottom-up input as well as top-down input from the Integration layer (induced by the previous word). In addition, the authors operationalized the later P600 as the degree of update induced by a lexical input within the hidden Integration layer — the "difficulty in integration" — again between time point t and time-point t+1.

Using this cosine dissimilarity measure at the Semantic Retrieval layer, the authors simulated N400 effects at the sentence-final verb in the four experimental conditions described by Hoeks' et al. (2004). In the expected condition (Condition 1 above), they found that the final verb (e.g. *prepared*) induced a fairly large shift at the Semantic Retrieval layer. Given that expected words are known to evoke a small amplitude N400, this relatively large shift was somewhat surprising. It also stands in contrast with the minimal shift observed at the Integration layer (see Figure 5C in Brouwer et al., 2017), which had already converged on the correct event prior to encountering the verb *prepared* (see Limitations below for further discussion).

Critically, however, when the anomalous critical verbs were not semantically associated with the set of words in the prior context (Conditions 2 and 4), they induced a shift in the Semantic Retrieval layer that was larger than the shift induced in the expected condition (*sung* > *prepared*), mirroring the results of Hoeks et al. (2004). This occurred because the model had learned to associate the semantic features of Agents and Patients and Actions around stereotypical events (*cook – food – prepared*). Therefore the associated expected verbs (*prepared*) produced a smaller shift at the Semantic Retrieval layer than non-associated anomalous verbs (e.g. *sung*).

Finally, the authors also simulated the effect of the role reversal anomalies. They found that the degree of shift induced by the sentence-final verbs in this condition (*prepared*) was the same as in the expected sentences (*prepared*), and smaller than that observed when the anomalous verbs were not semantically associated with words in the prior context (*sung*). Again, the authors attributed the attenuation of the N400 in this condition to semantic associative priming: the model had learned links between the verb with the prior set of semanticaly

associated lexical items, based on its prior experience with stereotypical events, regardless of word order.

Importantly, the authors also showed that the reduced N400 to the associated role reversal anomalies was not due to a mis-assignment of thematic roles at the final verb (a "semantic illusion"): Because the model had been trained on anomalies, even prior to verb onset, the Integration Output layer showed Agent and Patient activations that were consistent with the actual (anomalous) interpretation of the sentence (Agent: <meal>, Action: <prepared>, Patient: <cook>) (see Figure 3 in Brouwer et al., 2017). However, upon encountering the reversal anomaly, the Integration layer nonetheless produced a large shift. This was because the final verb (*prepared*) shifted the event representation from a relatively uncertain prior distribution (representing multiple possible low-probability events) to a *single* anomalous interpretation. The author's argued that this higher-level shift reflected "difficulty in integration" and that this difficulty was indexed by the large P600 response produced in this condition[5].

Insights and Limitations

The computational model proposed by Brouwer et al. (2017) represents an ambitious attempt to simulate ERPs at multiple levels of linguistic representation. The authors' inclusion of recurrent connections allowed the model to represent events beyond the time scale of individual lexical items and allowed the hidden Integration layer to compute implicit predictions of specific upcoming Agents, Actions and Patients. Because of the two-stage training procedure, the model

[5] The authors also found that the sentence-final verbs in the two non-associated anomalous conditions (Conditions 2 and 4) produced a large shift at the Integration layer (the shift from the unpredicted prior event state to the new anomalous event state). They argued that the reason why these two conditions did not produce a visible P600 at the scalp surface in the study by Hoeks et al. (2004) was because of component overlap from the earlier large N400 effect.

also learned to represent semantic associations between the words used to describe stereotypical events, allowing for some facilitation on the simulated N400 produced by associated (*versus* non-associated) words at the lower-level Semantic Retrieval layer. However, the model's N400 simulations had several limitations, which we discuss below (a discussion of the P600 is beyond the scope of this review).

The first set of problems relates to the authors' conceptualization of feedback from the Integration layer to the Semantic Retrieval layer. Although the authors describe this feedback as providing "pre-activation" that influenced lexico-semantic processing of expected words, the model does *not* actually implement top-down pre-activation as it is typically understood in most psycholinguistic frameworks. In top-down predictive processing models of language comprehension, the basic assumption is that higher levels of representation provide top-down feedback that pre-activates lower-level lexico-semantic representations (encoded at a smaller spatiotemporal scale) *before* new bottom-up input becomes available (e.g. Federmeier, 2007; Kuperberg & Jaeger 2016, section 3.5). Following pre-activation, lexico-semantic processing is facilitated if the incoming word that matches these prior top-down predictions, with highly predictable words eliciting little to no N400 activity.

However, in Brouwer et al.'s Retrieval-Integration model, there was no intermediate time-step before word onset in which the Integration layer could influence the state of the Semantic Retrieval layer. The top-down context was only operative in a brief time-window in which the bottom-up input was presented, and so the longer time-scale of the implicit event representation at the Integration layer was not exploited. This failure of predictable contexts to actually pre-activate the Semantic Retrieval layer meant that when the expected words were encountered (Condition 1), they produced a relatively large update at the Semantic Retrieval

layer as it shifted from representing the semantics of the previous word to that of the new word. This, however, is at odds with the very small amplitude N400 that is usually produced by expected words in constraining sentence contexts.

Instead of being driven by top-down predictive pre-activation, the smaller shift at the Semantic Retrieval layer to both the expected associated verbs (Condition 1) and to the role reversed anomalous verbs (Condition 3), relative to the non-associated anomalies (Conditions 2 and 4), occurred because the model had learned associative relationships between the words that constituted stereotypical events. The authors attributed this smaller shift to facilitatory effects of associative lexico-semantic "priming" on the N400. This interpretation, however, is at odds with a large body of empirical work showing that, although associative priming across individual words can lead to *some* facilitation on the N400 during sentence comprehension, these effects are relatively small (e.g. Van Petten, 1993; Camblin, Gordon, & Swaab, 2007), and they cannot account for all the effects of higher-level context. For example, in many situations, even when the lexico-semantic associations across the "bags of words" within sentences are matched across experimental conditions, the N400 is still smaller to expected than to unexpected or incongruous words (e.g. Van Petten & Kutas, 1990; Van Petten & Kutas, 1991; Nieuwland & Kuperberg, 2008; Otten & Van Berkum, 2007; Urbach, DeLong & Kutas, 2015; Xiang & Kuperberg, 2015; Kuperberg, Paczynski, & Ditman, 2011, Paczynski & Kuperberg, 2011; Paczynski & Kuperberg, 2012; see Kuperberg, 2016 and Shetreet, Alexander, Romoli, Chierchia & Kuperberg, 2019, Supplementary Materials section 3 for a recent review and discussion).

A second major issue with the model concerns its treatment of the semantically anomalous inputs. As discussed above, despite these sentences being anomalous, the model was explicitly trained with these inputs (e.g. *The meal has prepared the cook*) and so it had no

difficulty assigning anomalous Agents and Patients at the Event layer, in both reversible and non-reversible anomalous sentences. This feature of the model was essential for the authors' argument that the reduction of the N400 on the associated semantically reversible anomalies did not reflect a "semantic illusion" (a mis-assignment of thematic roles at the event level), but, instead reflected simple associative priming. However, human behavioral experiments have shown that readers *do* sometimes incorrectly assign thematic roles in non-canonical sentence structures, like passives. This is even more likely to occur when role assignment results in an implausible interpretation or when the roles are reversible (e.g., Ferreira, 2003; Gibson, Bergen & Piantadosi, 2013).

A final issue involves the model's operationalization of the N400 as a shift in state. This shift in state was computed *outside* the model by recording and comparing the state of the Semantic Retrieval layer at two time points. Moreover, because the Semantic Retrieval layer's activation values were not retained from the presentation of one word to the next, it is not clear *how* the model itself would be able to compute this dissimilarity measure without additional assumptions. In addition, it is also not altogether clear *why* the model would compute this measure, although one possibility is that the N400 arises as an epiphenomenon, resulting from process of implicitly shifting activation from a prior to a new state, based on new bottom-up input.

Sentence Gestalt Model: Rabovsky, Hansen & McClelland, 2018

Introduction

Another computational model by Rabovsky, Hanson and McClelland (RHM18) had many similarities with Brouwer et al.'s Retrieval-Integration model. Like Brouwer et al.'s model,

RHM18's model transformed a sequence of word inputs (a sentence) into an implicit event representation encoded in a hidden layer. However, unlike the Retrieval-Integration model, RHM18's architecture was not divided into separate modules that implemented semantic retrieval and integration, and, instead of simulating the N400 as a shift within a hidden Semantic Retrieval layer, RHM18 operationalized it as a shift of the implicit event representation itself (analogous to how Brouwer et al. operationalized "integration" and the P600 effect).

Like Brouwer et al. (2017), RHM18 simulated N400 effects of semantic incongruity (Kutas & Hillyard, 1980) and thematic role reversal (Kuperberg, Sitnikova, Caplan & Holcomb, 2003; Kolk, Chwilla, van Herten, & Oor, 2003; Hoeks, Stowe and Doedens, 2004; Kim & Osterhout, 2005; Chow, Smith, Lau & Phillips, 2016). However, they also tackled several other sentence-level phenomena, including the *graded* effects of congruity/predictability on the N400 (Kutas & Hillyard, 1984), the so-called "related anomaly effect", i.e. the N400 reduction to incongruous words that are semantically related to an expected completion (Federmeier & Kutas, 1999; Kutas & Hillyard, 1984), the effect of word position in sentences, i.e. the smaller N400 on content words that appear later *versus* earlier in coherent sentences (Van Petten & Kutas, 1990; Van Petten & Kutas, 1991; Payne, Lee & Federmeier, 2015), and the null effects of contextual constraint (Kutas & Hillyard, 1984; Federmeier, Wlotko, De Ochoa-Dewald & Kutas, 2007; Kuperberg, Brothers, & Wlotko, 2020) and word order violation (Hagoort & Brown, 2000) on the N400.

In addition to these sentence-level effects, the authors simulated the effects of lexical frequency and priming that had been captured by previous word-level models (Laszlo & Armstrong, 2014; Cheyette & Plaut, 2017). Finally, they carried out a simulation that began to address the question of *why* the model might track the magnitude of the shift induced by bottom-

up input (the simulated N400). Building on the insights of the earlier word-level model by Rabovsky & McRae (2014), RHM18 suggested that the magnitude of this shift might function as a "prediction error" that influenced downstream learning, analogous to how "temporal differences" are used to drive learning in reinforcement learning models (Sutton & Barto, 2018). To test this hypothesis, RHM18 simulated the interaction between long-term repetition and sentence congruity – the finding that long-term repetition effects on the N400 are larger on semantically incongruous (*versus* congruous) sentence continuations (Besson, Kutas & Van Petten, 1992).

Model Characteristics

The architecture, shown in Figure 2B, was based on earlier computational models of event comprehension, developed by McClelland and colleagues (McClelland, St. John & Taraban, 1989; St. John & McClelland, 1990). Localist lexical inputs (74 units, one unit per word) fed forward to a hidden layer (Hidden 1), which then activated another hidden layer that was trained to encode an implicit event representation (i.e. the semantic-thematic roles of the current sentence). In RHM18's model, this second hidden layer was called the Sentence Gestalt (SG) layer, and it was partly analogous to the hidden Integration layer used by Brouwer et al. (2017). Again, the SG layer provided feedback through recurrent connections to an earlier layer (in this case, the Hidden 1 layer), which allowed it to dynamically represent an event state that implicitly anticipated upcoming semantic-thematic roles. This network was called the Update Network and was used for all simulations. As in Brouwer et al.'s model, during training, the SG hidden layer interfaced with an Event Output layer, which was used to train the model and explicitly encoded the semantic features associated with an event's thematic roles (although, as

we describe below, this interface was indirect as the model was trained via a separate Query Network). This Event Output layer was separated into two partitions, with five units specifying one of five possible thematic roles (e.g. Agent) and 171 units specifying the distributed semantic features which could be associated with each role.

Unlike the Retrieval-Integration model, RHM18's Sentence Gestalt model was not hierarchically organized: the authors made no distinction between the layers that represented the semantics associated with individual words, and events. Therefore, instead of defining the N400 as a change in state at a lower hidden layer (representing lexico-semantic information), it was operationalized as the degree of change in the SG layer itself (which implicitly represented event-based information), analogous to how Brouwer et al., operationalized the P600.

These architectural assumptions were also reflected by how RHM18 trained their model. Unlike the Retrieval-Integration model, which was trained in two separate stages, the Sentence Gestalt model was trained in a single stage by mapping sequences of lexical inputs on to complete events. During training, a downstream Query Network played the role of "teacher", which allowed the hidden SG layer to learn the appropriate semantic-thematic mappings of each sentence. On a typical training trial, a sentence was presented to the model's Update Network as a sequence of localist lexical inputs (e.g. *At breakfast the man eats eggs*). As each word activated the SG layer, this activity was fed into the model's Query Network, which included a Probe layer with the same structure as the Event Output layer. After each word, the modeler presented the Query Network with a set of partial event representations via this Probe layer by leaving different partitions (semantic or thematic role) blank. Following each "query", the correct answer was then provided to the Event Output layer. Because the model was queried after each input

about semantic features or thematic roles that had not yet been presented, the SG layer was encouraged to implicitly anticipate an event's upcoming semantic-thematic roles.

N400 Simulations

As in Brouwer et al (2017), the authors simulated the N400 by presenting sentences to the model, word by word. At each time-point (t), the model combined the new lexical input with the prior state of the SG layer to compute a new, updated pattern of SG activations. The N400 was defined as the absolute value of the change in activation within the SG layer (from t *to* t+1), summed across all SG units. As for Brouwer et al. (2017), the magnitude of this update was computed externally by the modelers.

Using this measure, the authors first examined the effects of contextual congruity/predictability, going beyond Brouwer et al. (2017) by simulating *graded* effects on the N400. During training, the authors used sentences in which Agents performed Actions on stereotypical Patients. For example, in a breakfast scenario, the *eat* action always occurred with either *eggs* or *toast*, while in a dinner context, *eat* occurred with either *pizza* or *soup* (see Supplementary Figure 12 in RHM18 for details). However, in a given scenario, certain Action-Object combinations were presented more frequently to the model during training (e.g. "…*eats eggs*" appeared more often than *"...eats toast"*). Therefore, when processing the sequence "*At breakfast the man eats*…", the model's SG layer was able to represent the complete event (Agent: <man>, Action: <eat>, Patient: <egg>) and to implicitly anticipate the upcoming word (its semantic features and Patient role) before the onset of the final noun. When the model encountered an expected Patient (*eggs*), this resulted in little to no change in the internal state of the SG layer, consistent with the small N400 evoked by contextually predictable words. In

contrast, a larger SG update was triggered by lower-probability words ("…*eats toast*"), and still larger updates were triggered by completely "anomalous" words ("…*eats oak"*) that were never encountered in this context during training.

The Sentence Gestalt model also successfully simulated the smaller N400 response to role reversal incongruities *versus* unrelated incongruities (specifically, the findings reported by Kuperberg, Sitnikova, Caplan & Holcomb, 2003, and Kolk, Chwilla, van Herten & Oor, 2003). However, this was for a different reason from that discussed above in relation to Brouwer et al.'s Retrieval-Integration model. In RHM18's simulations, role-reversed verbs like "*eats*" in the sentence, "*At breakfast the egg *eats*" were treated by the model as truly anomalous because, during training, "*egg*" was *never* presented in an Agent role. Therefore, in these role-reversed sentences, the model incorrectly interpreted "*egg*" as a Patient, both before and after encountering the verb "*eats*", leading to a minimal SG update, which mirrored the empirical result. This pattern is consistent with a classic "semantic illusion" interpretation of the N400 reduction in reversal anomalies, in which comprehenders, at least temporarily, entertain a semantically-driven interpretation of the sentence, based on stereotypical thematic roles (Kuperberg, 2007; see Kuperberg, 2016 for a more recent discussion). In contrast to these role-reversed verbs, semantically anomalous verbs produced a large SG update.

The authors also simulated several other sentence-level phenomena. To simulate the effects of contextual constraint, certain actions appeared with restricted object sets during training *("eat eggs", "eat toast"),* while other actions were relatively non-constraining (e.g. *"to like").* Low-probability endings elicited a similar SG update across high-constraint *("...eats oak")* and low-constraint *("...likes oak")* sentences. In both cases, the final word *("oak")* caused the SG layer to update its internal representation to the same degree, in order to successfully

encode the semantic features of the unexpected object. This mirrored the empirical result that cloze probability, not constraint, is the primary determinant of N400 amplitudes in sentence contexts (Kutas & Hillyard, 1984; Federmeier, Wlotko, De Ochoa-Dewald & Kutas, 2007; Kuperberg, Brothers, & Wlotko, 2020).

The SG model was also able to simulate the related anomaly effect (Federmeier & Kutas, 1999). To simulate this effect, following the presentation of high constraint contexts, the network was then presented with three types of endings: expected words, anomalous words that were semantically *related* to the expected words, or anomalous words that were semantically *unrelated* to the expected completion. During training, although these "anomalous" nouns had never been presented with the preceding verb, the authors introduced different degrees of overlap between these words' distributed semantic features, represented in the Probe and Event Output layers. This ensured that the "related" anomalous words that shared partially overlapping sets of semantic features with the expected words, while the "unrelated" anomalous words had no semantic features in common with the expected word. Consistent with the empirical findings, the semantically related anomalous endings elicited a smaller update at the SG layer than the semantically unrelated anomalous completions.

To simulate the effects of word position during sentence processing (Van Petten & Kutas, 1991), the authors presented the longest sentences in their training set (e.g., *At breakfast, the man eats egg in the kitchen*), and simulated the N400 produced at each word position. All else kept equal, words presented earlier in the sentence induced a larger update at the SG layer, resulting in an approximately linear decline in N400 amplitude as the sentence progressed. Note, however, that the true probability of the incoming words did not decrease smoothly across the course of the sentence. For example, although actions (*eats*) were presented relatively early in the

sentence, they were 100% predictable in the model's training set when paired with a specific situation (*breakfast*). This surprising finding showed that the model (and the SG layer in particular) did not perfectly track corpus probabilities, but rather that it built up more confidence in its predictions as probabilistic cues accumulated over the course of the sentence.

Despite being trained on complete sentences, RHM18 were also able to simulate some of the lexical phenomena that were simulated in the word-level models described in Section 2 (Laszlo & Armstrong, 2014; Cheyette & Plaut, 2017; Rabovsky & McRae, 2014). During training, some words were presented more frequently than others. Therefore, the model was able to successfully simulate the effects of lexical frequency, independent of context. Specifically, when presented in isolation, higher frequency words (e.g. *egg*) elicited a smaller SG update than low frequency words (e.g. *toast*). In addition, when the model was presented with two-word sequences, it was able to simulate N400 priming effects, including repetition priming (*egg – egg* vs. *oak – egg*), associative priming (*eat – egg* vs. *play – egg*), and semantic feature priming (*cereal – egg* vs. *oak – egg*). Each of these priming effects was implemented somewhat differently. As noted above, during training, associatively related words commonly appeared together, while semantically related words shared overlapping units in the Event Output and Probe layers. In all three manipulations, target words produced a smaller update at the SG layer when preceded by related (*versus* unrelated) prime words.

Finally, the authors carried out a simulation of a set of results reported by Besson, Kutas & Van Petten (1992) to determine whether the updates in the SG layer, used to operationalize the N400, could also drive longer-term learning. Sentences with either congruous or incongruous endings (e.g. *The man eats eggs/*oaks*) were presented to the model, word by word. Each sentence was presented twice so that the authors could assess the effects of delayed repetition on

the simulated N400. As described above, the pattern of activity that was produced by word n (e.g. *eats*) implicitly predicted the semantic-thematic role associated with the subsequent word. Then, when this next word was presented (n+1, e.g. "*eggs*"), it produced its own pattern of SG activity. For the purposes of this simulation, the pattern of activation produced by this next word was considered the "target" for learning. The shift in SG activity between word n and word n+1 (corresponding to the N400 on word n) served as a "prediction error" signal that was backpropagated to modulate the model's weights (see RHM18, Supplementary Discussion for details). The assumption here was that, because the model was already trained, instead of learning to map a sequence of words to an event representation, the SG layer activation at the end of the sentence could be trusted to represent the "ground truth". Therefore, by minimizing the change in SG activation induced by incoming words, this "ground truth" SG state could be reached as early as possible, allowing the model to learn from new lexical inputs without relying on an explicitly available event representation.

With this learning procedure, sentence-final words elicited a smaller N400 on their second presentation (after learning). Moreover, the magnitude of this repetition effect was larger on semantically incongruous than congruous words, mirroring the empirical data (Besson, Kutas & Van Petten, 1992). The authors argued that this was because the larger N400 induced by the initial presentation of the incongruous words served as a "temporal difference error" that led to greater weight updates, i.e. increased learning.

Sentence Gestalt Model (2): Rabovsky, 2020

In a follow-up paper, Rabovsky (2020) reported that, with minimal modifications, the model developed by RHM18 was able to simulate and re-interpret an additional effect reported

in the prior literature: the so-called article-induced N400 effect. This effect was first described in an experiment by DeLong, Urbach and Kutas (2005), who observed larger N400 amplitudes to articles that were phonologically incompatible with an expected noun – for example, in the sentence "*The day was breezy so the boy went outside to fly an …*", *an* is phonologically incompatible with the expected continuation (*kite)* because this expected completion begins with a consonant (i.e., "*...a kite*" vs. " *...an airplane*"). The fact that the neural response was sensitive to the article's phonological incompatibility with the *upcoming* word was interpreted as evidence that comprehenders predicted this upcoming word's phonological form. However, an alternative explanation is that, instead of reflecting the prediction of the upcoming word's form, the N400 on the phonologically marked article reflected a shift in state at the event-level, which updated the probabilities of the upcoming noun.

To investigate this possibility, Rabovsky (2020) set up the Sentence Gestalt model so that the form of the indefinite article (*a* vs. *an*) reliably predicted the form of the upcoming noun. Specifically, the training procedure was designed so that articles could rule out particular upcoming nouns (e.g., *an* was never followed by *kite*). Mirroring the empirical results of DeLong, Urbach & Kutas (2005), articles that were incompatible with strongly expected upcoming nouns induced larger updates at the SG layer than articles that were compatible. This is because the incompatible article provided a highly reliable cue that the predicted event representation (e.g., a representation in which <kite> played the role of Patient) was probably incorrect, and therefore needed to be modified, inducing a shift at the SG layer.

Notably, subsequent studies (e.g., Ito, Martin, & Nieuwland, 2016), including a large-scale replication attempt (Nieuwland et al., 2018, but see Urbach, DeLong, Chan, & Kutas, 2020) found that the article-induced N400 effect was either absent or much weaker than the

effect originally reported by DeLong, Urbach and Kutas (2005). Rabovsky (2020) suggested that the discrepancy across studies may have arisen because the indefinite articles varied in how *reliably* they predicted the upcoming word's phonological form.

To investigate the effect of the article's predictive reliability, Rabovsky (2020) modified the model's training procedure so that the form of the indefinite article (*a* vs. *an*) was an *unreliable* predictor of the form of the upcoming noun. Specifically, the correlation between the form of the article and that of the upcoming noun was removed by introducing an intervening adjective ("*...an old kite*"; "*...a new airplane*"). In line with the cue validity interpretation, in this simulation, articles that were incompatible with the expected noun did *not* elicit a larger N400 than compatible articles.

Based on these observations, Rabovsky (2020) suggested that the article-induced N400 effect may not directly reflect the pre-activation of phonological information. Instead, they argued that this effect may be driven by a shift at the event level when articles provide a reliable cue that ruled out an otherwise high probability continuation (but see below for discussion of empirical evidence against this interpretation).

Insights and Limitations

RHM18 and Rabovsky (2020) were able to simulate an impressive range of N400 effects. Given that the model was trained on full sentences, its ability to simulate N400 effects in arbitrary word sequences (e.g., individual words, word pairs, word lists, etc.) is particularly striking, providing strong support for the theoretical perspective that words function as *cues* to sentence meaning (Rumelhart, 1979; Elman, 1990; Elman, 2009), rather than lexico-semantic

entities that must be retrieved independently and then composed into a higher-order event representation.

The model also addressed some of the limitations of Brouwer et al.'s Retrieval-Integration model. For example, in RHM18's Sentence Gestalt model, the shift of the SG layer was minimal to highly expected words and to semantic reversal anomalies, mirroring the empirical finding of a minimal N400 in these situations. In addition, the model's alternative account of role-reversal effects is consistent with the behavioral literature on the mis-assignment of thematic roles in reversible sentences (e.g., Ferreira, 2003; Gibson, Bergen & Piantadosi, 2013), and semantic illusions more generally (see Erickson & Mattson, 1981; see also Sanford, Leuthold, Bohan & Sanford, 2011).

Another strength of RHM18's approach is that it begins to address the question of *why* the brain might compute the internal change-in-state induced by each incoming word (the measure used to simulate the N400). Specifically, the model's successful simulation of long-term repetition effect on the N400 illustrates how, in principle, a shift in state can be used as a "temporal difference" signal to drive downstream learning. This builds on the work by Rabovsky and McRae (2014) by providing a more biologically and cognitive plausible mechanism that links comprehension and learning/adaptation.

Despite these successes, RHM18's Sentence Gestalt model also has some limitations. First, the claim that the N400 reflects the degree of shift induced by new inputs at the *event level* is at odds with recent empirical evidence, which suggests that large shifts in event probability alone are *not* necessarily a strong predictor of N400 modulation (e.g. Szewczyk, Mech, & Federmeier, 2021; see also Szewczyk & Wodniecka, 2020). In most situations, the degree of event-level shift induced by an incoming word will covary with the lexico-semantic

predictability of that word. However, these constructs can be dissociated. For example, consider the sentence context, "His skin was red from spending the day at the...". At this point in the sentence, it is likely that the comprehender will have already inferred a beach-related event. If, however, the comprehender then encounters the adjective "neighborhood", they are likely to shift their belief to an event about a community pool. However, Szewczyk, Mech, and Federmeier (2021), showed that a metric that operationalized the magnitude of this type of event-level shift on the adjective did not predict N400 amplitude as well as the adjective's lexico-semantic predictability (see also Federmeier, 2022, Box 2, page 16 for discussion).

Second, if the N400 effect reflects shifts within a single event-level state, this would predict that the effects of priming, predictability, and plausibility on this component should all localize to the same neuroanatomical regions. Again, however, this is at odds with the empirical data. Whereas the N400 effects of semantic priming (Lau, Phillips, Poeppel, 2008; Lau, Weber, Gramfort, Hämäläinen & Kuperberg; Lau, Gramfort, Hämäläinen & Kuperberg, 2013) and of predictability in plausible sentences (Wang et al., under review) both localize to regions of the left *temporal* lobe that support lexico-semantic processing, implausible words evoke an additional effect within left inferior frontal cortex (e.g. Wang et al., under review; Halgren et al., 2002; Marinkovic et al., 2003; Maess, Herrmann, Hahne, Nakamura, & Friederici, 2006; Ihara, Hayakawa, Wei, Munetsuna, & Fujimaki, 2007; Pylkkänen & McElree, 2007), which is thought to support updates at the event level (see Hagoort & Indefrey, 2014). These observations are inconsistent with a unitary event-updating account of the N400.

Relatedly, although the Sentence Gestalt model captures the effects of contextual facilitation via implicit predictions at the event level, these implicit predictions did not lead to *top-down* pre-activation of lexico-semantic information, encoded at a shorter time-scale at a

lower level of representation. This at odds with evidence that regions of the left temporal cortex can be *pre-activated* in constraining contexts (Dikker & Pylkkänen, 2013; Piai, Roelofs, Rommers, & Maris, 2015; Wang, Hagoort, & Jensen, 2018; Wang, Kuperberg, & Jensen, 2018).

A third limitation of the Sentence Gestalt's model concerns the link with learning. In their final simulation, the authors demonstrated that their N400 update measure could be used as a secondary teaching signal, allowing the model to simulate the effects of longer-term repetition priming. However, this learning procedure was not applied consistently across their simulations, and it is not clear *what* this alternative learning procedure was actually teaching the model. For example, in the absence of explicit event labels, the model could learn to minimize changes in the SG layer's simply by mapping all word inputs to the exact same pattern of SG activations (e.g. a string of zeros). Although this training would efficiently minimize the magnitude of state updates, the model would also "unlearn" the correspondence between words and their associated thematic-semantic roles. More generally, it is unclear how the model would calculate this update measure, which, as noted above, had to be computed externally by the modeler.

Error Propagation Model: Fitz & Chang, 2019

Introduction

In the two sentence-level architectures described above (Brouwer et al., 2017 and RHM18/Rabovsky 2020), the authors operationalized the N400 as an internal change-in-state induced by incoming words. However, neither of these models provided a compelling account of *how* (mechanistically) or *why* (functionally) the language comprehension system would compute these changes. As noted above, one possibility is that the N400 is computed implicitly as a byproduct of the process of shifting from a prior to a new state upon encountering new bottom-

up input. However, this would be at odds with RHM18's idea that the magnitude of this shift is explicitly computed and tracked for the purpose of downstream learning.

This issue was tackled head-on in a computational model by Fitz and Chang (2019). Similar to the word-level Semantic Attractor model by Rabovsky and McRae (2014), the authors proposed a "prediction error" account of the N400. Their model explicitly predicted each upcoming word, compared these lexical predictions to the word that was next encountered in a sentence, and used this difference – the prediction error – as a backpropagation signal to drive long-term learning. In a stark departure from all previous models, Fitz and Chang (2019) proposed that linguistic predictions (and the N400) are actually generated within the language production system ("prediction as production", see Federmeier, 2007; Pickering & Garrod, 2013), which runs independently and in parallel with sentence comprehension. Their model was therefore designed as a production model, and did not attempt to simulate comprehension per se.

Similar to RHM18, the authors attempted to simulate a wide range of sentence-level N400 phenomena, including the effects of lexical predictability, sentence constraint, word position, role reversals, as well as several syntactic phenomena (e.g. noun-verb agreement). They also aimed to simulate the influence of these factors on the P600 response, although a description of these additional simulations is outside the scope of this review. As discussed under Limitations, not all their simulations on the N400 were successful, and they were unable to model other findings, including the effects of lexical frequency, priming, or the effects of semantic overlap between expected and encountered words during sentence processing.

Model Characteristics

Unlike the previous sentence-level models of the N400, Fitz & Chang's model's was *not* trained to infer an event representation, but rather to accurately predict the next word of an unfolding sequence of lexical inputs. The model's architecture is depicted in Figure 2C. It consisted of a "core" Sequencing System that was trained to perform next-word prediction, and a Message System that, during training, sometimes provided additional event information that supported this process. The Message System was switched off altogether during the N400 simulations.

The Sequencing System performed next-word prediction by passing lexical inputs through a series of four layers: PrevWord → Hidden → Compress → NextWord. The pattern of activation within the final NextWord layer represented the model's next-word prediction. The central Hidden layer was recurrent and received activation from its own previous state. Another component of the Sequencing System was a PrevWordHistory layer that represented a running average of the pattern of activation based on previous inputs.

The Message System functioned to constrain next-word predictions in the Sequencing System during training. Events were encoded in the Message System through the connections between a Role layer (where each unit represented a thematic role: one unit for Agent, another for Action, etc.) and a Concept layer (where each unit corresponded to the semantic properties of a word: one unit for <dog>, another for <chase>, etc.). In this scheme, an event was uniquely represented by forming a temporary connection between a thematic role unit and its corresponding semantic units and removing all other connections between the Role and Concept layers. For example, the event corresponding to the sentence, "*A dog chases a cat*" would be expressed by connecting the Agent, Action and Patient units in the Role layer to <dog>, <chase> and <cat> in the Concept layer, with one-to-one connections. In addition to these Role and

Concept layers, which represented the role and concept of the word currently being predicted, the Message System also included CRole and CConcept Layers, which were structured identically, but represented the role and concept of the *previous* word in a sequence. The Message System also included a CRoleHistory layer, which represented a running average of the pattern of activation based on previously assigned thematic roles. Finally, the Message System included an EventSemantics layer that encoded verb alternations (cf. Levin, 1993) that were relevant for thematic role assignment.

During training, the model's weights were trained to predict the next word of sentences which were presented word by word to the PrevWord layer of the Sequencing System. In 30% of training trials, just before the first word in a sentence was presented (e.g. the sentence, "*A dog chases a cat*"), the model's Message System was given an event representation of the upcoming sentence, which included the appropriate set of thematic roles at the Role layer (Agent-Action-Patient) and their corresponding concepts at the Concept layer (<dog> <chase> <cat>). To illustrate how training worked, consider the stage of training after the model had generated a pattern of activation at the NextWord layer, which corresponded to the prediction of the second word in the sentence, "*She walk -ed*". Note, however, that everything we describe next occurred for every single word in a sentence.

First, the model explicitly compared the predicted pattern (the pattern generated at the NextWord layer before *walk* was encountered) with the target pattern (the vector corresponding to the observed word, *walk)*. The difference in activity – the prediction error – was backpropagated through the model in order to later update its weights such that error was minimized over future iterations of the model.

The vector corresponding to the presented word (*walk*) was then summed with its prediction, and this single vector was clamped at the PrevWord layer. This input at the PrevWord layer then activated the corresponding semantics and thematic role of the word, *walk,* at the CConcept layers and CRole (respectively) of the Message System, which, in turn, fed into the Hidden layer of the Sequencing System. This Hidden layer also received information from the CRoleHistory and from the EventSemantics layer of the Message System, as well as information about the full set of prior lexical inputs from the PrevWordHistory layer in the Sequencing System. The Hidden layer then combined all these inputs, and, from here, generated new predictions about the next word through *two routes*.

The first route was within the Sequencing System itself. Information in the Hidden layer was passed, via the Compress layer, to the NextWord layer. The second route was through the Message System. Information in the Hidden layer passed to the Role and Concept layers, which then passed information to the NextWord layer. Together, these two routes generated a predicted output pattern within the NextWord layer in the Sequencing System. The cycle then began again for the next word in the sentence (in this case, this "word" was the marker *-ed*).

The model's lexicon consisted of 88 tokens which were either words (e.g., *dog*) or marker morphemes (e.g., the progressive *-ing*). The model was trained on 50,000 sentences (each presented twice during training), produced by a symbolic generative grammar, which was designed to teach the model a range of semantic and syntactic regularities (cloze probability, noun-verb number agreement, tense inflection, etc.).

N400 Simulations

For all N400 simulations, the Message System was switched off, and the model no longer had access to event representations, prior to sentence onset. Instead, the Sequencing System had to predict the next word based only on the preceding context. The authors operationalized the N400 as the difference between the target pattern (the observed word) and the model's prior prediction (the pattern of activation produced at the NextWord layer, at the previous time-step), which they referred to as a "prediction error". As discussed above, during training, this prediction error played a crucial role in triggering the necessary weight updates that would minimize the average prediction error produced over future iterations.

As in RHM18's model, the authors simulated the effects of predictability and contextual constraint by manipulating how often particular verb-object pairs were presented during training. For example, the verbs *drink*, *taste* and *take* were presented with the object *water* 60%, 15% and 4% of the time, respectively. As expected, after training, the model showed progressively smaller prediction errors on the final critical word as contextual predictability increased (e.g., *A teacher was drink/taste/take-ing the* <u>*water*</u>) (cf. Kutas & Hillyard, 1984; DeLong, Urbach & Kutas, 1995). Moreover, when the model was presented with high and low constraint sentences (e.g., high constraint: *the woman will sip*….most expected ending: *tea* (60%); low constraint: *the woman will sniff*…most expected ending: *wine* (40%)), the prediction error elicited by an *unexpected* critical word (e.g. *water*) did not differ between the high and low constraint conditions. This is again consistent with the empirical literature (Kutas & Hillyard, 1984; Federmeier, Wlotko, De Ochoa-Dewald & Kutas, 2007; Kuperberg, Brothers, & Wlotko, 2020).

Also similar to RHM18, the authors were able to simulate the effect of word position on the N400 (Van Petten & Kutas, 1990; Van Petten & Kutas, 1991; Payne, Lee & Federmeier, 2015): they found that the prediction errors tended to decrease for later word positions in

congruous sentences (*A grandma give -ed the clerk a beer*). Extending RHM18's findings, they further showed that this word position effect was not produced in semantically incoherent sentences (*A pencil give -ed the coffee a friend*), again mirroring the empirical findings. These findings suggest that the model was able to make increasingly more accurate predictions as the sentence progressed, but only when the sentence was semantically coherent.

The authors also attempted to simulate patterns of neural activity (N400/P600) when processing different classes of syntactic violations (noun-verb number agreement violations: Hagoort, Brown & Groothusen, 1993; tense inflection violations: Allen et al., 2003; word category violations: Friederici, Hahne & Mecklinger, 1996, and verb subcategorization violations: Osterhout & Holcomb, 1992), as well as role reversal anomalies (Kuperberg, Sitnikova, Caplan & Holcomb, 2003; Kolk, Chwilla, van Herten, & Oor, 2003; Hoeks, Stowe & Doedens, 2004; Kim & Osterhout, 2005), albeit unsuccessfully. To simulate the effects of role reversal on the N400, the authors compared active sentences (*The woman is take -ing the pencil*) with their role-reversed counterparts (*The pencil is take -ing the woman*). However, at the critical morpheme (*-ing*), the model showed a larger prediction error on the implausible reversals, which is inconsistent with the null N400 effects reported in the ERP literature. Similarly, the model also produced larger prediction errors to certain syntactic violations (*The man will sip -ed the beer*) than non-violated controls. Although these effects were small, and somewhat variable across syntactic manipulations, they are generally inconsistent with the empirical findings, which report no N400 modulation on purely syntactic violations.

Finally, like RHM18, Fitz & Chang (2019) successfully simulated the interaction between predictability and long-term repetition on the N400, this time focusing on empirical findings reported by Rommers and Federmeier (2018). In this simulation, the model was

presented, word by word, with sentences that either had predictable endings (e.g. *Alfonso has started biking to work instead of driving his car*) or unpredictable endings (e.g. *Jason tried to make space for others by moving his car*). After each sentence, the model's weights were updated, and then a second sentence was presented. This second sentence either had the same sentence-final word as the original sentences (Repeated) or it had a different sentence-final word (New). Smaller prediction errors were produced by the sentence-final words in the Repeated than the New condition. Moreover, mirroring the empirical findings, this repetition effect on the simulated N400 was larger when the initially-presented word was unpredictable than when it was predictable. This effect can be explained intuitively within this prediction error framework because larger prediction errors drive greater weight updates. Therefore, a large prediction error produced by an unpredictable word on the first presentation of a sentence would have led the model to generate stronger predictions for this word on subsequent trials, resulting in a larger reduction of the N400 when these predictions were confirmed in Repeated (versus New) sentences.

Insights and Limitations

By linking the N400 to prediction error, computed for the purpose of downstream learning, Fitz and Chang (2019)'s Error Propagation model provides an important extension of the previous sentence-level approaches. Building on the word-level model by Rabovsky and McRae (2014), and one of the simulations carried out by RHM18, they showed that the N400, operationalized as lexical prediction error, can serve as signal that drives longer-term learning. Like any supervised connectionist model, the targets for learning in this model had to be provided externally by the modeler. However, unlike Rabovsky & McRae (2014) in which each semantic target was provided somewhat arbitrarily, in Fitz and Chang's model, each target was

provided as the subsequent word in a connected sentence, mirroring the process of "natural" language comprehension.

There were, however, important limitations. The model was trained to carry out next word prediction. Therefore, it was successful in being able to explain the sensitivity of the N400 to contextual probability, i.e. graded cloze effects (Kutas & Hillyard, 1984; DeLong, Urbach & Kutas, 2005) and word position effects (Van Petten & Kutas, 1990; Van Petten & Kutas, 1991; Payne, Lee & Federmeier, 2015). However, it was less successful in simulating other sentence-level phenomena. For example, in the empirical literature, semantic role reversal anomalies (Kuperberg, Sitnikova, Caplan & Holcomb, 2003; Kolk, Chwilla, van Herten, & Oor, 2003; Hoeks, Stowe & Doedens, 2004; Kim & Osterhout, 2005) and syntactic violations (Allen et al., 2003 ; Osterhout & Mobley, 1995 ; Osterhout & Holcomb, 1992) generally show no effect on the N400. However, in Fitz and Chang's Error Propagation model, these manipulations produced clear simulated N400 effects. In the case of reversal anomalies, the authors suggested that their simple grammar may have artificially inflated the predictability of congruous continuations, resulting in an N400 effect. They also note that small N400 differences are sometimes observed on reversal anomalies. However, these effects are quite weak and typically only appear with longer distances between verbs and arguments (e.g. Chow, Lau, Wang, & Phillips, 2018).

The authors also made no attempt to simulate some of the other phenomena that were successfully simulated by RHM18: the related anomaly effect (cf. Federmeier & Kutas, 1999) and simple semantic or associative priming effects. This is not surprising, and brings us to the fundamental problem with Fitz and Chang's Error Propagation model: By framing the N400 as a byproduct of the language *production* system, it ignores the relationship between the N400 and language *comprehension* altogether. To *comprehend* language accurately and efficiently, we must

predict and infer not only upcoming individual lexical items, but also the semantic features associated with these words, and, ultimately, the events that are conveyed by sequences of words. Fitz and Chang's Error Propagation model, however, was trained *only* to predict specific lexical items. In contrast to all previous models of the N400 discussed thus far, these words were not linked to *distributed* semantic features; no features were shared between words, and the model did not actually infer word meanings. This explains why the model would be unable to simulate effects of semantic overlap either during priming or sentence comprehension. Even more fundamentally, and in contrast to RHM18's Sentence Gestalt model, upcoming word prediction played no role in inferring whole events (the purpose of comprehension). This failure to infer events explains why the model was unable to simulate effects like associative priming that rely on the co-occurrence of words around canonical events.

Section 4: Summary of Models in relation to Cognitive and Biological constraints

In the previous sections, we reviewed several computational models of the N400, each with their own set of strengths and weaknesses. Although in combination, these word-level and sentence-level models were able to account for an impressive range of empirical phenomena, none of them provided a complete account of the N400 that was both cognitively and biologically plausible. In this section, we summarize these models in relation to a set of cognitive and biological constraints that have informed our understanding of both language comprehension and the N400 over the past few decades. In the following section, we will turn to a new framework of understanding the N400 known as *predictive coding,* which satisfies many of these constraints.

1. Explanation of empirical phenomena

As we have seen, the N400 is influenced by a large number of factors, including the lexical properties of single words, minimal contexts (priming), and wider sentence contexts (including the incremental effects of contextual predictability). Ideally, a computational model of the N400 should be able to account for this broad range of empirical findings parsimoniously, within a single framework; that is, it should be able to process arbitrary orthographic (or phonological) inputs (*both* words and non-words), and it should be possible to independently manipulate both lexical properties (e.g. frequency or concreteness) and contextual factors, as well as interactions between these variables, to examine their effects on the simulated N400.

Currently, no single model has been able to fulfill this criterion. Indeed, there is a strong divide in the empirical coverage of word-level and sentence-level models. On the one hand, the models by LPAC and Rabovsky & McRae (2014) were very successful in simulating a wide range of lexical and priming effects. However, the ability to simulate sentence-level N400 effects (cloze probability, semantic congruity effects, related anomaly effects, etc.) was completely outside the scope of these models. Conversely, the sentence-level models were able to simulate an impressive range of contextual phenomena. However, they were unable to simulate all the effects simulated by the word-level models. For example, even though RHM18 were able to explain the effects of lexical frequency and priming, like all other sentence-models, it had simplified (localist) lexical representations that didn't connect to orthography (or phonology). It was therefore unable to simulate the effects of orthographic neighborhood and was unable to process non-words.

2. Incorporation of interactive, hierarchically organized representations

In order to incorporate interactions between lexical and contextual factors, we believe that models of the N400 should incorporate a *hierarchical structure* that enables information to

be represented at different temporal scales (e.g. letters, words, events), and for levels of this hierarchy to continuously *interact* through both top-down and bottom-up connections. This type of architecture would be consistent with what we know about the hierarchical organization of language (Jackendoff, 2002), the strong interactivity across levels of linguistic representation during processing (e.g. Tanenhaus & Trueswell, 1995; MacDonald, Pearlmutter & Seidenberg, 1994), and the continuous interaction between linguistic representations and the higher-level event (and broader situation model) that is being incrementally inferred from the input (Kuperberg, 2013; McRae & Matsuki, 2009). It would also be consistent with what we know about the hierarchical organization of the cortex, and the neurobiology of language processing, with different areas of cortex being specialized for encoding and processing information at different levels of information at different time scales (e.g., Hickok & Peoppel, 2007; Dehaene, Cohen, Sigman, & Vinckier, 2005; Price & Devlin, 2011; Wang et al., under review).

The notion of continuous top-down/bottom-up interactions across hierarchically-organized layers of representation was captured in the seminal *interactive activation* models of written word recognition (McClelland & Rumelhart, 1981) and speech perception (McClelland & Elman, 1986) that have inspired a huge body of psycholinguistic work over the past few decades. However, the word-level N400 models discussed above (LPAC; Rabovsky & McRae, 2014) relied exclusively on bottom-up inputs, and these models lacked a contextual layer that could influence ongoing lexico-semantic processing.

Conversely, none of the sentence-level N400 models reviewed above included a lower-level orthographic layer, and none of these models fully capture *both* hierarchical structure and interactive principles. For example, in Fitz and Chang's Error Propagation model, even though there were multiple types of representation (lexical items, semantic-thematic roles, events), these

did not influence the comprehension system at all. RHM18's Sentence Gestalt network did not distinguish lexical, semantic and event information; that is, lexical inputs were mapped directly on to event representations via an unstructured hidden layer, with no intermediate levels of representation, and the N400 reflected incremental updates within a *single* dynamic event state. The notion of hierarchy was best expressed in Brouwer et al.'s Retrieval-Integration model, which captured linguistic information at different time-scales (word-level semantics *vs.* sentence-level events) by incorporating distinct Retrieval and Integration modules. However, as discussed earlier, the Integration layer in this model did not actually provide top-down pre-activation of upcoming semantic representations in the Semantic Retrieval layer.

The idea that higher-level event representations can generate predictions that can be propagated down to lower levels of a representational hierarchy is a central component of predictive language comprehension (DeLong, Urbach, & Kutas, 2005; Federmeier, 2007; Kuperberg & Jaeger 2016). This type of *top-down predictive pre-activation* goes beyond the type of *implicit* temporal predictions of upcoming semantic-thematic roles that are implemented as recurrent feedback connections at the event level in Brouwer et al.'s Retrieval-Integration model and in RHM18's Sentence Gestalt model. These top-down predictive frameworks posit that, because higher-level event information is encoded at a longer spatiotemporal scale than lower-level lexico-semantic information, then this is exploited to provide top-down pre-activation at lower lexico-semantic levels *before* an anticipated input appears (see Kuperberg & Jaeger, 2016, page Section 3 for a detailed discussion). Indeed there is neurobiological evidence that supports this type of top-down pre-activation. For example, MEG studies have reported differences in oscillatory activity within left temporal regions, which are thought to support lexico-semantic processing, following highly predictive *versus* less predictive contexts (Dikker & Pylkkänen,

2013; Piai, Roelofs, Rommers, & Maris, 2015; Wang, Hagoort, & Jensen, 2018). In addition, we have shown that the left ventromedial temporal lobe produces *item-specific temporal patterns* of neural activity that correspond to the pre-activation of specific individual words (e.g. "baby" in the context of "In the crib, there is a sleeping …") (Wang, Kuperberg, & Jensen, 2018). None of sentence-level models reviewed above fully capture this type of top-down predictive lexico-semantic pre-activation.

3. Cognitive plausibility: Facilitatory effects of prior context during language processing

A central claim of all cognitive theories of the N400 is that the amplitude reductions produced by priming and contextual predictability are linked to *facilitated* processing of an incoming target word. This link between N400 attenuation and cognitive facilitation has been discussed in different ways in the literature. In the priming literature, the reduced N400 to primed (*versus* unprimed) targets has been taken to reflect easier access to the lexico-semantic representation of that target. In the sentence processing literature, it has been variously argued that the smaller N400s to predictable or congruous sentence continuations is linked to facilitated "integration" of words into their prior context (e.g. Hagoort, Baggio, & Willems, 2009), to facilitated access/retrieval of lexico-semantic information that has already been pre-activated (Federmeier & Kutas, 1999; Lau, Phillips, Poeppel, 2008), or to both facilitated integration and retrieval within a dynamically interactive system (e.g. Baggio & Hagoort, 2011). What is common to all of these accounts is the assumption that more predictable words produce a smaller N400 because they are relatively "easier" to process[6]. At a computational level, this facilitated

[6] We note two points here. First, it is, of course, logically possible that the N400 reflects a neural signal that is only superficially correlated with facilitation. However, to the extent that this correlation systematically occurs in such a wide variety of contexts, we think that there is likely

processing would translate on to how easily a model settles on a particular target representation in the 300-500ms (N400) time window. However, not all the models that we have discussed capture this central idea.

The discrepancy between contextual facilitation and the simulated N400 response was most apparent in how the Semantic Activation models simulated priming (Laszlo & Armstrong, 2014; Cheyette & Plaut, 2017). Recall that, in these models, without additional assumptions, primed words would produce a *larger* simulated N400 response (more total semantic activation) than unrelated words. To remedy this issue, the authors introduced a decay-driven inhibition mechanism. However, while this correctly resulted in a reduction in the simulated N400, the inhibition of semantic features also made primed target words *more* difficult to access (rather than easier to access), running contrary to all current theoretical models of priming.

This particular issue was side-stepped by Rabovsky and McRae (2014) and Fitz and Chang (2019), who linked the N400 to "prediction error" rather than to the total activation of semantic features. However, in both these models, prediction errors failed to influence the activation of semantic features. Specifically, in Rabovsky and McRae's word-level model, prediction error was defined as the "distance" between the current state of semantic activity induced by the input and an external target pattern that was only available to the modeler. Therefore, this error had no influence on the model's own activations during online processing. Similarly, in Fitz & Chang (2019), prediction errors were calculated outside the language comprehension system, and they were not used to update the model's internal weights until the

to be a deeper relationship between the attenuation of the N400 and facilitation. Second, the notion of contextual facilitation on the N400 is closely related to contextual facilitation on behavior. We return to this point in Section 6 (Future Directions).

end of the sentence. Therefore, it was unclear how this could lead to facilitation, or, indeed, have any influence on online processing.

Brouwer et al.'s model begins to capture the idea of cognitive facilitation by simulating the N400 reduction to associated (*versus* non-associated) words in sentence contexts as a smaller shift at the Semantic Retrieval layer. However, as discussed earlier, a failure to *pre-activate* information at the Semantic Retrieval layer, prior to word onset, meant that even highly predictable sentence-final verbs induced large shifts in state, which is inconsistent with most conceptions of facilitated processing.

Of all the sentence-level models, the one that best captures facilitatory effects on the N400 is RHM18's Sentence Gestalt Model. When the sentence-level meaning was fully predictable, incoming words induced only a minimal shift in the SG layer. In other words, it was easier for the model to settle on an event state when the preceding context was predictable than unpredictable. On the other hand, as discussed in Section 3, this formalization of the N400 as a shift within a single-state system is not supported by MEG and intracranial data, which show that the facilitatory effects of priming, predictability and plausibility on the N400 do not all localize to the same neuroanatomical regions.

4. Biological plausibility: An explicit and biologically plausible linking function

Another important constraint in developing any model of the N400 is its biological plausibility. In other words, the structure and computations of these models should, in principle, be implementable in neural tissue. In particular, the model should specify a biologically plausible mechanism that links its N400 construct (semantic activation, prediction error, shift in state) to

differences in neural activity that can be recorded at the surface of the scalp. Moreover, this construct should ideally be able to capture the morphology and latency of the N400 waveform.

The Semantic Activation models satisfy this criterion best. Taking a step in the direction of biological realism, LPAC explicitly constrained their model architecture to reflect certain established neurobiological constraints. In addition, the linking function that translated the model's activation into a neural signature is clear. Specifically, if individual processing units in this model correspond to cortical neurons (p. 273, Laszlo & Plaut, 2012), any process that increases the total activation (or firing rate) of these units would plausibly give rise to a larger ERP response at the scalp. Finally, this was the only model in which the simulated N400 traced out a rise-and-fall trajectory, with a relatively constant peak latency across different classes of stimuli, similar to the N400 itself. This contrasts with all the other models we reviewed, which used architectures that did not take biologically-motivated constraints into account.

In the error-based accounts, the N400 was computed explicitly as the difference between the current semantic state and an ideal target (Rabovsky & McRae, 2014), or between a word input and the model's prior lexical prediction (Fitz & Chang, 2019). However, as discussed above, because these error measures were calculated outside the model, the direct link with ongoing neural activity was unclear. In the change-of-state accounts (Brouwer et al., 2017; RHM18), the assumption was that the construct reflecting the N400 was generated implicitly as an update between two successive time points. However, it was left unspecified how or why this change in state would lead to an overall increase in neural activity at the scalp surface.

Finally, another important criterion is the biological plausibility of the model's learning procedure. In all the models of the N400 reviewed above, the model's weights were trained using backpropagation, and, in three of these models, the authors directly linked the errors used for

backpropagation to the simulated N400 response (Rabovsky & McRae, 2014, Fitz & Chang, 2019, RHM18 in one simulation). However, several aspects of the backpropagation algorithm are known to be biologically implausible (Grossberg, 1987; Whittington & Bogacz, 2019). For example, it is unclear how error information computed outside the model would be transmitted "backwards" across multiple layers of cortex to update weights. As we discuss in the next section, algorithms that rely on more local error representations (e.g. Lillicrap et al., 2020; see Whittington & Bogacz, 2019, for discussion) may be necessary to develop a more biologically realistic account of the N400.

Section 5: A Predictive Coding Account of the N400

We now turn to a computational framework known as *predictive coding* that accommodates many of the constraints described above. In the language comprehension literature, the term, "predictive coding" has sometimes been used loosely to refer *any* form of top-down predictive processing in the brain, or to any framework of language comprehension that involves the production of "prediction error". However, the term "predictive coding" actually describes a specific computational algorithm and architecture, with a particular arrangement of feed-forward and feedback connections, that was first developed in the visual system to simulate extra-classical receptive field effects (Rao & Ballard, 1999; Spratling, 2012, 2013, 2014; see also Mumford, 1992). Over time, predictive coding has been expanded into a more general theory for how information is transmitted between cortical areas, allowing the brain to perform probabilistic inference in multiple domains of perception and cognition (Friston, 2005; Clark, 2013; Spratling, 2016), including lower levels of language processing, such as

speech perception (Blank & Davis, 2016; Sohoglu & Davis, 2020) and visual word recognition (see Price & Devlin, 2011; Heilbron, Richter, Ekman, Hagoort, & de Lange, 2020).

Predictive coding is fundamentally an optimization algorithm. One of its central claims is that each level of cortical representation has a distinct population of "state units" that encode its internal representations and "error units" that pass information between cortical areas (Rao & Ballard, 1999; Spratling, 2017; Friston, 2005). The patterns of state activity encoded in a given layer can be thought of as a dynamically changing "target" pattern for a higher-level state. At each point in time, state units at higher cortical layers generate a top-down *prediction* (or reconstruction) of this target pattern at the level below. Lower-level "error units" then calculate the *residual difference* in information between this top-down prediction and the target state pattern, either by subtraction (cf. Rao & Ballard, 1999) or division (cf. Spratling, 2009), depending on the precise predictive coding algorithm.

The error units compute two types of residual information. The first is *prediction error* — the information that is encoded in the target state, but *not* in the current top-down prediction. The second is *top-down bias* — the information that is encoded in the top-down predicted state but not in the target state. These two types of residual information play different roles in the algorithm. First, prediction error is passed up to the cortical level above where it is used to update higher-level state units. This allows these higher-level states to generate more accurate top-down predictions/reconstructions on the next iteration of the algorithm. Second, the *top-down bias* modifies the target state pattern at the same cortical level, bringing it closer to the prediction from the level above. Therefore, at each iteration of the algorithm, the state at each level of the cortical hierarchy is modified in two ways: one that helps it better predict its lower-level target pattern (driven by bottom-up prediction error), and another that helps it serve as a

better target to a yet higher-level state (driven by top-down bias). Over multiple iterations of the algorithm, the magnitude of prediction error and top-down bias gradually decreases, and the model settles into a global state that can accurately explain the bottom-up input at multiple levels of representation.

As several researchers have pointed out, the N400 is very naturally interpreted as the new lexico-semantic information, encoded within the bottom-up input, which has not already been predicted by the prior context, i.e. as the magnitude of lexico-semantic prediction error within a predictive coding framework (e.g. Xiang & Kuperberg, 2015; Kuperberg 2016; Bornkessel-Schlesewsky & Schlesewsky, 2019; Kuperberg, Brothers & Wlotko, 2020). Note that, in this framework, the term "prediction error" does *not* correspond to the detection of a linguistic anomaly or to a *violation* of a strong top-down prediction. Instead, similar to the previous prediction error computational frameworks discussed above (Rabovsky & McRae, 2014; Fitz & Chang, 2019), it simply refers to the difference between a "target" pattern of lexico-semantic activity and the lexico-semantic activity that was predicted by the model. However, there are four important differences between the prediction errors that are computed during predictive coding and the prediction errors or shifts in state have been used to simulate the N400 in previous computational models.

First, in these previous models, the "target" pattern of activity constituted a single input that was provided outside the model, whereas in predictive coding, the target pattern constitutes a dynamic state *within* the model that is continually updated by the model's algorithm. Second, in these previous models, errors were computed outside the model. In predictive coding, however, errors are computed *within* the model, locally at each level of representation. Third, in these previous models, the error produced by a particular input was computed *after* a lexical item or its

semantic features were activated, in order to drive subsequent learning. In predictive coding, the computation of prediction error drives comprehension itself (i.e. the process of *inferring* the semantic features corresponding to the input), although as we discuss below, these errors can, in principle, *also* be used for downstream learning.

Finally, in previous models, "prediction error" was used to simulated *either* the effects of lexical factors (e.g., orthographic neighborhood and semantic richness) *or* the effects of context on the N400, raising questions of how this single univariate signal can explain the "multiplicity of factors" that are known to affect N400 amplitude (see Federmeier, 2022, Box 2, page 16). Hierarchical predictive coding begins to address these concerns by positing a continuous *interaction* across levels of representation, allowing *both* higher-level and lower-level sources of information to influence the magnitude of prediction error computed during real time comprehension (Rao & Ballard, 1999; Mumford, 1992; Friston, 2005).

In the sections below we describe a hierarchical Predictive Coding model of lexico-semantic processing, implemented in our own lab, which simulated the N400 as the magnitude of lexico-semantic prediction error computed by a predictive coding algorithm. Similar to the models of the N400 described above, we first explain our motivation for this work; we then describe the specific architecture of this model and our N400 simulations. Finally, we discuss the insights of this approach in relation to the four criteria outlined above.

Predictive Coding Model: Nour-Eddine, Brothers, Wang, Spratling & Kuperberg, 2022

Introduction

In developing our Predictive Coding model, we had three major goals. First, we wanted to determine if the same predictive coding principles that can explain neural activity in the visual

and auditory systems could be used to simulate the N400. We therefore adopted the same processing algorithm and the same network architecture that has previously successfully simulated various perceptual and cognitive phenomena (Spratling, 2012, 2013, 2014, 2016), changing only the model's internal representations (orthographic, lexical and semantic).

Second, we were interested in whether a single model can account for both lexical and *and* contextual-level effects on the N400. Specifically, we aimed to account for the effects that were captured by the word-level models reviewed in Section 2, including lexical frequency, orthographic neighborhood size, semantic richness, repetition priming and semantic priming. We also aimed to simulate an additional effect that these previous models were not able tosimulate—the larger N400 elicited in response to pseudowords than to words (cf. Bentin, 1987; Holcomb et al., 2002; Meade et al., 2018). In addition, at the sentence-level, we aimed to simulate several of the contextual effects that were successfully simulated by RHM18, including the effects of cloze probability (Kutas & Hillyard, 1984; DeLong et al., 2005), the null effect of constraint (Kutas & Hillyard, 1984; Federmeier et al., 2007), and the effect of semantic overlap between expected and presented words (Kutas & Hillyard, 1984; Federmeier & Kutas, 1999). We also aimed to simulate the effect of *orthographic overlap* between expected and presented words (Laszlo & Federmeier, 2009; Ito, Corley, Pickering, Martin & Nieuwland, 2016; DeLong, Chan & Kutas, 2019), which was not simulated in previous models. For example, Laszlo and Federmeier (2009) showed that the N400 was attenuated on both words and non-words that were anomalous, but shared orthography with an expected completion (e.g., a reduced N400 to *dish* or *wush*, if the sentence constrained strongly for *wish*). Finally, in contrast to previous models, we aimed to simulate *interactions* between the top-down effects of contextual predictability and lexical factors.

Our third goal was to determine whether the lexico-semantic prediction error, computed by error units, would mirror the morphology of the N400 effect as it unfolded over time. As noted above, thus far, only the Semantic Activation models succeeded in accounting for the N400's morphology, and they did so by inducing an initial excitatory imbalance in the network followed by a delayed inhibition. In contrast, predictive coding explains the upslope and downslope of ERP components in terms of the rise and fall in prediction error as the model converges on an accurate state-unit representation that "switches off" lower-level error (Friston, 2005).

Model Characteristics

The model's architecture is shown in Figure 3A. It was organized hierarchically, with four layers and three levels of linguistic representation (orthographic, lexical, semantic). As in all predictive coding architectures, each level of representation had distinct populations of error units and state units. The lowest orthographic level included a set of 104 error units and 104 state units, which encoded 26 letter identities (A-Z) at four possible spatial positions (cf. McClelland and Rumelhart, 1981). The middle lexical level consisted of a set of 1,579 *lexical* state/error units; the units in each set corresponded to the 1,579 four-letter words in the model's lexicon (e.g., *lime*, *corn*). The third semantic level consisted of a set of 12,929 *semantic* state/error units, which each represented a unique semantic feature (e.g. <plant>, <sour>; cf. Rabovsky & McRae, 2014). The highest layer of the hierarchy, which only had state units, was the "dummy layer"; these 1,579 units allowed the modeler to provide top-down predictions that are thought to be generated based on a higher-level event state during incremental sentence comprehension.

Critically, the precise connections between error and state units, within and across layers, was based on predictive coding principles. In particular, state units at consecutive levels communicated exclusively through error units. Therefore, at each level, error units were connected one-to-one to state units at the same level, and had symmetric feedforward and feedback connections to and from the state units at the higher level.

The connection weights were hand-coded as matrices that described the contingencies/mappings between levels. For example, orthographic error units coding for L, I, M, and E had feedforward connections to, and feedback connections from, the lexical state unit coding for *lime*. Similarly, each lexical error unit (e.g. *lime*) was only connected to its corresponding semantic state units (e.g. <sour>, <plant>, etc.). As described further below, the model was set up to simultaneously capture lexical frequency, semantic richness, orthographic neighborhood and semantic relatedness.

Note that, unlike classic Interactive Activation connectionist models (e.g. Chen & Mirman, 2012; McClelland & Rumelhart, 1981; McClelland & Elman, 1986), predictive coding architectures have no lateral inhibitory connections between state units. Instead, state units at a given level compete by inhibiting the inputs of their neighbors, i.e. by suppressing bottom-up prediction errors (see Spratling, 2008 for discussion).

In this model, we implemented the Predictive Coding/Biased Competition-Divisive Input Modulation algorithm (Spratling, 2009). In this formulation, prediction errors are calculated by division rather than subtraction (Prediction Error = State / Prediction; Top-down Bias = Prediction / State). This ensured rapid convergence of the algorithm, and also guaranteed that the activity across all units remained non-negative, similar to biological neurons. We operationalized

the N400 as the magnitude of lexico-semantic prediction error produced by lexical and semantic error units in the model.

N400 Simulations

To simulate the N400 on each trial, an orthographic state (L-I-M-E) was clamped at the bottom of the network. We then computed the total prediction error produced by both lexical and semantic error units at each iteration of the algorithm. This resulted in a full time course of the simulated N400, as the model settled into a stable state at each level.

In all simulations, we found that the time course of these prediction errors mirrored the time course of the N400, see Figure 3B. After stimulus onset, lexico-semantic prediction error rose quickly to a peak, as new state units were activated in the model, and could not be explained (suppressed) by predictions/reconstructions from the level above. These errors then gradually fell as the model converged on a set of lexical, semantic and dummy event states that could accurately explain the pattern of orthographic activity.

For all of the simulations described below, we selected a set of 512 critical words from the model's lexicon whose lexical variables were uncorrelated (orthographic neighborhood size, lexical frequency, semantic richness). This allowed us to examine the effects of each lexical variable, while holding the others constant.

Bottom-up Lexical simulations

Similar to previous models, the orthographic neighborhood size of each word was determined based on the number of overlapping words in the model's internal lexicon. We found that words with more orthographic neighbors (e.g. *core*) elicited a larger lexico-semantic error than words with few neighbors (e.g. *kiwi*) (cf. Holcomb, Grainger & O'Rouke, 2002; Laszlo &

Federmeier, 2007; Laszlo & Federmeier, 2011; Laszlo & Federmeier, 2014). This occurred because high neighborhood words partially activated the lexical and semantic state units of their orthographic neighbors, resulting in lexico-semantic prediction errors that were relatively non-specific. For example, when the model was presented with the input C-O-R-E, it attempted to minimize the prediction error on the C, O and R orthographic units by partially activating a large number of partially compatible lexical state units (*corn*, *corp*, *cork*, etc). These lexical errors were then passed to the semantic level, generating larger prediction error responses throughout the network. Eventually, the model always settled on the correct set of lexical and semantic states, but this process triggered larger prediction errors in the presence of orthographic competitors.

Replicating Laszlo and Plaut (2012), the model also produced robust neighborhood effects even when processing non-word stimuli (cf. Laszlo & Federmeier, 2007): for both words and non-words, we observed a strong positive correlation between the magnitude of lexico-semantic error and orthographic neighborhood size. Notably, unlike any previous model, we also found that non-word inputs (e.g., W-U-S-H) elicited a greater lexico-semantic prediction error than real words, even when matched on orthographic neighborhood (cf. Holcomb, Grainger & O'Rourke, 2002; Meade, Midgley, Dijkstra & Holcomb, 2018; Meade, Grainger & Holcomb, 2019; but see Laszlo & Federmeier, 2011). This is because pseudowords activated a combination of lexical and semantic state units (e.g., *wish*, *wash*, *rush*, *lush*) in order to minimize orthographic prediction errors at each letter position. Because there was no single lexico-semantic state to suppress these prediction errors, the simulated N400 remained higher, overall, compared to real-word inputs.

Lexical frequency was simulated by introducing a stable top-down "prior" in the model. This was implemented by modifying the top-down connections between higher-level state units and lower-level error units; specifically, we increased the strength of each non-zero feedback connection weight by a positive value, proportional to each word's SUBTLEX-US frequency (Brysbaert & New, 2009). At a given level of state unit activity, units associated with higher frequency words produced stronger predictions, allowing this word to be inferred more easily. As a consequence, we found that higher frequency inputs elicited a smaller lexico-semantic prediction error than lower frequency inputs (cf. Rugg, 1990; Van Petten & Kutas, 1990).

Semantic richness was simulated by varying the number of semantic features linked to each word (9 *vs.* 18 features). Mirroring the empirical findings (e.g. Kounios & Holcomb, 1994; Holcomb, Kounios, Anderson & West, 1999; Lee & Federmeier, 2008), words with a larger number of semantic features produced larger lexico-semantic prediction errors. This is because activating additional state units (with the same amount of unpredicted information per unit) activated a larger number of prediction error units, and this error activity summed to produce a larger simulated N400.

Priming effects

Following previous studies, we simulated priming by presenting the model with an initial "prime" input, followed by two blank iterations, and then a "target" input. In addition to simulating the effects of repetition priming (cf. Rugg, 1985), we were also able to simulate semantic priming (cf. Bentin, McCarthy &Wood, 1985; Rugg, 1985; Holcomb, 1988; Holcomb & Neville, 1990) because semantic features were sometimes shared between words (e.g., the semantic feature, <is-plant> is shared by the lexical items, *lime* and *corn*). As expected, primed

target words elicited smaller lexico-semantic prediction errors than unprimed targets. This is because, when processing the prime, the model's state units settled on a set of semantic features, which "lingered" in the intervening period between the prime and target. If the semantic features of the target either partially (semantic priming) or fully (repetition priming) overlapped with the prime, this overlap allowed the model to settle more quickly and to suppress the prediction error produced at the level below.

Top-down contextual effect simulations

We were also interested in the effects of broader context on the N400. To this end, we simulated predictive *pre-activation* by clamping the highest dummy event layer to a desired cloze probability distribution[7]. During this pre-activation phase, this layer produced semantic predictions, which induced a top-down bias in the model's semantic state units. Newly activated semantic states then led to the pre-activation of lexical state units, and so on down the network. When pre-activation was complete, the model's state units at each level were aligned with the probability distribution of the state units at the dummy event layer, at which point we presented the model with expected or unexpected orthographic inputs.

To simulate the graded effects of cloze (Kutas & Hillyard, 1984; DeLong, Urbach & Kutas, 2005; Wlotko & Federmeier, 2012; Morgan, Brothers, & Kuperberg, under review), we presented each of the 512 critical words at four levels of probability: 99%, 50%, 25% and uniform (1/[total words] = 1/1579 = 0.06%). As expected, there was a graded, inverse relationship between word probability and the magnitude of lexico-semantic prediction error.

[7] As described earlier, each state unit in this layer represented a word in the model's lexicon. Therefore, an activity pattern over units in this layer approximated a probability distribution over words.

Critically, similar to the sentence-level models implemented by RHM18 and Fitz & Chang (2019), prediction errors were only sensitive to the cloze probability of the input, and not to the *constraint* of the prior context (cf. Kutas & Hillyard, 1984; Federmeier, Wlotko, De Ochoa-Dewald & Kutas, 2007; Kuperberg, Brothers, & Wlotko, 2020). In other words, unexpected words generated the same simulated N400 response when they violated a strong prior prediction (99% constraint) as when no particular word was highly predictable (uniform pre-activation).

Similar to RHM18, we also showed that the simulated N400 was sensitive to the degree of semantic overlap between an expected input and an otherwise low-probability target word. Moreover, we further extended this simulation by showing that the semantic overlap effect was larger in sentence contexts that were more *versus* less constraining, again mirroring the empirical findings (Federmeier & Kutas, 1999).

Another novel contribution of our approach was the simulation of orthographic overlap effects during sentence comprehension (Laszlo & Federmeier, 2009; Ito, Corley, Pickering, Martin & Nieuwland, 2016; DeLong, Chan & Kutas, 2019). When one word was highly predictable (*wish*), violations that were orthographically related to this expected completion produced less lexico-semantic prediction error than unrelated inputs (F-R-O-G), regardless of whether the violating input was a word (D-I-S-H) or a pseudoword (W-U-S-H*;* cf. Laszlo & Federmeier, 2009). Taken together, these effects of contextual predictability and overlap suggest that the magnitude of lexico-semantic prediction error is directly linked to the amount of new, unpredicted information carried by the bottom-up input, regardless of whether this information is orthographic, lexical, or semantic.

Interactions between bottom-up lexical variables and top-down context

In addition to these main effects, our simulations also captured interactions between lexical and contextual factors. Similar to Cheyette & Plaut (2017) and Rabovsky & McRae (2014), we showed that repetition priming was modulated by both frequency (cf. Rugg, 1990) and semantic richness (cf. Kounios & Holcomb, 1994). We further extended these findings by successfully simulating the reduced effects of frequency and semantic richness on high cloze (*versus* low cloze) continuations, again mirroring the empirical findings (Cloze x Frequency: Dambacher, Kliegl, Hofmann, & Jacobs, 2006; Cloze x Richness: Holcomb, Kounios, Anderson, & West, 1999). In the case of word frequency, differences in feedback weights imposed an implicit prior on the model that reduced the lexico-semantic prediction error produced by more frequent words. However, repetition and cloze probability were able to override this prior, such that high and low frequency words produced equally small lexico-semantic prediction errors. In the case of semantic richness, words with a greater number of semantic features carried additional unpredicted semantic information, but, again, this difference was eliminated when these words were repeated or strongly pre-activated by prior top-down predictions.

Insights & Limitations

Our model demonstrated, for the first time, that the basic principles of predictive coding – a general computational theory of brain function – can successfully explain key properties of the N400 response, an important neural signature of language comprehension. As noted above, we used a predictive coding algorithm and architecture that was originally developed to explain low-level phenomena in the visual system (Spratling, 2012, 2013, 2014, 2016). The fact that the same algorithm and architecture was able to simulate effects on the N400 suggests that the same computational principles that are employed in other perceptual and cognitive domains may also

support language processing. We now consider the insights and limitations of our model in relation to the four major constraints that we discussed in Section 4.

1. Explaining a wide range of empirical phenomena

We first argued that any successful model of the N400 should be able to simulate a wide range of empirical findings, using a single outcome measure, with few additional assumptions. In this regard, our model did impressively well. The structure of our model was largely *prescribed* by the principles of predictive coding, both in terms of the precise connectivity between units, and the stages of the algorithm itself. In this sense, our model was relatively constrained. Yet even with these constraints, the model was able to simulate a broad set of N400 effects.

Like the *word-level* models reviewed in Section 2, the model was able to explain how the lexical characteristics of words presented in isolation (semantic richness, lexical frequency, orthographic neighborhood) can influence the N400, and why the amplitude of this component is reduced when words are presented after a repeated or semantically related prime. Moreover, even though the model included explicit lexical representations, the presence of an orthographic layer allowed us to simulate the processing of *non-word* stimuli: Like Laszlo & Plaut (2012), we observed effects of orthographic neighborhood size on *both* word and pseudoword strings. Moreover, unlike previous models, our Predictive Coding model was able to accurately simulate the *larger* N400 response to pseudowords, relative to real words, even when controlling for orthographic neighborhood size (cf. Holcomb, Grainger & O'Rourke, 2002; Meade, Midgley, Dijkstra & Holcomb, 2018; Meade, Grainger & Holcomb, 2019; but see Laszlo & Federmeier, 2011). This is because, in contrast to Laszlo & Plaut (2012), the model was not trained to reduce semantic activation in response to non-words; instead, the pseudowords were free to spread

activation to the semantic representations of their orthographic neighbors, producing a larger error overall.

In addition to these single word effects, the model was able to simulate several sentence-level effects, including the graded effect of contextual predictability on the N400 (similar to RHM18 and Fitz & Chang, 2019). Importantly, the model went beyond RHM18 by not only simulating the effects of *semantic* prediction overlap (cf. Federmeier & Kutas, 1999; Kutas & Hillyard, 1984), but *also* the effects of *orthographic* prediction overlap on the N400 evoked by both words and pseudoword strings during sentence comprehension (see Lazlo and Federmeier, 2009). Taken together, these findings all point to a comprehension system in which the amplitude of the N400 reflects the amount of latent lexico-semantic information encoded in the bottom-up input that has not already been predicted by the system as a whole.

Perhaps the most important success of this computational framework, however, is that it is the first to capture *both* word-level *and* sentence-level manipulations – as well as their interactions –using a single dependent measure: lexico-semantic prediction error. A key reason for this model's success is its hierarchical and interactive architecture, which we turn to next.

2. Incorporation of interactive, hierarchically organized representations

As discussed, a central precept in cognitive and neurobiological models of language processing is the continuous interaction between higher and lower levels of representation. This type of continuous interaction across levels of representation is an inherent component of hierarchical predictive coding (Rao & Ballard, 1999; Mumford, 1992; Friston, 2005). In our Predictive Coding model, each level of representation interacted continuously with the levels above and below, through both top-down and bottom-up connections. More specifically, the top-down effects took two forms: First, higher layers continually generated top-down predictions (or

reconstructions) of states at the level below, which were used by error units to compute prediction error. This prediction error provides the bottom-up input to the layer above, where it is used to update state representations, allowing the model to incrementally converge on the correct interpretation. Second, higher layers provided a top-down bias that could activate and "fill-in" state representations at the level below, making the system robust to perceptual noise or ambiguous inputs (see Mumford, 1992 for discussion). Unlike previous models of the N400, the combination of top-down and bottom-up flow of information across hierarchical levels allowed for direct *interactions* between low-level lexical factors (such as orthographic neighborhood size) and higher-level contextual variables (cloze probability).

As we also discussed, central to many predictive processing models of language comprehension is the idea that the brain can exploit the implicit predictions of event representations, encoded at longer time scales at higher levels of the processing hierarchy. The assumption is that these anticipatory states generate *top-down* predictions that *pre-activate* lexico-semantic representations *before* new bottom-up inputs are encoded at this level (DeLong, Urbach, & Kutas, 2005; Kuperberg & Jaeger, 2016; Wang, Kuperberg, & Jensen, 2018). In our Predictive Coding model, we simulated these temporal *and* hierarchical top-down predictions by pre-activating state units at the top layer of our model and allowing these predictions to filter down to lower-level state units, prior to presenting the model with a bottom-up orthographic input.

One limitation of this current implementation is that we provided this top-down pre-activation to the model through a "dummy" event layer. In reality, we believe the source of this top-down input would correspond to an event state that is incrementally inferred from a preceding sentence context. In order to implement this, the highest level of the model would

need to encode event representations over a longer time span than individual words (analogous to the event-level representations implemented in the models by Brouwer et al. 2017 and RHM18). Note that a lower-level analogue of this situation already exists in our model: If we present one letter at a time to our model, the remaining letters can be pre-activated because lexical units are connected to "future" letter positions that have not yet been encountered. In general, this approach is somewhat analogous to how the speech recognition model TRACE (McClelland & Elman, 1986) includes word units that span longer durations than phoneme units. Therefore, in future versions of this Predictive Coding model, it will be important to incorporate explicit event representations that are built up incrementally over the course of a sentence.

3. Cognitive plausibility: Facilitatory effect of context on facilitation during language comprehension within N400 time window

Predictive coding also captures the central idea, common to all cognitive theories of the N400, that the attenuation of this component by prior context is closely linked to facilitated processing. In our model, regardless of whether inputs were repeated, semantically primed, or higher cloze, they produced a smaller lexico-semantic prediction error (i.e. a smaller simulated N400). Critically, however, in all these cases, despite producing less neural activity within *error units*, the lexico-semantic *state units* were able to converge more quickly on the correct interpretation of the bottom-up input. In other words, despite the reduction in the N400 (driven by the reduction in activity within the error units), the semantic features of the expected words would still be easier to access as the model converged[8]. This stands in contrast to Semantic

[8] The process of converging on a specific *expected* state representation is sometimes referred to as representational "sharpening". Sharpening is a feature of several other classic computational theories of brain function, such as interactive activation architectures (e.g. McClelland & Rumelhart, 1981), adaptive resonance theory (Carpenter & Grossberg, 1987) and biased competition models of top-down attention (Desimone & Duncan, 1995; see also Harpur &

Activation models, where an activity-dependent decay function resulted in *more difficult* semantic retrieval following a related prime (Laszlo & Armstrong, 2014; Cheyette & Plaut, 2017). Our approach was also distinct from previous prediction error accounts, where errors were computed outside the model (Rabovsky & McRae, 2014), or even outside the comprehension system itself (Fitz & Chang, 2019) at a stage when word processing was already complete.

The link between reduced N400 responses and facilitated processing was also captured by RHM18's Sentence Gestalt model in which predictable inputs triggered a smaller shift in an internal event representation. However, as discussed earlier, this model was unable to explain the neuroanatomical sources of the N400 effect. Predictive coding can account for these source-localization data. This is because it attributes the evoked effect of priming and contextual predictability in plausible sentences to differences in the magnitude of prediction error produced by error units at the *lexico-semantic level*. As such, it correctly predicts that semantic priming effects, and, in plausible sentences, predictability effects on the N400 should all localize to regions that support lexico-semantic processing within the left temporal lobe (*semantic priming effects*: Nobre & McCarthy, 1995; Halgren et al., 2006; Lau, Weber, Gramfort, Hämäläinen & Kuperberg; Lau, Gramfort, Hämäläinen & Kuperberg, 2013; *predictability effects in plausible*

Prager, 1994). Early discussions of predictive coding assumed that a reduction of prediction error to expected inputs was incompatible with a sharpening of neural activity to expected inputs (e.g. Murray, Schrater, & Kersten, 2004). However, others pointed out that, because of the functional distinction between "error units" (that compute prediction error at each level of cortical representation) and "state units" (that infer the underlying representation of the input at each level of cortical representation), post-stimulus sharpening is fully compatible with predictive coding (Spratling, 2008; Friston, 2005; Kok & de Lange, 2015; Walsh, McGovern, Clark, & O'Connell, 2020). In other words, expected bottom-up input reduces the magnitude of the prediction error computed by error units, while simultaneously increasing activity produced within state units that encode the expected representation.

sentences: Wang et al., under review). Moreover, predictive coding can also explain why *implausible* words additionally produce a larger evoked response within left inferior cortex within the N400 time-window (e.g. Wang et al., under review; Halgren et al., 2002; Marinkovic et al., 2003; Maess, Herrmann, Hahne, Nakamura, & Friederici, 2006; Ihara, Hayakawa, Wei, Munetsuna, & Fujimaki, 2007; Pylkkänen & McElree, 2007). According to this framework, higher-level prediction error at the event level is only produced by inputs whose statistics strongly deviate from the statistics of natural environmental inputs (see Rao & Ballard, 1999), i.e. when top-down predictions based on stored real-world knowledge fail to suppress event-level prediction error produced within left inferior frontal cortex.

4. Biological plausibility: An explicit and biologically plausible linking function

We also emphasized that a comprehensive model of the N400 should provide an explanation for *why* its particular linking function would give rise to neural activity at the scalp's surface. In the case of predictive coding, the linking function is clear: prediction errors are computed as an inherent component of semantic inference; that is, when inputs are unexpected, this leads to increased activity in error units, and therefore increased post-synaptic potentials, leading to a larger N400.

Moreover, like previous Semantic Activation accounts (LPAC), our Predictive Coding model was able to simulate the characteristic rise-and-fall trajectory of the N400. As pointed out by Friston (2005), the basic morphology of event-related potentials is intuitively explained by the activation dynamics of the predictive coding algorithm. First, prediction errors rise as new inputs are presented to the system that cannot be explained by predictions/reconstructions at the level above. Then, gradually, these errors begin to fall as higher level regions update their

representations to generate more accurate reconstructions, thereby suppressing lower-level prediction errors. In our model, this updating process always required a relatively fixed number of iterations, which may explain the relatively constant latency of the N400 response.

At a more general level, there is a growing body of evidence that the computations performed in predictive coding can, in principle, be implemented at the level of cortical microcircuits (see Bastos et al., 2012; Shipp, 2016). Moreover, studies using single cell recordings (Bell et al., 2016; Fiser et al., 2016) and fMRI (de Gardelle et al., 2012, 2013) provide evidence for functionally distinct populations of state and error units. Indeed, consistent with the theory that distinct lexico-semantic error and state units are differentially activated by expected and unexpected inputs during sentence comprehension, in a recent MEG study, we were able to detect spatially distinct patterns of neural activity within the left temporal lobe to expected and unexpected words between 300-500ms (Wang, Nour-Eddine, Brothers, Jensen & Kuperberg, in preparation).

A limitation of our current Predictive Coding implementation is that it did not include a learning procedure. Instead, we provided the model with a fixed set of connection weights. Therefore, unlike the models implemented by RHM18 and Fitz and Chang (2019), we were unable to simulate the interaction between predictability and *longer-term* repetition effects on the N400 (Besson, Kutas & Van Petten, 1992; Rommers & Federmeier, 2018). It is important to note, however, that this is not an inherent limitation of predictive coding more generally. In principle, the prediction error computed during predictive coding can function not only to perform perceptual *inference*, but also to gradually update connection weights for longer-term *learning* (see Rao and Ballard, 1999). Under such circumstances, error units would function to

minimize prediction errors *both* in the short-term (through inference) and in the longer term (by updating the model's weights).

Indeed, under certain conditions, the prediction errors computed during predictive coding can be shown to converge to the learning signal that is backpropagated in connectionist networks (Whittington & Bogacz, 2017; Millidge, Tschantz & Buckley, 2020; Song, Lukasiewicz, Xu, & Bogacz, 2020). However, unlike backpropagation, where updates require information from all downstream neurons, weight updates in predictive coding rely exclusively on locally generated error. Therefore, predictive coding approaches may provide a biologically plausible alternative to backpropagation. Incorporating long-term learning will be an important goal for future iterations of our Predictive Coding model, serving to highlight the important links between prediction error and learning, not just during language production (as highlighted by Fitz and Chang, 2019), but also during language comprehension (e.g. Elman, 1990; Kleinschmidt & Jaeger, 2015; Fine, Jaeger, Farmer & Qian, 2013; Fine & Jaeger, 2016; Myslin & Levy, 2016).

Section 6: Future Directions

In this review, we have discussed several computational models that have deepened our understanding of the N400. Over time, these models have shifted from exploring lower-level lexical phenomena and simple priming effects, to the effects of higher-level context on the N400 produced during sentence comprehension. Each of these models provided unique insights, which often built upon each other over time. For example, our recent Predictive Coding model (Section 5) incorporates many aspects of prior approaches, including: a) a biologically plausible linking function (Laszlo and Plaut, 2012), b) a central role for prediction error in generating the N400 response (Rabovsky & McRae, 2014; Fitz & Chang, 2019), c) a hierarchical structure that

separates lexico-semantic and higher-level event states (Brouwer et al., 2017), and, d) the core assumption that words function as "cues to meaning" that induce shifts in the model's internal state (RHM18). Despite this progress, there is obviously much work to be done. In this final section, we discuss some further challenges for existing models, while recommending additional directions for future research.

In addition to the large literature documenting the effects of lexical and contextual factors on the N400, there is also an extensive *behavioral* literature that documents the effects of these factors on the speed and accuracy of word identification (see Rayner, 1998; Rastle, 2016 for reviews). Therefore, a critical challenge for future models will be to simulate the effects of these factors on *both* the N400 and behavior, with minimal additional assumptions.

We believe that one promising approach for simulating the effects of context on behavior is to adopt an evidence accumulation/sequential sampling framework (cf. Ratcliff, 1978; Usher & McClelland, 2001; Forstmann, Ratcliff & Wagenmakers, 2016). In this framework, lexico-semantic units associated with different words can accumulate activation over time, triggering behavioral responses through a boundary-crossing decision criterion[9]. This evidence accumulation approach has been successfully used to model the effects of context on cloze response times through a race model (Staub, Grant, Astheimer & Cohen, 2015), and has also

[9] We note that Cheyette and Plaut (2017) attempted to simulate certain effects on both the N400 and behavior (lexical decision). However, although the semantic units accumulated activation, the lexical decision output itself was not based on whether or not this accumulated value passed a threshold. Instead, what determined this behavioral response was a non-linear combination of the instantaneous (decayed) and accumulated (undecayed) activation. As discussed in Section 2, this approach raises several concerns. We also note that, in their original model, Laszlo & Plaut (2012) did simulate lexical decision based on a threshold-crossing criterion (consistent with neurally plausible decision-making). However, there was no attempt to simulate the effects of lexical variables on lexical decision times.

been used to model the effects of context on eye movement decisions during reading (Bicknell & Levy, 2010). Importantly, this approach is biologically plausible: There is strong evidence from primate neurophysiology that neurons can accumulate evidence that guides behavioral responses (see Gold & Shadlen, 2007 for a review). Although at the time of writing we have not yet attempted to simulate the behavioral effects of context, we believe that our Predictive Coding model offers a particularly promising framework for simulating contextual effects on *both* the N400 and behavior. Specifically, while the rise and fall of lexical and semantic *error* activity (prediction error) can explain the rise and fall of the N400, lexical and semantic *state units* accumulate information about the identity of a specific input. Thus, a simple activation threshold on this accumulated state activation could provide a method for simulating effects of context on measures such as word recognition times.

We should emphasize, however, that simulating both neural and behavioral effects using the same computational model is far from trivial. Although many *contextual* factors lead to both reductions in the N400 and reductions in reaction times, other variables can influence the N400 and behavior in opposite directions. For example, semantic concreteness leads to an increase in N400 amplitude but shorter reaction times (Kounios & Holcomb, 1994). Moreover, in addition to the influence of linguistic factors, the link between lexico-semantic activation and behavior will also depend on several other factors, including the nature of task (e.g. lexical decision *versus* semantic categorization) and higher-order utility functions (e.g. speed-accuracy tradeoffs). Ultimately, future computational models may need to adopt a *decision theoretic* approach in which behavioral responses are guided by a *decision variable* that accumulates lexical and/or semantic evidence that is specifically relevant for the comprehender's current goals (cf. Gold & Shadlen, 2007).

Another important goal for future work will be to incorporate more realistic *sublexical* representations into models of the N400. Currently, computational models either *omit* sublexical representations altogether (Brouwer et al., 2017; RHM18; Fitz & Chang, 2019), or rely on a "slot-based" coding scheme that maps directly from orthography to semantics (Semantic Activation models, Predictive Coding[10]). Although these slot-based schemes are simple to implement, they are inconsistent with our current knowledge of orthographic representations in the brain, and they are unable to account for transposed-letter priming effects on behavior and the N400 (Grainger, 2008; Grainger, Kiyonaga & Holcomb, 2006; Carreiras, Vergara, & Perea, 2009).

Even more importantly, no current model of the N400 has yet incorporated *phonological* representations, which are thought to be co-activated in parallel with orthographic representations, even during silent reading (Seidenberg & McClelland, 1989; Harm & Seidenberg, 2004; Grainger & Holcomb, 2009). Indeed, several empirical N400 phenomena can only be explained by appealing to an "indirect pathway" that maps from orthography to semantics *via* phonological representations. For example, the phonological neighborhood of a visually-presented word can also modulate the N400 (Carrasco-Ortiz, Midgley, Grainger & Holcomb, 2017), and larger N400 priming effects are observed on targets that are phonologically related to pseudo-homophone primes (*brane* – *brain*) compared to orthographically matched control primes (*brans* – *brain*; Grainger, Kiyonaga & Holcomb, 2006). Without phonological representations, current models of the N400 produced during reading are unable to account for these results.

[10] In Rabovsky & McRae (2014), the input representations did not strictly correspond to either orthography or phonology; the authors did not specify the relationship between the input representation and the word it represented.

While previous computational models of the N400 have focused on visual word processing, spoken language comprehension arguably plays an even more important role in day-to-day communication. Empirically, spoken words are known to produce a robust N400 response, which is sensitive to many of the same lexical and contextual factors discussed in this review (e.g. Diaz & Swaab, 2007; Winsler, Midgley, Grainger & Holcomb 2018). However, because auditory inputs unfold sequentially, phoneme by phoneme, studies of spoken language comprehension have also revealed additional N400 phenomena that are specific to online speech comprehension (e.g. cohort *versus* rhyme overlap effects, Van Petten, Coulson, Rubin, Plane, & Parks, 1999). These findings could serve as important benchmarks for future computational models of spoken word comprehension.

In future computational models, it will also be important to simulate earlier language-related ERP components (evoked between 150-300ms) that are thought to support sublexical processing, including the N170 (Bentin, Mouchetant-Rostaing, Giard, Echallier, & Pernier, 1999) and the N250 (Holcomb & Grainger, 2006; Kiyonaga, Midgley & Holcomb, 2007; Grainger & Holcomb, 2009). These early negative-going ERP components have been linked to the activation of sublexical orthographic or phonological information in a hierarchical interactive activation framework (Grainger & Holcomb, 2009). As such, hierarchical models, like predictive coding, may be particularly well-suited for simulating these early ERP responses. For example, just as the N400 is simulated as prediction error at the level of lexico-semantic features, the N250 could be simulated as lower-level prediction error at the level of orthographic form.

By the same token, it will also be important to extend future computational models to simulate ERP components beyond the N400 time-window. Two previous computational models (Brouwer et al., 2017; Fitz & Chang, 2019) have attempted to model the P600 – a late

posteriorly-distributed positivity observed between 600-1000ms following the onset of syntactic (Osterhout & Holcomb, 1992; 1993; Hagoort, Brown, & Groothusen, 1993) and semantic (Kuperberg, 2007) violations. However, both these modeling approaches had limitations. In Brouwer et al's Retrieval-Integration model, the P600 was simulated as a shift in state at the higher event level, which is nearly identical to the proposed linking function for the N400 in RHM's Sentence Gestalt model. Because of this assumed linking function for the P600, the Retrieval-Integration model is unable to explain why linguistic anomalies that are relatively uninformative at the event-level (and that do not produce a large N400) can still produce robust P600 responses (e.g., agreement errors). In Fitz & Chang's Error Propagation model, although the P600 construct was able to capture both syntactic and semantic anomalies, this component was always dependent on a preceding N400 response. Specifically, because it was defined as the backpropagated error signal, it was impossible for the P600 construct to take on a large value if the N400 (reflecting the total absolute error) was minimal. This assumption prevented the model from being able to correctly simulate the effects of factors that produce a large P600 effect but minimal differences on the N400 (e.g., the effects of semantic reversal anomalies).

Ultimately, we believe that any computational model of the P600 must incorporate the sensitivity of this component to multiple types of linguistic errors, as well as the fact that this component is enhanced when comprehenders are engaged in deep comprehension; that is, they have established a prior "situation model" (for recent discussion, see Brothers, Wlotko, Warnke, & Kuperberg, 2020). We have suggested that the posterior P600 is evoked when comprehenders are initially *unable* to integrate a word into their prior situation model, resulting in a reallocation of attention and reprocessing of the input (Kuperberg, Brothers, & Wlotko, 2020; Brothers, Wlotko, Warnke, & Kuperberg, 2020; Alexander, Brothers, & Kuperberg, Under review). In a

predictive coding framework, this late-stage reprocessing could occur when the model fails to converge, resulting in the continued generation of inaccurate top-down predictions, which fail to switch off prediction error produced at lower levels of linguistic representation (Wang et al., under review).

In addition, it will be important for future models to distinguish between the posteriorly distributed P600 effect, which is produced by highly *implausible* or *anomalous* inputs, and a distinct late frontally-distributed positivity that is triggered by unexpected but *plausible* inputs (Federmeier et al., 2007; Van Petten & Luka, 2012). In recent studies, we have argued that this late frontal positivity component is produced when readers *are* able to successfully update or revise their prior situation model in response to new, unexpected information (Kuperberg, Wlotko & Brothers, 2020; Brothers, Wlotko, Warnke & Kuperberg, 2020). In a predictive coding framework, this late-stage activity could be driven by the generation of *new* top-down predictions that are passed down the hierarchy, updating event-level representations and activating lexico-semantic information, in order to facilitate the processing of new upcoming inputs (Wang et al., under review).

Finally, while previous computational models have focused on N400 effects in young, healthy monolingual readers, these models could also provide important insights into how the N400 is modulated in different populations. For example, one exciting avenue of research will be to develop N400 models of bilingual language processing. A central question in this field is how the bilingual lexicon is organized, and whether cross-linguistic representations are activated automatically during online comprehension. For example, in bilingual studies of sentence processing, a sudden switch between languages results in an increase in the N400, which has been taken as evidence for top-down suppression of the non-target language (Moreno,

Federmeier & Kutas, 2002). However, other ERP studies have shown a surprising degree of permeability ("non-selectivity") across languages. For example, English words produce a larger N400 when they have more orthographic neighbors in a bilingual's second language (e.g. French: Midgley, Holcomb, Walter & Grainger, 2008). In addition, bilingual readers show N400 facilitation when they process *cognate* words that have the same form and meaning across languages (relative to frequency-matched controls; Midgley, Holcomb, Grainger, 2011; Peeters, Dijkstra, Grainger, 2013). Accounting for these results within a single model of the N400 will be an important direction for future research.

Similarly, computational models of the N400 can also play an important role in informing our understanding of different clinical disorders characterized by language processing disturbances. In previous work, researchers have applied artificial "lesions" to simple connectionist models of word reading in order to account for the specific pattern of behavioral deficits in certain language disorders (e.g., Plaut & Shallice, 1993a; Dell, Schwartz, Martin, Saffran, & Gagnon, 1996; Plaut & Shallice, 1993b). A similar approach could be taken to explain abnormalities in the N400 across different language disorders including aphasia (Swaab, Brown & Hagoort, 1997), dementia (Iragui, Kutas & Salmon, 1996), and dyslexia (Rüsseler, Becker, Johannes & Münte, 2007). In addition, both autism and schizophrenia are characterized by disturbances in language processing (American Psychiatric Association, 2013), and researchers have theorized that deficits in predictive coding may play a central role in both these disorders (*autism*: van Boxtel & Lu, 2013; Van de Cruys, Evers, Van der Hallen, Van Eylen, Boets, de-Wit & Wagemans, 2014; *schizophrenia:* Griffin & Fletcher, 2017; Sterzer, Adams, Fletcher, Frith, Lawrie, Muckli, Petrovic, Uhlhaas, Voss & Corlett, 2018; Corlett, Taylor, Wang, Fletcher & Krystal, 2010 ; see also Brown & Kuperberg, 2015). Future computational models

can explore whether these language processing deficits are directly linked to abnormalities in predictive coding mechanisms.

In conclusion, modeling ERP components like the N400 offers a unique opportunity to bridge across the algorithmic and implementational levels of analysis (cf. Marr, 1982). In this review, we have described the strengths and limitations of different computational theories, highlighting the unique insights that each modeling framework has contributed to our understanding of language comprehension and the N400. More generally, this body of work demonstrates how basic neurophysiological principles can inform our cognitive theories, as well as how empirical phenomena in the neural and behavioral literatures can constrain our theories about cortical organization. Given the clear progress the field has made over the past decade, we predict that future iterations of these computational models will provide an even more precise understanding of semantic processing in the human brain.

Acknowledgments

This work was funded by the National Institute of Child Health and Human Development (2R01 HD082527 to G.R.K.). We are grateful for Jeff Stibel for his support of Samer Nour Eddine and Gina Kuperberg. We thank Kara Federmeier and Lin Wang for their insightful comments on the manuscript. We also thank Arim Choi Perrachione for her assistance with manuscript formatting.

Figure Legends

Figure 1. The neural network architectures of two "word-level" models of the N400.

* is used to indicate layers/representations that were used to compute the N400.

A. Semantic Activation Model. The particular architecture depicted here is adapted from Cheyette & Plaut, 2017, Figure 3. The N400 was operationalized as the total semantic activation at the Semantics layer. The architectures used by Laszlo & Plaut (2012) and Laszlo & Armstrong (2014) were very similar, but had random, fixed outgoing connections from the inhibitory (INH) units (i.e., the inhibitory weights were not learned), and no clean-up layer associated with the Hidden layer.

B. Semantic Attractor Model. The architecture depicted here is adapted from Rabovsky & McRae, 2014, Figure 1. The N400 was operationalized as the cross-entropy between the model's semantic output and a target semantic representation defined by the modeler.

Arrows mapping from one layer to another indicate connection weights that are learned during training. *Pink ovals* indicate layers encoding distributed semantic features. *Yellow ovals* indicate layers encoding word-form features. *Gray ovals* indicate layers whose encoded representations do not have a clear linguistic interpretation. *Ovals with solid outlines* indicate Input or Output layers. *Ovals with dashed outlines* indicate Hidden layers.

Figure 2. The neural network architectures of three "sentence-level" models of the N400.

* is used to indicate layers/representations that were used to compute the N400.

A. Retrieval-Integration Model. The depicted architecture is adapted from Brouwer et al., 2017, Figure 2. The architecture consists of a Retrieval Module (bottom) and an Integration Module (top) whose weights were trained separately (see text). The N400 was operationalized as the degree of shift in activation at the Semantic Retrieval layer after processing a given input.

B. Sentence Gestalt Model. The depicted architecture is adapted from Rabovsky, Hansen & McClelland, 2018, Figure 1A. The architecture consists of an Update Network (left) and a Query Network (right). The N400 was operationalized as the degree of shift in activation at the Sentence Gestalt layer after processing a given input.

C. Error Propagation Model. The depicted architecture is adapted from Fitz & Chang, 2019, Figure 9. The architecture consists of a Sequencing System (left) and a Message System (right). The Sequencing System processed a sequence of lexical inputs (encoded in the PrevWord and PrevWordHistory layers) to predict the next word. The N400 was operationalized as the difference (absolute value) between predicted activity at the NextWord layer and the observed target word defined by the modeler. The Message System encoded events in the fast-changing links between the Role and Concept layer (and equivalently, CConcept to CRole), indicated by a thick green-pink line; the EventSemantics layer (in blue) encoded the relative prominence of different arguments in the sentence.

Solid arrows mapping from one layer to another indicate connection weights that are learned during training. *Dashed arrows* denote a "copy" operation. *Black ovals* indicate layers encoding lexical information. *Pink ovals* indicate layers encoding distributed semantic features. *Gray ovals* indicate layers whose encoded representations do not have a clear linguistic interpretation. *Ovals with solid outlines* indicate Input or Output layers. *Ovals with dashed outlines* indicate Hidden layers.

Figure 3. Predictive Coding Model.

A. Architecture. Each level of representation has a distinct population of state and error units (except for the highest level). State units pass a copy of their state activity (St) to the error units at the same level (dashed arrows), and send a top-down Prediction (Pr) to the level below (blue

arrows). Error units compute two quantities: Prediction Error (PE) and top-down Bias (tdB). Prediction Error is computed at each level by dividing the state at that level by the top-down Prediction, and this is passed up to update the state at the level above (red arrows). Top-down Bias is the reciprocal of this, computed by dividing the top-down prediction by the state at that level, and the result is copied to the state units at that level (dashed arrows). At each iteration, the N400 is operationalized as the sum of the Prediction Error at the lexical and semantic levels (indicated with an asterisk).

Solid arrows denote a hand-coded mapping from one layer to another. *Dashed arrows* denote a "copy" operation. *Ovals* denote state units. *Half-arches* on top of the ovals denote error units. Color is used to denote the level of representation, with *yellow* denoting orthographic form, *black* denoting lexical, and *pink* denoting the semantic level. The dummy event state units are indicated with a *gray oval*.

B. The time course of the lexico-semantic Prediction Error (simulated N400) produced in two simulations, exploring the influence of lexical frequency (top) and cloze probability (bottom). High frequency words elicited a smaller prediction error than low frequency words; and Prediction Error was inversely graded with contextual predictability. Shading represents ±1 standard error of the mean across items.

References

Alexander, E., Brothers, T., & Kuperberg, G. R. (Under review). The P600 reflects reanalysis – not error correction.

Allen, M., Badecker, W., & Osterhout, L. (2003). Morphological analysis in sentence processing: An ERP study. *Language and Cognitive Processes, 18*(4), 405-430.

Altmann, G. T., & Kamide, Y. (1999). Incremental interpretation at verbs: restricting the domain of subsequent reference. *Cognition, 73*(3), 247-264.

American Psychiatric Association. (2013). *DSM-V: Diagnostic and Statistical Manual of Mental Disorders* (5th ed.). Arlington, VA: American Psychiatric Publishing.

Amsel, B. D. (2011). Tracking real-time neural activation of conceptual knowledge using single-trial event-related potentials. *Neuropsychologia, 49*(5), 970-983.

Baggio, G., & Hagoort, P. (2011). The balance between memory and unification in semantics: A dynamic account of the N400. *Language and Cognitive Processes, 26*(9), 1338-1367.

Bastos, A. M., Usrey, W. M., Adams, R. A., Mangun, G. R., Fries, P., & Friston, K. J. (2012). Canonical microcircuits for predictive coding. *Neuron, 76*(4), 695-711.

Bell, A. H., Summerfield, C., Morin, E. L., Malecek, N. J., & Ungerleider, L. G. (2016). Encoding of stimulus probability in macaque inferior temporal cortex. *Curr Biol, 26*(17), 2280-2290.

Bentin, S. (1987). Event-related potentials, semantic processes, and expectancy factors in word recognition. *Brain and Language, 31*(2), 308-327.

Bentin, S., McCarthy, G., & Wood, C. C. (1985). Event-related potentials, lexical decision and semantic priming. *Electroencephalography and Clinical Neurophysiology, 60*, 343-355.

Bentin, S., Mouchetant-Rostaing, Y., Giard, M. H., Echallier, J. F., & Pernier, J. (1999). ERP manifestations of processing printed words at different psycholinguistic levels: time course and scalp distribution. *Journal of Cognitive Neuroscience, 11*(3), 235-260.

Besson, M., Kutas, M., & Van Petten, C. (1992). An event-related potential (ERP) analysis of semantic congruity and repetition effects in sentences. *J Cogn Neurosci, 4*(2), 132-149.

Bicknell, K., & Levy, R. (2010). *A rational model of eye movement control in reading.* Paper presented at the Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL '10), Uppsala, Sweden.

Blank, H., & Davis, M. H. (2016). Prediction errors but not sharpened signals simulate multivoxel fMRI patterns during speech perception. *PLoS Biology, 14*(11), e1002577.

Bornkessel-Schlesewsky, I., & Schlesewsky, M. (2019). Toward a neurobiologically plausible model of language-related, negative event-related potentials. *Front Psychol, 10*, 298.

Brothers, T., Wlotko, E. W., Warnke, L., & Kuperberg, G. R. (2020). Going the extra mile: Effects of discourse context on two late positivities during language comprehension. *Neurobiology of Language, 1*(1), 135-160.

Brouwer, H., Crocker, M. W., Venhuizen, N. J., & Hoeks, J. C. J. (2017). A neurocomputational model of the N400 and the P600 in language processing. *Cogn Sci, 41 Suppl 6*, 1318-1352.

Brown, M., & Kuperberg, G. R. (2015). A hierarchical generative framework of language processing: Linking language perception, interpretation, and production abnormalities in schizophrenia. *Frontiers in Human Neuroscience, 9*, 643.

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods, 41*, 977-990.

Bugmann, G. (1997). Biologically plausible neural computation. *Biosystems, 40*(1-2), 11-19.

Camblin, C. C., Gordon, P. C., & Swaab, T. Y. (2007). The interplay of discourse congruence and lexical association during sentence processing: Evidence from ERPs and eye tracking. *Journal of Memory and Language, 56*(1), 103-128.

Carpenter, G. A., & Grossberg, S. (1987). A massively parallel architecture for a self-organizing neural pattern recognition machine. *Computer Vision, Graphics, and Image Processing, 37*(1), 54-115.

Carrasco-Ortiz, H., Midgley, K. J., Grainger, J., & Holcomb, P. J. (2017). Interactions in the neighborhood: Effects of orthographic and phonological neighbors on N400 amplitude. *J Neurolinguistics, 41*, 1-10.

Carreiras, M., Vergara, M., & Perea, M. (2009). ERP correlates of transposed-letter priming effects: the role of vowels versus consonants. *Psychophysiology, 46*(1), 34-42.

Chang, F., Dell, G. S., & Bock, J. K. (2006). Becoming syntactic. *Psychological Review, 113*(2), 234-272.

Chen, Q., & Mirman, D. (2012). "Competition and cooperation among similar representations: Toward a unified account of facilitative and inhibitory effects of lexical neighbors": Correction to Chen and Mirman (2012). *Psychological Review, 119*(4), 898-898.

Cheyette, S. J., & Plaut, D. C. (2017). Modeling the N400 ERP component as transient semantic over-activation within a neural network model of word comprehension. *Cognition, 162*, 153-166.

Chow, W. Y., Lau, E. F., Wang, S., & Phillips, C. (2018). Wait a second! delayed impact of argument roles on on-line verb prediction. *Lang Cogn Neurosci, 33*(7), 803-828.

Chow, W. Y., Smith, C., Lau, E. F., & Phillips, C. (2016). A "bag-of-arguments" mechanism for initial verb predictions. *Lang Cogn Neurosci, 31*(5), 577-596.

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

Corlett, P. R., Taylor, J. R., Wang, X. J., Fletcher, P. C., & Krystal, J. H. (2010). Toward a neurobiology of delusions. *Progress in Neurobiology, 92*(3), 345-369.

Cree, G. S., McNorgan, C., & McRae, K. (2006). Distinctive features hold a privileged status in the computation of word meaning: Implications for theories of semantic memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32*(4), 643-658.

Crick, F. H. C., & Asanuma, C. (1986). Certain aspects of the anatomy and physiology of the cerebral cortex. In J. L. McClelland, D. E. Rumelhart, & PDP Research Group (Eds.), *Parallel Distributed Processing: Explorations in the Microstructures of Cognition, Vol 2: Psychological and Biological Models* (pp. 333-371). Cambridge, MA: MIT Press.

Dambacher, M., Kliegl, R., Hofmann, M., & Jacobs, A. M. (2006). Frequency and predictability effects on event-related potentials during reading. *Brain Research, 1084*(1), 89-103.

de Gardelle, V., Stokes, M., Johnen, V. M., Wyart, V., & Summerfield, C. (2013). Overlapping multivoxel patterns for two levels of visual expectation. *Front Hum Neurosci, 7*, 158.

de Gardelle, V., Waszczuk, M., Egner, T., & Summerfield, C. (2013). Concurrent repetition enhancement and suppression responses in extrastriate visual cortex. *Cerebral Cortex, 23*(9), 2235-2244.

Dehaene, S., Cohen, L., Sigman, M., & Vinckier, F. (2005). The neural code for written words: A proposal. *Trends in Cognitive Sciences, 9*(7), 335-341.

Dell, G. S., & Chang, F. (2014). The P-chain: relating sentence production and its disorders to comprehension and acquisition. *Philosophical Transactions of the Royal Society B: Biological Sciences, 369*(1634), 20120394.

Dell, G. S., Schwartz, M. F., Martin, N., Saffran, E. M., & Gagnon, D. A. (1996). A Connectionist Model of Naming Errors in Aphasia. In J. A. Reggia, E. Ruppin, & R. S. Berndt (Eds.), *Neural Modeling of Brain and Cognitive Disorders* (pp. 135-156). Singapore: World Scientific.

DeLong, K. A., Chan, W. H., & Kutas, M. (2019). Similar time courses for word form and meaning preactivation during sentence comprehension. *Psychophysiology, 56*(4), e13312.

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

Desimone, R., & Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annu Rev Neurosci, 18*, 193-222.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Diaz, M. T., & Swaab, T. Y. (2007). Electrophysiological differentiation of phonological and semantic integration in word and sentence contexts. *Brain Research, 1146*, 85-100.

Dikker, S., & Pylkkänen, L. (2013). Predicting language: MEG evidence for lexical preactivation. *Brain and Language, 127*(1), 55-64.

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

Elman, J. L. (2009). On the meaning of words and dinosaur bones: Lexical knowledge without a lexicon. *Cognitive Science, 33*, 1-36.

Erickson, T. D., & Mattson, M. E. (1981). From words to meaning: A semantic illusion. *Journal of Verbal Learning and Verbal Behavior, 20*, 540–551.

Federmeier, K. D. (2007). Thinking ahead: the role and roots of prediction in language comprehension. *Psychophysiology, 44*(4), 491-505.

Federmeier, K. D. (2022). Connecting and considering: Electrophysiology provides insights into comprehension. *Psychophysiology, 59*(1), e13940.

Federmeier, K. D., & Kutas, M. (1999). A rose by any other name: Long-term memory structure and sentence processing. *J Mem Lang, 41*(4), 469-495.

Federmeier, K. D., & Laszlo, S. (2009). Time for meaning: Electrophysiology provides insights into the dynamics of representation and processing in semantic memory. *Psychology of Learning and Motivation, 51*, 1-44.

Federmeier, K. D., Wlotko, E. W., De Ochoa-Dewald, E., & Kutas, M. (2007). Multiple effects of sentential constraint on word processing. *Brain Research, 1146*, 75-84.

Ferreira, F. (2003). The misinterpretation of noncanonical sentences. *Cognitive Psychology, 47*, 164–203.

Fine, A. B., & Jaeger, T. F. (2016). The role of verb repetition in cumulative structural priming in comprehension. *J Exp Psychol Learn Mem Cogn, 42*(9), 1362-1376.

Fine, A. B., Jaeger, T. F., Farmer, T. A., & Qian, T. (2013). Rapid expectation adaptation during syntactic comprehension. *PLoS One, 8*(10), e77661.

Fiser, A., Mahringer, D., Oyibo, H. K., Petersen, A. V., Leinweber, M., & Keller, G. B. (2016). Experience-dependent spatial expectations in mouse visual cortex. *Nat Neurosci, 19*(12), 1658-1664.

Fitz, H., & Chang, F. (2019). Language ERPs reflect learning through prediction error propagation. *Cogn Psychol, 111*, 15-52.

Forster, K. I. (1979). Levels of processing and the structure of the language processor. In W. E. Cooper & E. C. T. Walker (Eds.), *Sentence processing: Psycholinguistic studies presented to Merrill Garrett*. Cambridge: Mass.: MIT Press.

Forster, K. I. (1981). Priming and the effects of sentence and lexical contexts on naming time: Evidence for autonomous lexical processing. *Quarterly Journal of Experimental Psychology. A: Human Experimental Psychology, 33*(4), 465-495.

Forstmann, B., Ratcliff, R., & Wagenmakers, E. (2016). Sequential sampling models in cognitive neuroscience: Advantages, applications, and extensions. *Annual Review of Psychology, 67*, 641-666.

Friederici, A. D., Hahne, A., & Mecklinger, A. (1996). Temporal structure of syntactic parsing: Early and late event-related brain potential effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22*(5), 1219-1248.

Friston, K. J. (2005). A theory of cortical responses. *Philos Trans R Soc Lond B Biol Sci, 360*(1456), 815-836.

Gibson, E., Bergen, L., & Piantadosi, S. T. (2013). Rational integration of noisy evidence and prior semantic expectations in sentence interpretation. *Proceedings of the National Academy of Sciences, USA, 110*(20), 8051-8056.

Gold, J. I., & Shadlen, M. N. (2007). The neural basis of decision making. *Annu Rev Neurosci, 30*, 535-574.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*: MIT Press.

Grainger, J. (2008). Cracking the orthographic code: An introduction. *Language and Cognitive Processes, 23*(1), 1-35.

Grainger, J., & Holcomb, P. J. (2009). Watching the word go by: On the time-course of component processes in visual word recognition. *Lang Linguist Compass, 3*, 128-156.

Grainger, J., Kiyonaga, K., & Holcomb, P. J. (2006). The time course of orthographic and phonological code activation. *Psychological Science, 17*(12), 1021-1026.

Griffin, J. D., & Fletcher, P. C. (2017). Predictive processing, source monitoring, and psychosis. *Annu Rev Clin Psychol, 13*, 265-289.

Grill-Spector, K., Henson, R., & Martin, A. (2006). Repetition and the brain: neural models of stimulus-specific effects. *Trends in Cognitive Sciences, 10*(1), 14-23.

Grossberg, S. (1987). Competitive learning: From interactive activation to adaptive resonance. *Cognitive Science, 11*(1), 23-63.

Hagoort, P., Baggio, G., & Willems, R. M. (2009). Semantic unification. In M. S. Gazzaniga (Ed.), *The Cognitive Neurosciences, 4th Ed.* (pp. 819-836). Cambridge, MA: MIT Press.

Hagoort, P., Brown, C., & Groothusen, J. (1993). The syntactic positive shift (SPS) as an ERP measure of syntactic processing. *Lang Cogn Process, 8*(4), 439-483.

Hagoort, P., & Brown, C. M. (2000). ERP effects of listening to speech compared to reading: The P600/SPS to syntactic violations in spoken sentences and rapid serial visual presentation. *Neuropsychologia, 38*(11), 1531-1549.

Hagoort, P., & Indefrey, P. (2014). The neurobiology of language beyond single words. *Annu Rev Neurosci, 37*, 347-362.

Halgren, E., Dhond, R. P., Christensen, N., Van Petten, C., Marinkovic, K., Lewine, J. D., & Dale, A. M. (2002). N400-like magnetoencephalography responses modulated by semantic context, word frequency, and lexical class in sentences. *NeuroImage, 17*(3), 1101-1116.

Halgren, E., Wang, C., Schomer, D. L., Knake, S., Marinkovic, K., Wu, J., & Ulbert, I. (2006). Processing stages underlying word recognition in the anteroventral temporal lobe. *NeuroImage*.

Harm, M. W., & Seidenberg, M. S. (2004). Computing the meanings of words in reading: Cooperative division of labor between visual and phonological processes. *Psychological Review, 111*(3), 662-720.

Harpur, G. F., & Prager, R. W. (1994). *Experiments with simple Hebbian-based learning rules in pattern classification tasks*. Citeseer.

Heilbron, M., Armeni, K., Schoffelen, J.-M., Hagoort, P., & de Lange, F. P. (2020). *A hierarchy of linguistic predictions during natural language comprehension*. bioRxiv.

Heilbron, M., Richter, D., Ekman, M., Hagoort, P., & de Lange, F. P. (2020). Word contexts enhance the neural representation of individual letters in early visual cortex. *Nature Communications, 11*(1).

Hickok, G., & Poeppel, D. (2007). The cortical organization of speech processing. *Nature Reviews Neuroscience, 8*(5), 393-402.

Hinton, G. E., McClelland, J. L., & Rumelhart, D. E. (1986). Distributed representations. In D. E. Rumelhart, J. L. McClelland, & PDP Research Group (Eds.), *Parallel Distributed*

Processing: Explorations in the Microstructure of Cognition, Vol. 1: Foundations (pp. 77-109). Cambridge, MA: MIT Press.

Hinton, G. E., & Shallice, T. (1991). Lesioning an attractor network: Investigations of acquired dyslexia. *Psychological Review, 98*(1), 74-95.

Hoeks, J. C. J., Stowe, L. A., & Doedens, G. (2004). Seeing words in context: The interaction of lexical and sentence level information during reading. *Brain Res Cogn Brain Res, 19*, 59-73.

Holcomb, P. J. (1988). Automatic and attentional processing: an event-related brain potential analysis of semantic priming. *Brain and Language, 35*(1), 66-85.

Holcomb, P. J., & Grainger, J. (2006). On the time course of visual word recognition: An event-related potential investigation using masked repetition priming. *Journal of Cognitive Neuroscience, 18*(10), 1631-1643.

Holcomb, P. J., Grainger, J., & O'Rourke, T. (2002). An electrophysiological study of the effects of orthographic neighborhood size on printed word perception. *Journal of Cognitive Neuroscience, 14*(6), 938-950.

Holcomb, P. J., Kounios, J., Anderson, J. E., & West, W. C. (1999). Dual-coding, context-availability, and concreteness effects in sentence comprehension: An electrophysiological investigation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25*(3), 721-742.

Holcomb, P. J., & Neville, H. J. (1990). Auditory and visual semantic priming in lexical decision: A comparison using event-related brain potentials. *Language and Cognitive Processes, 5*(4), 281-312.

Ihara, A., Hayakawa, T., Wei, Q., Munetsuna, S., & Fujimaki, N. (2007). Lexical access and selection of contextually appropriate meaning for ambiguous words. *Neuroimage, 38*(3), 576-588.

Iragui, V., Kutas, M., & Salmon, D. P. (1996). Event-related brain potentials during semantic categorization in normal aging and senile dementia of the Alzheimer's type. *Electroencephalography and Clinical Neurophysiology, 100*(5), 392-406.

Ito, A., Corley, M., Pickering, M., Martin, A. E., & Nieuwland, M. S. (2016). Predicting form and meaning: Evidence from brain potentials. *Journal of Memory and Language, 86*, 157-171.

Ito, A., Martin, A. E., & Nieuwland, M. S. (2016). How robust are prediction effects in language comprehension? Failure to replicate article-elicited N400 effects. *Language, Cognition and Neuroscience, 32*(8), 954-965.

Jackendoff, R. (2002). *Foundations of Language: Brain, Meaning, Grammar, Evolution*. New York: Oxford University Press.

Kim, A., & Osterhout, L. (2005). The independence of combinatory semantic processing: Evidence from event-related potentials. *J Mem Lang, 52*(2), 205-225.

Kiyonaga, K., Grainger, J., Midgley, K., & Holcomb, P. J. (2007). Masked cross-modal repetition priming: An event-related potential investigation. *Language and Cognitive Processes, 22*(3), 337-376.

Kleinschmidt, D. F., & Jaeger, F. T. (2015). Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel. *Psychol Rev, 122*(2), 148-203.

Knoeferle, P., & Crocker, M. W. (2006). The coordinated interplay of scene, utterance, and world knowledge: evidence from eye tracking. *Cognitive Science, 30*(3), 481-529.

Kok, P., & de Lange, F. P. (2015). Predictive Coding in Sensory Cortex. In *An Introduction to Model-Based Cognitive Neuroscience* (pp. 221-244). New York, NY: Springer.

Kolk, H. H. J., Chwilla, D. J., van Herten, M., & Oor, P. J. (2003). Structure and limited capacity in verbal working memory: A study with event-related potentials. *Brain Lang, 85*(1), 1-36.

Kounios, J., Green, D. L., Payne, L., Fleck, J. I., Grondin, R., & McRae, K. (2009). Semantic richness and the activation of concepts in semantic memory: evidence from event-related potentials. *Brain Research, 1282*, 95-102.

Kounios, J., & Holcomb, P. J. (1994). Concreteness effects in semantic processing: ERP evidence supporting dual-coding theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*(4), 804-823.

Kuperberg, G. R. (2007). Neural mechanisms of language comprehension: Challenges to syntax. *Brain Research, 1146*, 23-49.

Kuperberg, G. R. (2013). The proactive comprehender: What event-related potentials tell us about the dynamics of reading comprehension. In B. Miller, L. Cutting, & P. McCardle (Eds.), *Unraveling Reading Comprehension: Behavioral, Neurobiological, and Genetic Components* (pp. 176-192). Baltimore, MD: Paul Brookes Publishing.

Kuperberg, G. R. (2016). Separate streams or probabilistic inference? What the N400 can tell us about the comprehension of events. *Lang Cogn Neurosci, 31*(5), 602-616.

Kuperberg, G. R., Brothers, T., & Wlotko, E. (2020). A tale of two positivities and the N400: Distinct neural signatures are evoked by confirmed and violated predictions at different levels of representation. *Journal of Cognitive Neuroscience, 32*(1), 12-35.

Kuperberg, G. R., & Jaeger, T. F. (2016). What do we mean by prediction in language comprehension? *Lang Cogn Neurosci, 31*(1), 32-59.

Kuperberg, G. R., Paczynski, M., & Ditman, T. (2011). Establishing causal coherence across sentences: an ERP study. *Journal of Cognitive Neuroscience, 23*(5), 1230-1246.

Kuperberg, G. R., Sitnikova, T., Caplan, D., & Holcomb, P. J. (2003). Electrophysiological distinctions in processing conceptual relationships within simple sentences. *Brain Res Cogn Brain Res, 17*(1), 117-129.

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

Kutas, M., & Hillyard, S. A. (1980). Reading senseless sentences: brain potentials reflect semantic incongruity. *Science, 207*(4427), 203-205.

Kutas, M., & Hillyard, S. A. (1984). Brain potentials during reading reflect word expectancy and semantic association. *Nature, 307*(5947), 161-163.

Laszlo, S., & Armstrong, B. C. (2014). PSPs and ERPs: applying the dynamics of post-synaptic potentials to individual units in simulation of temporally extended Event-Related Potential reading data. *Brain Lang, 132*, 22-27.

Laszlo, S., & Federmeier, K. D. (2007). Better the DVL you know: Acronyms reveal the contribution of familiarity to single-word reading. *Psychological Science, 18*(2), 122-126.

Laszlo, S., & Federmeier, K. D. (2009). A beautiful day in the neighborhood: An event-related potential study of lexical relationships and prediction in context. *J Mem Lang, 61*(3), 326-338.

Laszlo, S., & Federmeier, K. D. (2011). The N400 as a snapshot of interactive processing: Evidence from regression analyses of orthographic neighbor and lexical associate effects. *Psychophysiology, 48*(2), 176-186.

Laszlo, S., & Federmeier, K. D. (2014). Never seem to find the time: Evaluating the physiological time course of visual word recognition with regression analysis of single-item event-related potentials. *Language, Cognition and Neuroscience, 29*(5), 642-661.

Laszlo, S., & Plaut, D. C. (2012). A neurally plausible parallel distributed processing model of event-related potential word reading data. *Brain and Language, 120*(3), 271-281.

Lau, E. F., Phillips, C., & Poeppel, D. (2008). A cortical network for semantics: (De)constructing the N400. *Nature Rev Neurosci, 9*(12), 920-933.

Lau, E. F., Weber, K., Gramfort, A., Hämäläinen, M. S., & Kuperberg, G. R. (2016). Spatiotemporal signatures of lexico-semantic prediction. *Cerebral Cortex, 26*(4), 1377-1387.

Lee, C.-l., & Federmeier, K. D. (2008). To watch, to see, and to differ: An event-related potential study of concreteness effects as a function of word class and lexical ambiguity. *Brain and Language, 104*(2), 145-158.

Levin, B. (1993). *English Verb Classes And Alternations: A Preliminary Investigation*. Chicago: University of Chicago Press.

Lillicrap, T. P., Santoro, A., Marris, L., Akerman, C. J., & Hinton, G. (2020). Backpropagation and the brain. *Nat Rev Neurosci, 21*(6), 335-346.

Lindborg, A., & Rabovsky, M. (2021). Meaning in brains and machines: Internal activation update in large-scale language model partially reflects the N400 brain potential. *Proceedings of the Annual Meeting of the Cognitive Science Society, 43*, 1049-1055.

MacDonald, M. C., Pearlmutter, N. J., & Seidenberg, M. S. (1994). The lexical nature of syntactic ambiguity resolution. *Psychological Review, 101*(4), 676-703.

Maess, B., Herrmann, C. S., Hahne, A., Nakamura, A., & Friederici, A. D. (2006). Localizing the distributed language network responsible for the N400 measured by MEG during auditory sentence processing. *Brain Res, 1096*(1), 163-172.

Marinkovic, K., Dhond, R. P., Dale, A. M., Glessner, M., Carr, V., & Halgren, E. (2003). Spatiotemporal dynamics of modality-specific and supramodal word processing. *Neuron, 38*(3), 487-497.

Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. San Francisico: W. H. Freeman.

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology, 18*(1), 1-86.

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: I. An account of basic findings. *Psychological Review, 88*(5), 375-407.

McClelland, J. L., St. John, M., & Taraban, R. (1989). Sentence comprehension: A parallel distributed processing approach. *Language and Cognitive Processes, 4*, 287-336.

McRae, K., & Matsuki, K. (2009). People use their knowledge of common events to understand language, and do so as quickly as possible. *Lang Linguist Compass, 3*(6), 1417-1429.

Meade, G., Grainger, J., & Holcomb, P. J. (2019). Task modulates ERP effects of orthographic neighborhood for pseudowords but not words. *Neuropsychologia, 129*, 385-396.

Meade, G., Midgley, K. J., Dijkstra, T., & Holcomb, P. J. (2018). Cross-language neighborhood effects in learners indicative of an integrated lexicon. *J Cogn Neurosci, 30*(1), 70-85.

Michaelov, J. A., Coulson, S., & Bergen, B. K. (2021). So Cloze yet so Far: N400 amplitude is better predicted by distributional information than human predictability judgements. *arXiv preprint arXiv:2109.01226*.

Midgley, K. J., Holcomb, P. J., & Grainger, J. (2011). Effects of cognate status on word comprehension in second language learners: an ERP investigation. *J Cogn Neurosci, 23*(7), 1634-1647.

Midgley, K. J., Holcomb, P. J., VanHeuven, W. J., & Grainger, J. (2008). An electrophysiological investigation of cross-language effects of orthographic neighborhood. *Brain Research, 1246*, 123-135.

Millidge, B., Tschantz, A., & Buckley, C. J. (2020). *Predictive coding approximates backprop along arbitrary computation graphs*. Cornell University. arXiv.

Misra, M., & Holcomb, P. J. (2003). Event-related potential indices of masked repetition priming. *Psychophysiology, 40*(1), 115-130.

Moreno, E. M., Federmeier, K. D., & Kutas, M. (2002). Switching languages, switching palabras (words): an electrophysiological study of code switching. *Brain and Language, 80*(2), 188-207.

Morgan, E., Brothers, T., & Kuperberg, G. R. (under review). Parallel predictions during language comprehension: Contextual facilitation without competition.

Moss, H. E., Ostrin, R. K., Tyler, L. K., & Marslen-Wilson, W. D. (1995). Accessing different types of lexical semantic information: Evidence from priming. *Journal of Experimental Psychology: Learning, Memory and Cognition, 21*, 863-883.

Mumford, D. (1992). On the computational architecture of the neocortex. II. The role of cortico-cortical loops. *Biological Cybernetics, 66*(3), 241-251.

Murray, S. O., Schrater, P., & Kersten, D. (2004). Perceptual grouping and the interactions between visual cortical areas. *Neural Netw, 17*(5-6), 695-705.

Myslin, M., & Levy, R. (2016). Comprehension priming as rational expectation for repetition: Evidence from syntactic processing. *Cognition, 147*, 29-56.

Nieuwland, M. S., & Kuperberg, G. R. (2008). When the truth is not too hard to handle: an event-related potential study on the pragmatics of negation. *Psychological Science, 19*(12), 1213-1218.

Nieuwland, M. S., Politzer-Ahles, S., Heyselaar, E., Segaert, K., Darley, E., Kazanina, N., . . . Huettig, F. (2018). Large-scale replication study reveals a limit on probabilistic prediction in language comprehension. *eLife, 7*, e33468.

Nobre, A. C., & McCarthy, G. (1995). Language-related field potentials in the anterior-medial temporal lobe: II. Effects of word type and semantic priming. *Journal of Neuroscience, 15*(2), 1090-1098.

Osterhout, L., & Holcomb, P. J. (1992). Event-related brain potentials elicited by syntactic anomaly. *J Mem Lang, 31*(6), 785-806.

Osterhout, L., & Holcomb, P. J. (1993). Event-related potentials and syntactic anomaly: Evidence of anomaly detection during the perception of continuous speech. In S. M. Garnsey (Ed.), *Language and Cognitive Processes. Special Issue: Event-related brain potentials in the study of language* (Vol. 8 (4), pp. 413-437). Hove: Lawrence Erlbaum Associates.

Osterhout, L., & Mobley, L. A. (1995). Event-related brain potentials elicited by failure to agree. *Journal of Memory and Language, 34*, 739-773.

Otten, M., Nieuwland, M. S., & Van Berkum, J. J. (2007). Great expectations: Specific lexical anticipation influences the processing of spoken language. *BMC Neuroscience, 8*(1), 89-97.

Paczynski, M., & Kuperberg, G. R. (2011). Electrophysiological evidence for use of the animacy hierarchy, but not thematic role assignment, during verb argument processing. *Lang Cogn Process, 26*(9), 1402-1456.

Paczynski, M., & Kuperberg, G. R. (2012). Multiple influences of semantic memory on sentence processing: Distinct effects of semantic relatedness on violations of real-world event/state knowledge and animacy selection restrictions. *J Mem Lang, 67*(4), 426-448.

Payne, B. R., Lee, C. L., & Federmeier, K. D. (2015). Revisiting the incremental effects of context on word processing: Evidence from single-word event-related brain potentials. *Psychophysiology, 52*(11), 1456-1469.

Peeters, D., Dijkstra, T., & Grainger, J. (2013). The representation and processing of identical cognates by late bilinguals: RT and ERP effects. *Journal of Memory and Language, 68*(4), 315-332.

Piai, V., Roelofs, A., Rommers, J., & Maris, E. (2015). Beta oscillations reflect memory and motor aspects of spoken word production. *Human Brain Mapping, 36*(7), 2767-2780.

Pickering, M. J., & Garrod, S. (2013). An integrated theory of language production and comprehension. *Behavioral and Brain Sciences, 36*(04), 329-347.

Plaut, D. C., & Shallice, T. (1993). Deep dyslexia: A case study of connectionist neuropsychology. *Cognitive Neuropsychology, 10*(5), 377-500.

Plaut, D. C., & Shallice, T. (1993). Perseverative and semantic influences on visual object naming errors in optic aphasia: a connectionist account. *J Cogn Neurosci, 5*(1), 89-117.

Price, C. J., & Devlin, J. T. (2011). The interactive account of ventral occipitotemporal contributions to reading. *Trends in Cognitive Sciences, 15*(6), 246-253.

Pylkkänen, L., & McElree, B. (2007). An MEG study of silent meaning. *Journal of Cognitive Neuroscience, 19*(11), 1905-1921.

Rabovsky, M. (2020). Change in a probabilistic representation of meaning can account for N400 effects on articles: A neural network model. *Neuropsychologia, 143*, 107466.

Rabovsky, M., Hansen, S. S., & McClelland, J. L. (2018). Modelling the N400 brain potential as change in a probabilistic representation of meaning. *Nature Human Behaviour, 2*(9), 693-705.

Rabovsky, M., & McRae, K. (2014). Simulating the N400 ERP component as semantic network error: Insights from a feature-based connectionist attractor model of word meaning. *Cognition, 132*(1), 68-89.

Rabovsky, M., Sommer, W., & Abdel Rahman, R. (2012). The time course of semantic richness effects in visual word recognition. *Frontiers in Human Neuroscience, 6*, 11.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI blog, 1*(8), 9.

Rao, R. P. N., & Ballard, D. H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience, 2*(1), 79-87.

Rastle, K. (2016). Visual Word Recognition. In G. Hickok & S. L. Small (Eds.), *Neurobiology of Language* (pp. 255-264). San Diego, CA: Academic Press.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review, 85*(2), 59-108.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin, 124*(3), 372-422.

Rogers, T. T., & McClelland, J. L. (2008). Précis of semantic cognition: A parallel distributed processing approach. *Behavioral and Brain Sciences, 31*(06), 689-749.

Rommers, J., & Federmeier, K. D. (2018). Predictability's aftermath: Downstream consequences of word predictability as revealed by repetition effects. *Cortex, 101*, 16-30.

Rugg, M. D. (1985). The effects of semantic priming and word repetition on event-related potentials. *Psychophysiology, 22*, 642-647.

Rugg, M. D. (1990). Event-related potentials dissociate repetition effects of high and low frequency words. *Memory and Cognition, 18*, 367-379.

Rumelhart, D. E. (1979). Some problems with the notion of literal meanings. In A. Ortony (Ed.), *Metaphor and Thought* (pp. 71-82). Cambridge, UK: Cambridge University Press.

Russeler, J., Becker, P., Johannes, S., & Munte, T. F. (2007). Semantic, syntactic, and phonological processing of written words in adult developmental dyslexic readers: an event-related brain potential study. *BMC Neuroscience, 8*, 52.

Sanford, A. J., Leuthold, H., Bohan, J., & Sanford, A. J. S. (2011). Anomalies at the borderline of awareness: an ERP study. *Journal of Cognitive Neuroscience, 23*, 514-523.

Seidenberg, M. S., & McClelland, J. L. (1989). A distributed developmental model of word recognition and naming. *Psychological Review, 96*(4), 523-568.

Shetreet, E., Alexander, E. J., Romoli, J., Chierchia, G., & Kuperberg, G. R. (2019). What we know about knowing: Presuppositions generated by factive verbs influence downstream neural processing. *Cognition, 184*, 96-106.

Shipp, S. (2016). Neural elements for predictive coding. *Front Psychol, 7*, 1792.

Sohoglu, E., & Davis, M. H. (2020). Rapid computations of spectrotemporal prediction error support perception of degraded speech. *Elife, 9*.

Song, Y., Lukasiewicz, T., Xu, Z., & Bogacz, R. (2020). *Can the brain do backpropagation?---exact implementation of backpropagation in predictive coding networks.* Paper presented at the Advances in Neural Information Processing Systems 33 (NeurIPS 2020).

Spratling, M. W. (2008). Predictive coding as a model of biased competition in visual attention. *Vision Research, 48*(12), 1391-1408.

Spratling, M. W. (2012). Unsupervised learning of generative and discriminative weights encoding elementary image components in a predictive coding model of cortical function. *Neural Computation, 24*(1), 60-103.

Spratling, M. W. (2013). Image segmentation using a sparse coding model of cortical area V1. *IEEE Transactions on Image Processing, 22*(4), 1631-1643.

Spratling, M. W. (2014). A single functional model of drivers and modulators in cortex. *Journal of Computational Neuroscience, 36*(1), 97-118.

Spratling, M. W. (2016). Predictive coding as a model of cognition. *Cognitive Processing, 17*(3), 279-305.

Spratling, M. W. (2017). A review of predictive coding algorithms. *Brain and Cognition, 112*, 92-97.

Spratling, M. W., De Meyer, K., & Kompass, R. (2009). Unsupervised learning of overlapping image components using divisive input modulation. *Comput Intell Neurosci, 2009*, 381457.

St. John, M. F., & McClelland, J. L. (1990). Learning and applying contextual constraints in sentence comprehension. *Artificial Intelligence, 46*(1-2), 217-257.

Staub, A., Grant, M., Astheimer, L., & Cohen, A. (2015). The influence of cloze probability and item constraint on cloze task response time. *Journal of Memory and Language, 82*, 1-17.

Sterzer, P., Adams, R. A., Fletcher, P., Frith, C., Lawrie, S. M., Muckli, L., . . . Corlett, P. R. (2018). The predictive coding account of psychosis. *Biological Psychiatry, 84*(9), 634-643.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement Learning: An Introduction, Second Edition*. Cambridge, MA: MIT Press.

Swaab, T., Brown, C., & Hagoort, P. (1997). Spoken sentence comprehension in aphasia: Event-related potential evidence for a lexical integration deficit. *Journal of Cognitive Neuroscience, 9*, 39–66.

Szewczyk, J. M., Mech, E. N., & Federmeier, K. D. (2021). The power of "good": Can adjectives rapidly decrease as well as increase the availability of the upcoming noun? *J Exp Psychol Learn Mem Cogn*.

Szewczyk, J. M., & Wodniecka, Z. (2020). The mechanisms of prediction updating that impact the processing of upcoming word: An event-related potential study on sentence comprehension. *J Exp Psychol Learn Mem Cogn, 46*(9), 1714-1734.

Tanenhaus, M. K., & Trueswell, J. C. (1995). Sentence comprehension. In J. L. Miller & P. D. Eimas (Eds.), *Speech, Language, and Communication* (2 ed., Vol. 11, pp. 217-262). San Diego, CA: Academic Press.

Urbach, T. P., DeLong, K. A., Chan, W.-H., & Kutas, M. (2020). An exploratory data analysis of word form prediction during word-by-word reading. *Proceedings of the National Academy of Sciences, 117*(34), 20483-20494.

Urbach, T. P., DeLong, K. A., & Kutas, M. (2015). Quantifiers are incrementally interpreted in context, more than less. *Journal of Memory and Language, 83*, 79-96.

Usher, M., & McClelland, J. L. (2001). The time course of perceptual choice: The leaky, competing accumulator model. *Psychological Review, 108*(3), 550-592.

van Boxtel, J. J. A., & Lu, H. (2013). A predictive coding perspective on autism spectrum disorders. *Frontiers in Psychology, 4*.

Van de Cruys, S., Evers, K., Van der Hallen, R., Van Eylen, L., Boets, B., de-Wit, L., & Wagemans, J. (2014). Precise minds in uncertain worlds: Predictive coding in autism. *Psychological Review, 121*(4), 649-675.

Van Petten, C. (1993). A comparison of lexical and sentence-level context effects in event-related potentials. Special Issue: Event-related brain potentials in the study of language. *Language and Cognitive Processes, 8*, 485-531.

Van Petten, C., Coulson, S., Rubin, S., Plante, E., & Parks, M. (1999). Time course of word identification and semantic integration in spoken language. *Journal of Experimental Psychology: Learning, Memory and Cognition, 25*(2), 394-417.

Van Petten, C., & Kutas, M. (1990). Interactions between sentence context and word frequency in event-related brain potentials. *Memory and Cognition, 18*, 380-393.

Van Petten, C., & Kutas, M. (1991). Influences of semantic and syntactic context on open- and closed-class words. *Memory and Cognition, 19*(1), 95-112.

Van Petten, C., & Luka, B. J. (2012). Prediction during language comprehension: benefits, costs, and ERP components. *Int J Psychophysiol, 83*(2), 176-190.

Walsh, K. S., McGovern, D. P., Clark, A., & O'Connell, R. G. (2020). Evaluating the neurophysiological evidence for predictive processing as a model of perception. *Ann N Y Acad Sci, 1464*(1), 242-268.

Wang, L., Hagoort, P., & Jensen, O. (2018). Language prediction is reflected by coupling between frontal gamma and posterior alpha oscillations. *Journal of Cognitive Neuroscience, 30*(3), 432-447.

Wang, L., Jensen, O., & Kuperberg, G. R. (2018). *Representational Similarity Analysis reveals unique patterns associated with the fulfillment and violation of lexico-semantic prediction within the N400 time window: An MEG study*. Paper presented at the 25th Annual Meeting of the Cognitive Neuroscience Society, Boston, MA.

Wang, L., Kuperberg, G., & Jensen, O. (2018). Specific lexico-semantic predictions are associated with unique spatial and temporal patterns of neural activity. *eLife, 7*, e39061.

Wang, L., Schoot, L., Brothers, T., Alexander, E., Kim, M., Warnke, L., . . . Kuperberg, G. R. (Under review). Predictive coding across the left fronto-temporal hierarchy during language comprehension. *bioRxiv preprint*.

Whittington, J. C. R., & Bogacz, R. (2019). Theories of error back-propagation in the brain. *Trends Cogn Sci, 23*(3), 235-250.

Winsler, K., Midgley, K. J., Grainger, J., & Holcomb, P. J. (2018). An electrophysiological megastudy of spoken word recognition. *Lang Cogn Neurosci, 33*(8), 1063-1082.

Wlotko, E. W., & Federmeier, K. D. (2012). So that's what you meant! Event-related potentials reveal multiple aspects of context use during construction of message-level meaning. *NeuroImage, 62*(1), 356-366.

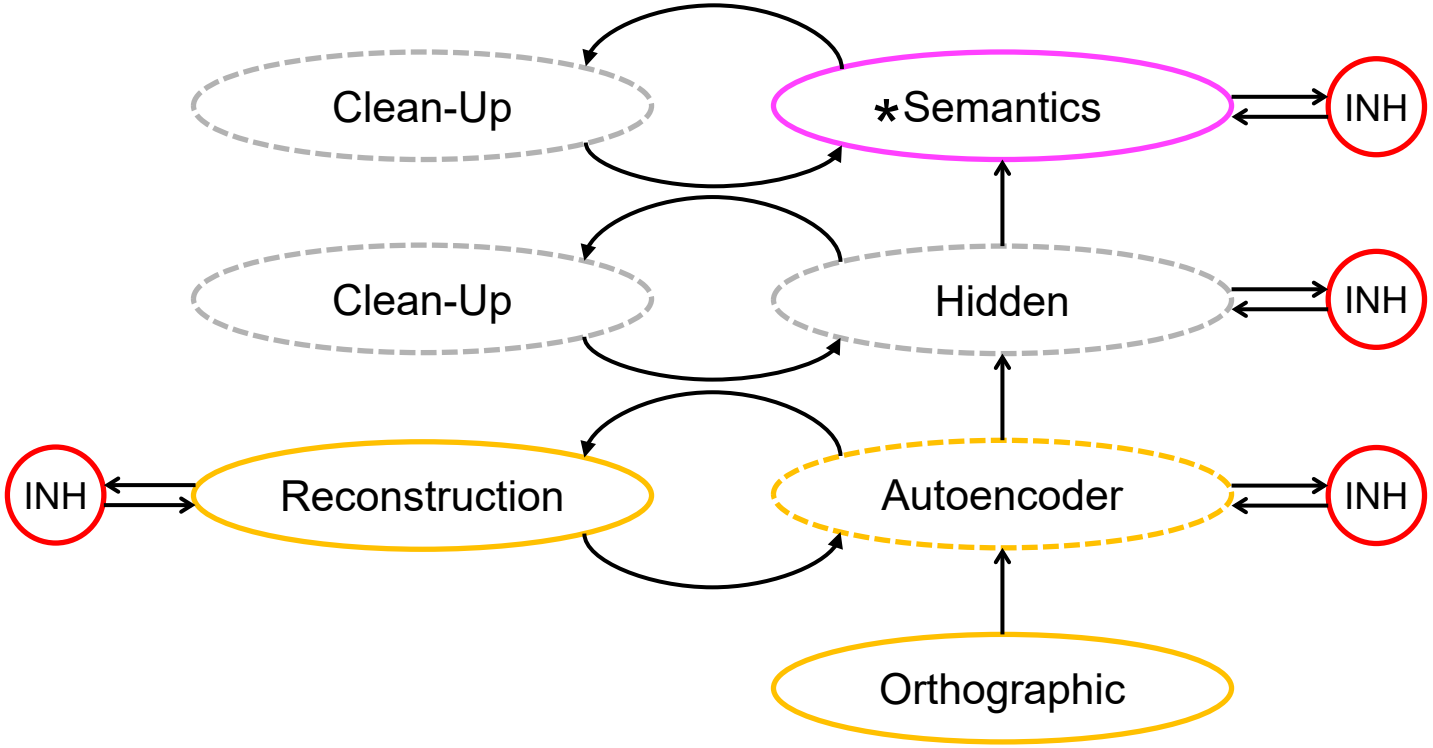
Xiang, M., & Kuperberg, G. (2015). Reversing expectations during discourse comprehension. *Lang Cogn Neurosci, 30*(6), 648-672.

Young, M. P., & Rugg, M. D. (1992). Word frequency and multiple repetition as determinants of the modulation of event-related potentials in a semantic classification task. *Psychophysiology, 29*(6), 664-676.

A.

Semantic Activation Model

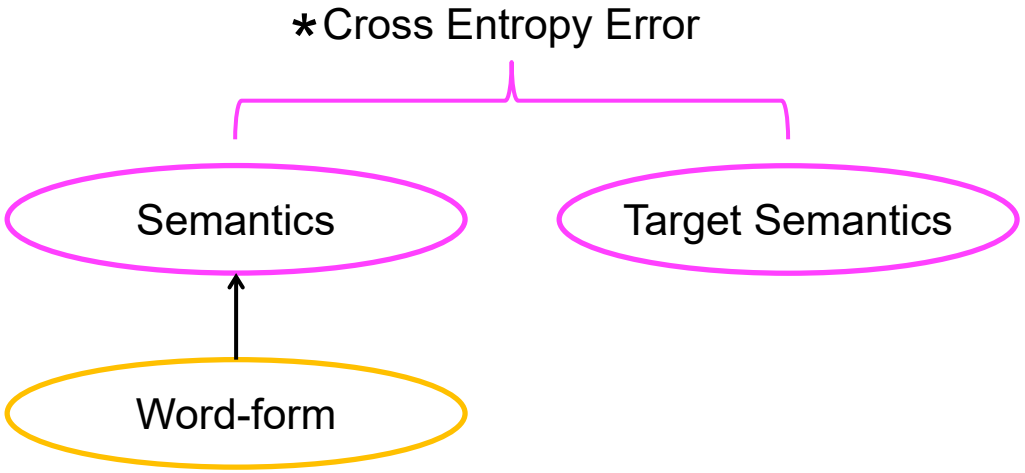
Cheyette & Plaut, 2017



B.

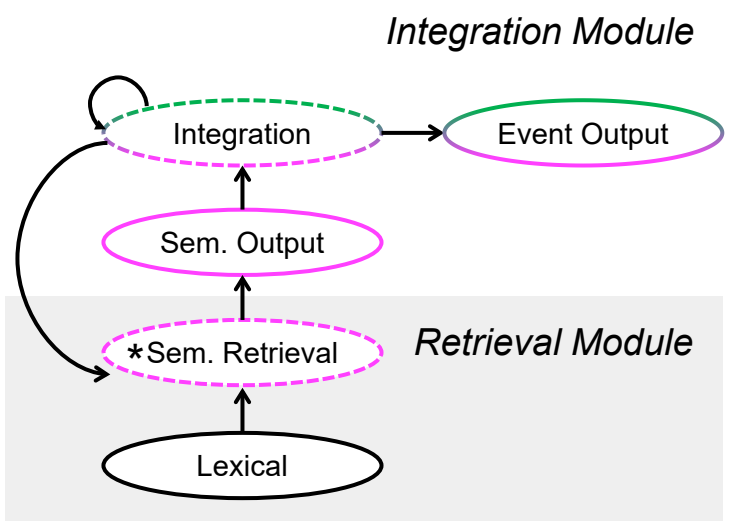
Semantic Attractor Model

Rabovsky & McRae, 2014



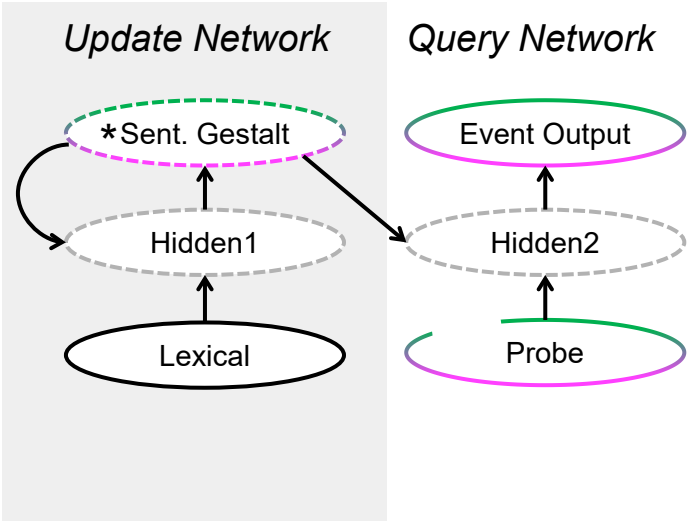
A. Retrieval-Integration Model

Brouwer, et al, 2017



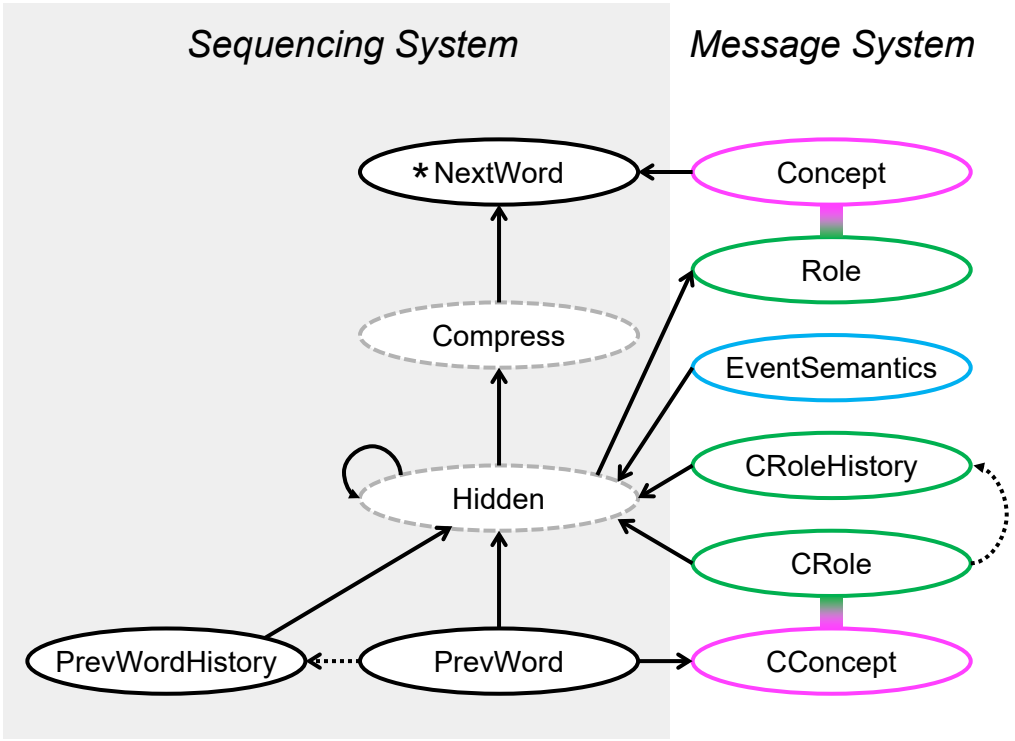
B. Sentence Gestalt Model

Rabovsky, Hansen & McClelland, 2018

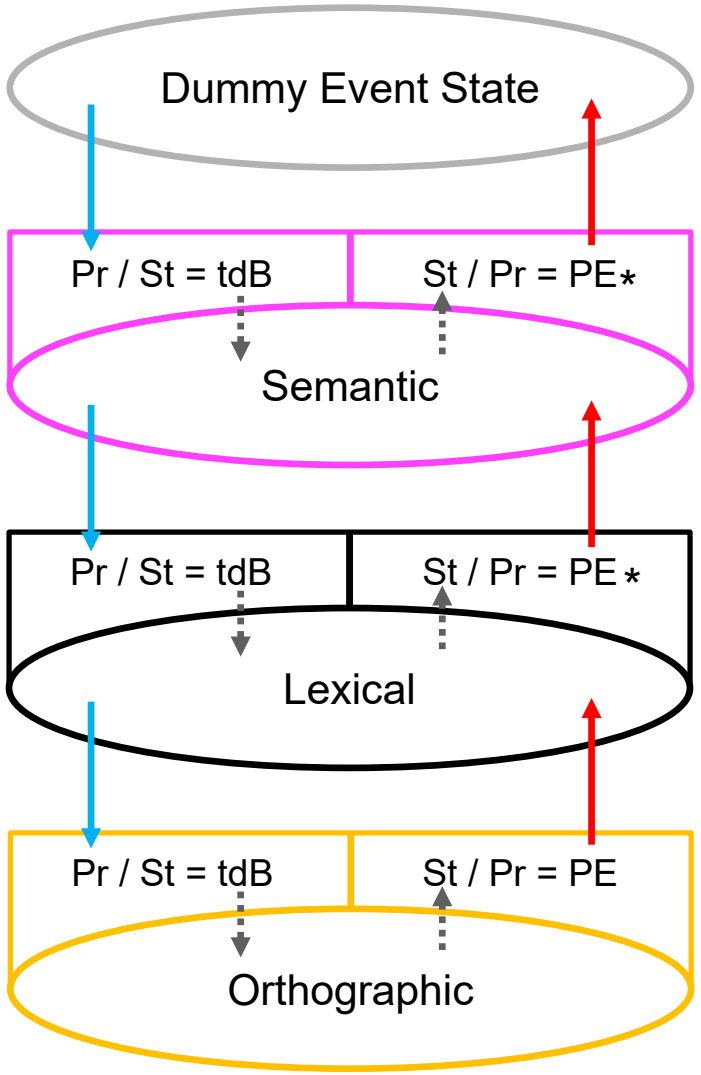


C. Error Propagation Model

Fitz & Chang, 2019



A. Predictive Coding Architecture



: *Error units*

: *State units*

Pr: *Prediction*
St: *State*
PE: *Prediction Error*
tdB: *Top-Down Bias*

B. N400 Simulations