Divide & Concur: A Predictive Coding Account of the N400 ERP Component

A thesis submitted by

Samer A. Nour Eddine

in partial fulfillment of the requirements for the degree of

Master of Science

in

Psychology

Tufts University

August 2021

Advisor: Gina Kuperberg

Abstract

The N400 event-related brain potential has provided core insights into the nature of on-line language comprehension, and its amplitude is modulated by a wide variety of lexical and contextual factors. Across many paradigms, the N400 amplitude appears to be a function of the mismatch between information that is anticipated based on learned regularities and information that is actually encountered. Various theories have been verbally formulated to explain what processes the N400 reflects, but much is left unspecified by these theories, making it difficult to test them experimentally. One way forward is to specify their assumptions explicitly in the form of a computational model. The current work proposes a hierarchical predictive coding model of perceptual inference that explicitly specifies how predictions and prediction errors are computed during on-line word comprehension, accounting for the influence of a range of lexical- and contextual-level factors on the N400.

Acknowledgments

One of the major difficulties in grad school is finding a direction of research that is appealing to you, relatively novel and promising, and within your advisor's area of expertise. I owe a debt of gratitude to my advisor Gina Kuperberg for going out of her way to help me in this process, while patiently allowing me to explore on my own. Her dedication and energy as a mentor is inspiring. I want to thank Trevor Brothers for his unwavering support in helping me navigate out of dead ends (writing and modeling alike), and for making sure I never lost sight of the big picture. I am grateful to Lin Wang for challenging my ideas (and the lab's) from interesting directions, for asking the right questions, and for identifying exciting new directions for the current work; and Arim Choi-Perrachione for patiently designing multiple drafts of this paper's figures. I want to thank Michael Spratling for being very responsive and supportive, and for teaching me the benefits of writing efficient code; and Ariel Goldberg for carefully reading and commenting on the dissertation, and for offering practical advice on how to deal with a large hyperparameter space. Finally, I want to thank my family for their unconditional love and support – Nadeen, Hazem, Hanan, Nagham, mom and dad; and my friends here in the US for being my family away from home – Majed, Jamal, Rami, and Rita.

Table of Contents

List of Tables

List of Figures

Divide & Concur: A Predictive Coding Account of the N400 ERP Component

A key discovery in the history of psycholinguistics was the presence of a signature — the N400 event-related component — that indexes online language processing in the brain (Kutas & Hillyard, 1980, 1984; DeLong, Urbach & Kutas, 2005). While there has been considerable interest in developing a theoretical framework for understanding the N400, this has proved a formidable challenge. Many theories and computational models have provided compelling explanations for many of its functional properties. However, a unifying, biologically plausible account remains elusive. In this study, we show that the N400 can be understood as the magnitude of lexico-semantic prediction error computed as a key step of a computational algorithm that has been proposed to carry out perceptual inference in the brain — predictive coding (Friston, 2005; Clark, 2013, see also Spratling, 2016; Rao & Ballard, 1999; Mumford, 1992; Lee & Mumford, 2003). Using an implemented predictive coding model of lexico-semantic processing, we show that the magnitude of lexico-semantic prediction error tracks both the temporal dynamics and the functional sensitivity of the N400 to both bottom-up lexical and top-down contextual information.

The N400 event-related potential (ERP) component is a negative-going waveform that is detected at the scalp surface using both electroencephalography (EEG) and magnetoencephalography (MEG) between 300 to 500 ms following the onset of any meaningful stimulus, such as a word or a picture (see Kutas & Federmeier, 2011 for a review). During sentence comprehension, the N400 is highly sensitive to the predictability of a word in relation to its prior context, whether this context is a single word (semantic and repetition priming paradigms, e.g. Bentin, McCarthy & Wood, 1985; Rugg, 1985), or a more extended sentence or discourse context (e.g., Kutas & Hillyard, 1984; DeLong, Urbach & Kutas, 2005; Van Berkum,

Hagoort & Brown, 1999). The N400 is also elicited by words presented in isolation, and its amplitude is sensitive to multiple lexico-semantic factors, including lexical frequency (e.g. *wart* (low frequency) > *cold* (high frequency); Rugg, 1990; Van Petten & Kutas, 1990), concreteness (e.g. *lime* > *know*; Kounios & Holcomb, 1994; Holcomb, Kounios, Anderson & West, 1999), and orthographic neighborhood size (e.g. *core* > *kiwi*; Holcomb, Grainger & O'Rourke, 2002; Laszlo & Federmeier, 2011).

Despite extensive work on the N400, there is still no general consensus on its functional significance. In the early 2000s, two competing theories dominated the debate: a lexical access and an integration account. Briefly, the lexical access view interpreted the N400 as reflecting the difficulty of accessing or "recognizing" a unique lexical item (e.g. Lau, Phillips, & Poeppel, 2008), while the integration account interpreted it as a "post-lexical" process that integrated the accessed item into its preceding context (Brown & Hagoort, 1993; Hagoort, 2009). However, as several researchers pointed out, this type of dichotomy between "access" and "integration" has difficulty in explaining both lexical and contextual effects on the N400 (e.g. Kutas & Federmeier, 2011; Baggio & Hagoort, 2011; Kuperberg 2016). More generally, it rests on the somewhat questionable assumption that lexical access and semantic integration are distinct, separable cognitive processes that occur in a fixed sequence (see Laszlo & Federmeier, 2011; Kuperberg, Brothers & Wlotko, 2020 for discussion).

These shortcomings led to the more general proposal that the N400 reflects the impact of stimulus-driven, feed-forward activation on the current state of semantic memory (Kutas & Federmeier, 2011). In this framework, semantic memory is conceptualized as a dynamic multimodal system that is interactively influenced by both the high-level incremental interpretation of the prior context, as well as the orthographic and phonological form of each

incoming word. This theory therefore intuitively explains a number of top-down and bottom-up influences on the amplitude of the N400. For example, as new bottom-up input becomes available, the co-activation of the semantic features associated with partially overlapping orthographic neighbors would result in an enhanced N400 response (see Laszlo & Federmeier, 2011), and the amplitude of the N400 would be reduced to the degree that the input matches semantic features had been pre-activated by the prior context (e.g. Federmeier & Kutas, 1999). On the other hand, the theory's flexibility leaves a number of cognitive processes unspecified. How do particular stimuli activate the correct set of semantic features in long-term memory? Why does lexical processing result in the partial activation of orthographic and semantic neighbors, and how does the brain ultimately suppress these to settle on a "correct" interpretation of the bottom-up input? What determines the characteristic rise and fall of the N400 response? Most importantly, how are these processes implemented in a biologically plausible fashion in the brain?

One way of addressing these questions is through the development of explicit computational models. Several researchers have risen to this challenge, and a number of connectionist models of the N400 have been described (Laszlo & Plaut, 2012; Laszlo & Armstrong, 2014; Cheyette & Plaut, 2017; Rabovsky & McRae, 2014; Brouwer, Crocker, Venhuizen & Hoeks, 2017; Rabovsky, Hansen & McClelland, 2018; Fitz & Chang, 2019). Broadly, these models of the N400 fall into two classes: word-level and sentence-level.

In the word-level models, the goal is to map a single word-form input (e.g., a letter-string), clamped at the input layer, to a pattern of activation that represents the word's meaning at the top layer (Laszlo & Plaut, 2012; Laszlo & Armstrong, 2014; Cheyette & Plaut, 2017; Rabovsky & McRae, 2014). In one set of studies, the authors describe a biologically motivated

architecture in which the dynamic activation of semantic features simulated the time course of the N400 (its rise and fall over time) (Laszlo & Plaut, 2012; Laszlo & Armstrong, 2014; Cheyette & Plaut, 2017). The authors also showed that the average semantic activity was sensitive to various lexico-semantic variables that are known to modulate the N400, including orthographic neighborhood size (Laszlo & Plaut, 2012), concreteness (Cheyette & Plaut, 2017), frequency (Cheyette & Plaut, 2017), as well as the relationship between the target word and preceding identity (Laszlo & Armstrong, 2014) or semantically related prime (Cheyette & Plaut, 2017); see Table 1. In another model, Rabovsky & McRae (2014) conceptualized the N400 amplitude as the mismatch between the model's observed semantic activity and an ideal "correct" semantic target, which they linked to "prediction error". They showed that, within their model, this operationalization of the N400 was able to account for a similar range of findings (see Table 1).

In sentence-level models of the N400, the goal is to incrementally map from a sequence of word inputs to a higher-level representation of meaning (Brouwer, Crocker, Venhuizen & Hoeks, 2017; Rabovsky, Hansen & McClelland, 2018; Fitz & Chang, 2019). Building this higher-level representation requires the model to retain a representation of the full sequence of prior inputs while processing new inputs, which is achieved by including a recurrent element in the neural network (cf. Elman, 1990; Jordan, 1986) that is trained jointly with the rest of the model to perform a task that relies on retaining this sequential representation. After the model is trained, the N400 is modeled either as the amount of change that a new word imposes on a particular latent representation in the network (Brouwer, Crocker, Venhuizen & Hoeks, 2017; Rabovsky, Hansen & McClelland, 2018), or as the deviation of the next-word prediction from the word that is subsequently presented (Fitz & Chang, 2019). Rather than focusing on the

influence of lower-level lexical factors on the N400, work in this direction has primarily focused on simulating the effects of prior context on the N400 (see Table 1).

The architectures and assumptions of these different computational models of the N400 are quite different from one another. However, it is worth emphasizing that, for nearly all of them, the N400 is conceptualized as a difference — the degree to which a state changes from before to after an incoming word is encountered (cf. McClelland, 1994). This difference is variously referred to as "state transition" (Brouwer et al., 2017), "semantic update" (Rabovsky et al., 2018), and "prediction error" (Rabovsky & McRae, 2014; Fitz & Chang, 2019). What remains unclear, however, is why the brain should actually compute a value that corresponds to these differences, and thus produce an N400 effect.

One possible answer to this question is that such differences in state are not explicitly computed, but are instead purely epiphenomenal — a byproduct of other computations (e.g., Brouwer et al., 2017). Another possibility, explicitly modeled by Rabovsky et al., (2018) and Fitz & Chang (2019), is that this difference serves as a learning signal that is computed post-inference.

Importantly, in none of these existing computational models is the surprise or prediction error signal that corresponds to the N400 evoked by each word computed as an inherent component of comprehension — the process of inferring meaning from bottom-up input. In the present study, we propose a different functional account of the N400 — that instead of reflecting a surprise signal that is a byproduct of inference or a post-inference learning signal, the N400 reflects the magnitude of a prediction error that is computed within an inference algorithm. As such, we argue that prediction error encodes lexico-semantic information that is actually passed from lower to higher levels of cortical representation during online language comprehension.

Predictive coding refers to a computational algorithm, implemented by a biologically plausible network architecture, that has been proposed to approximate probabilistic inference in the cortex. This algorithm and network architecture was first described in a seminal study by Rao & Ballard, 1999, who showed that it was able to simulate extra-classical receptive field effects in the visual system (see also Mumford, 1992; Lee & Mumford, 2003). It has since been used to simulate a wide range of basic perceptual neural phenomena, including end-stopping (Rao & Ballard, 1999), contour integration (Spratling, 2013; Spratling, 2014), binocular rivalry (Denison et al., 2011; Hohwy et al., 2008), and the mismatch negativity ERP component (Garrido et al., 2009; Wacongne et al., 2012). Predictive coding can also simulate a variety of cognitive phenomena, including top-down lexical effects on letter perception (Spratling, 2016). More generally, hierarchical predictive coding and its extensions have been proposed as an overarching computational account of how the brain carries out perceptual inference (Friston, 2005; Clark, 2013).

Like the classic connectionist networks that have transformed our understanding of language processing over the last few decades (McClelland & Rumelhart, 1981; Rumelhart & McClelland, 1982; Harm & Seidenberg, 2004), the predictive coding architecture is highly interactive, with both feedforward and feedback connections between levels of representation that allow for both the top-down and bottom-up flow of information across the hierarchy (see Fig. 1A). Where it is distinguished from these classic connectionist models is that it posits that, at each level of cortical representation, a set of "error units" explicitly computes "prediction error" — the difference or residual information between top-down predictions from the level above, and bottom-up input from the level below. These error units are distinct from the "state units" that actually represent the state of activity at any given level. Error and state units are

connected by a specific set of feedforward and feedback connections both within and across cortical levels of representation.

According to this algorithm, state units at a higher level of cortical representation continuously generate top-down predictions or "reconstructions" that are actively propagated to the error units of the level below. The prediction error computed by the error units is then passed back up and used to update the higher-level state unit representations, thereby allowing them to generate more accurate top-down reconstructions on the next iteration of the algorithm. Over multiple iterations, the magnitude of prediction error gradually decreases as the higher-level state units converge upon the internal representation that is most likely to explain the input ("explaining away" alternative, competing explanations encoded in state units, see Spratling, 2008). Analogous computations occur both at the cortical levels above and below until the prediction error is minimized across the entire generative hierarchy (see Figure 1B for more detailed schematic).

The claim that higher levels of representation continually predict activity at lower levels of representation — a central premise of predictive coding — intuitively maps on to more general predictive frameworks of language comprehension that emphasize the role of probabilistic pre-activation in facilitating neural processing of bottom-up linguistic inputs that match such predictions, manifest by a smaller amplitude N400 (Federmeier, 2007; DeLong, Urbach & Kutas, 2005; Kuperberg & Jaeger, 2016). Predictive coding attributes this attenuation of the N400 to the smaller prediction error produced by lexical and semantic error units in response to predicted (vs unpredicted) bottom-up inputs.

Predictive coding can also intuitively account for the influence of lexical variables on the N400 when these are associated with increased activity elicited by a bottom-up input (e.g.,

concreteness: more semantic features; orthographic neighborhood size: broader activation), which should enhance the magnitude of lexico-semantic prediction error produced by that input. Finally, as pointed out by Friston (2005), the incremental application of the predictive coding algorithm may also explain the characteristic rise and fall of evoked neural responses (such as the N400). This is because, as new unpredicted bottom-up information becomes available to error units, the magnitude of the prediction error they produce will increase, increasing N400 amplitude. However, as the higher-level state units converge on the underlying cause of the input, they will produce top-down reconstructions that switch off the lower-level prediction error, reducing N400 amplitude.

Despite these intuitions, and despite the fact that some researchers have appealed to the predictive coding framework in discussions of the N400 (e.g. Rabovsky & McRae, 2014; Bornkessel-Schlesewsky & Schlesewsky, 2019; Kuperberg, Brothers, & Wlotko, 2020), there have been no attempts to simulate the N400 (or any other language ERP component) using an implemented predictive coding model. In the present work, we explore that direction by implementing a lexico-semantic processing model based on the same predictive coding principles as those used to simulate low-level neural phenomena in the visual system (Rao & Ballard, 1999; Spratling, 2012; Spratling, 2013; Spratling, 2014). We used this model to simulate the effects of lexico-semantic variables (frequency, orthographic neighborhood and concreteness), single word repetition and semantic priming, contextual predictability, and interactions between these variables, operationalizing the N400 as the magnitude of lexico-semantic prediction error produced by error units on each iteration of the algorithm. This enabled us to determine whether predictive coding was able to account for the dynamics of the N400, as well as the effects of these bottom-up and top-down variables on its amplitude.

The Model

Model Architecture

The model computes activity at three hierarchically organized levels of linguistic representation — orthographic, lexical, and semantic (see Fig. 1A & 1B). As in all predictive coding architectures, each level of representation has two distinct types of connectionist units — *state units* and *error units*. While state unit activity encodes the internal representation at a given level of representation, error units encode the mismatch between observations and top-down predictions.

Our model incorporates two types of error units (Spratling, 2016). First, *bottom-up error units* compute "bottom-up error", otherwise referred to as "prediction error". This error reflects the residual information encoded within state units that is not present in the top-down reconstructions from the level above; this error is passed up the network to update higher-level state units. A separate set of *top-down error units* compute "top-down errors" or "bias": the residual information contained in top-down reconstructions that was not present in state units at the current level. These error units provide top-down biases during perceptual inference, while also allowing anticipatory information to be passed to lower levels of the network.

Bottom-up error units at a given level receive connections from both the higher-level and same-level state units, and they send feedforward connections to the higher level state units. Top-down error units also receive connections from both the higher-level and same-level state units, but send outgoing connections to the state units at the *same* level (i.e., feedback connections). Between levels, we provided a set of hand-coded weight matrices to map between levels of representation (see Supplementary Materials for details). In addition, in order to simulate the

effects of word familiarity, the strength of feedback weights at each layer were scaled as a function of each word's lexical frequency (Brysbaert et al., 2009; additional details below).

Finally, a fourth "contextual" layer was added to the top of the model. This has no error units and simply serves as a dummy layer to inject top-down activity into the model, enabling us to simulate the effects of contextual probability.

Representations and mappings between levels of representation

The *orthographic level* contains 104 units, encoding 26 letter identities (A-Z) at 4 possible spatial positions (cf. McClelland and Rumelhart, 1981).

At the *lexical level*, each unit (1579 in total) represented a four-letter word in the model's lexicon. Of these, 512 represented "critical words" that served as input stimuli in our simulations. An additional 1067 word units were also added to enrich the model's lexicon and to produce more realistic levels of competition among orthographically similar words.

At the *semantic level*, each unit (12,929 in total) represented a "toy" semantic feature (e.g. <plant>, <sour>; cf. Cheyette & Plaut, 2017; Rabovsky & McRae, 2014). Of the 512 critical words, half were *concrete*, with 18 semantic features, and half were abstract with 9 semantic features. These mappings were assigned such that each of the 512 lexical units shared between 0 and 8 semantic features with at least one other lexical unit. The remaining 1067 "filler" words were each assigned 9 unique semantic features. Table 2 provides an example of how shared features were assigned for four representative words in the lexicon.

The Predictive Coding Algorithm

The predictive coding algorithm implemented here is a modified version of the Predictive Coding/Biased Competition-Divisive Input Modulation algorithm(PC/BC-DIM) algorithm (Spratling, 2008; Spratling, 2016). Previously, this algorithm has successfully simulated neural phenomena in the early visual system (contour integration: Spratling, 2013, 2014; learning Gabor-like receptive fields: Spratling, 2012b) as well as higher-level cognitive processes, including top-down contextual effects on letter perception (Spratling, 2016). This model shares many processing principles with an influential predictive coding algorithm developed by Rao & Ballard, 1999, but, instead of computing bottom-up and top-down error through element-wise subtraction, it is computed through element-wise division (see Spratling, 2008; see Algorithm 1 and Supplementary Materials for details).

At each level of representation, on each iteration of the algorithm, (1) state units are updated, (2) two types of error are computed: bottom-up error (information in the input that is not in a reconstruction, also referred to as "prediction error") and top-down bias (information in a reconstruction that is not in the input, i.e. unmatched top-down prediction, also referred to as "pre-activation"), and (3) top-down reconstructions are computed. Each of these steps is described in detail in Algorithm 1 and summarized below:

(1) State vectors are updated based on three inputs: (a) the previously updated state value, (b) the newly computed bottom-up error from the level below ("prediction error"), and (c) the previously computed top-down error ("preactivation") from the same level. Specifically, the sum of the two error vectors serves as an update factor that multiplies the old state value elementwise to produce a new state value.

(2) Bottom-up error ("prediction error) is computed by dividing each element of the newly updated state by the corresponding element in the top-down reconstruction that was

computed on the previous iteration of the algorithm. Note that this ratio is largest when the newly updated state contains information absent in the reconstruction. Similarly, top-down error ("pre-activation") is computed by dividing each element in the reconstruction (computed on the previous iteration of the algorithm) by the corresponding element in the newly updated state. This value is largest when the reconstruction contains information that is absent from the newly updated state.

(3) Top-down reconstructions are computed based on the newly updated state.

The same three steps[1] then take place at the next level of the hierarchy. Iterating over these steps effectively carries out a gradient-based optimization procedure that minimizes the prediction error across the entire network (see Supplementary Materials for details). When error is minimized, the model has settled on a mutually compatible set of internal representations, and the model has effectively inferred a coherent latent cause of the bottom-up input.

During a typical iteration, either the orthographic or contextual state vector is clamped to an input value (orthographic clamping propagates bottom-up stimulus information to the rest of the model; contextual clamping propagates word probability expectations to the rest of the model), and all unclamped units are updated as described above. Note that the order of computations is fixed regardless of whether the orthographic or contextual state vector is clamped.

[1] Note that the orthographic level has no level below it – and so the state units do not receive bottom-up error, and do not generate reconstructions of the level below (however, it receives top-down pre-activation from the lexical level). Similarly, the contextual level has no level above it, and so it does not have error units associated with it (however, it receives bottom-up error from the semantic level). The algorithm can accommodate these edge cases with no issues; the terms corresponding to the missing variables are simply dropped from the general equations. Therefore, a simpler version than the general description above will apply to either edge case.

Simulations

In order to test whether predictive coding mechanisms can account for the functional sensitivity profile and time course of the N400, we simulated a wide range of benchmark phenomena in the N400 literature, including: (1) the effects of three lexico-semantic variables (frequency, orthographic neighborhood, and concreteness) to words presented in isolation; (2) the effects of lexical repetition (*lime* – *lime*) and semantic priming (*sour* – *lime*) for related versus unrelated word pairs; (3) the effects of top-down contextual predictability and anticipatory semantic overlap during observed during sentence processing; (4) interactions between priming/top-down contextual effects and the lexical variables described above.

In all simulations, the model was presented with 512 orthographic input vectors, each of which corresponded to one of the critical 512 words in the model's lexicon (described above under Architecture). For each simulation, our hypothesized index of the N400 was the summed magnitude of the bottom-up lexical and semantic errors produced by the predictive coding algorithm, consistent with our assumption that the N400 represents the total lexicosemantic error.

In order to visualize the rise-and-fall of prediction errors over multiple iterations of algorithm, we constructed prediction error time courses for each condition of interest. On each iteration of the algorithm, we summed the bottom-up semantic and lexical error across all bottom-up error units, and then averaged this summed "lexico-semantic prediction error" across items.

Similar to the N400, this error measure produced a rise-and-fall "waveform" after word onset (e.g., see Figure 3). Therefore, in analogy to ERP studies and roughly in line with other models of the N400 (e.g., Cheyette & Plaut, 2017), we calculated an N400 response for each

item by taking the average error in a 10-iteration widow surrounding the peak (2 -12 iterations after word onset):

$$\text{N400} = \frac{1}{10} \sum_{t=\text{onset}+2}^{\text{onset}+12} \left(\sum_i |\mathbf{e}^t_{\text{lex}_i}| + \sum_i |\mathbf{e}^t_{\text{sem}_i}| \right)$$

In order to test the effects of lexical variables, which varied between items, we carried out simple regression analyses. To examine the effects of priming and contextual predictability, which varied within items, as well as any interactions between these factors and lexical variables, we performed linear mixed effects regression using lme4 package version 1.1-21(Bates et al., 2015) in R (R Core Team, 2016). In all these models, we used random intercepts at the item level, and random slopes for all within-item predictors (the maximal identifiable random effects structure: Barr et al., 2013), and assessed statistical significance using a type-III sums of squares estimation, with p-values estimated using the Satterthwaite approximation (Satterthwaite, 1946) using lmerTest version 3.1-0 (Kuznetsova et al., 2015). Below we report the results that address our main theoretical questions, outlined above.

To examine the effects of frequency, orthographic neighborhood size and concreteness, we initialized the model so that units at each level were uniformly activated (see Algorithm 1 for details) and presented each of our 512 critical words, extracting the lexicosemantic prediction error at each iteration. As described in the Architecture section, we selected 512 critical words in a 2 x 2 x 2 design, ensuring a) that these stimuli included a wide range of lexical frequencies (Brysbaert et al., 2009), orthographic neighborhood sizes (OLD20, Orthographic Levenshtein Distance 20, Yarkoni, Balota & Yap, 2012), and concreteness values (half concrete, half abstract), and (b**)** that these three variables were fully orthogonal (all $r < 0.07$). To visualize the effects of frequency and orthographic neighborhood size on the prediction error time courses, we averaged prediction error values across the 256 words above and the 256 words below a cut-off value (e.g. high vs. low frequency; see Table 3). In our statistical analyses, frequency, orthographic neighborhood and concreteness were entered as continuous linear predictors in a single regression model.

Frequency

Background: It is well established that the amplitude of the N400 is smaller to high frequency than low frequency words (e.g., Rugg, 1990; Van Petten & Kutas, 1990; Laszlo & Federmeier, 2014; Hauk et al., 2006).

Methods: As noted under Architecture, we used between-level feedback connectionist weights to simulate the effect of frequency. Log-transformed frequency values (Brysbaert et al., 2009) were encoded in all feedback connection weights so that units associated with high frequency words received stronger feedback from higher-level state units (see Figure 2 for illustration of frequent *ball* vs infrequent *gall*). In particular, we started with a frequency-naive matrix and added a frequency-bias score to each non-zero feedback connection; this score was

obtained by scaling the softmax of the log frequencies of all 1579 words in the model lexicon (see Supplementary Materials for details).

There are multiple approaches to set up the model so that the processing of high frequency words is facilitated. Because knowledge of lexical frequency is acquired over a long period of time and reflects a stable prior, we chose to encode this bias into the weights of the network rather than in the state activations which are more volatile[2] (Spratling, 2016b; Vilares and Kording, 2011). This is also in line with the idea that lexical frequency alters the weighting of perceptual evidence (Norris, 2006), and with how others have modeled lexical frequency (Cheyette & Plaut, 2017; Rabovsky, Hansen & McClelland, 2018).

Results: Consistent with empirical studies of the N400, frequency predicted the magnitude of lexicosemantic prediction error (b = -3.98, t = -4.88, $p < .001$). High frequency words produced a smaller lexicosemantic prediction error than low frequency words (see Fig. 3).

Orthographic Neighborhood Size

Background: In previous studies, words with more orthographic neighbors (e.g. *ball*: *bull*, *call*, *bail*) produce larger N400 responses than words with few orthographic neighbors (e.g., *kiwi*; Holcomb, Grainger, & O'Rourke, 2002; Laszlo & Federmeier, 2011). Indeed, for individual words, the magnitude of the N400 responses shows a strong linear relationship with measures of orthographic neighborhood size (Laszlo & Federmeier, 2011).

[2] Encoding frequency as differences in the resting activation, prior to word onset, had no influence on error magnitudes in this multiplicative model. Modeling frequency by increasing the feedforward between-level weights would not bias the model's prior beliefs, but would instead serve as an update bias (such that a frequent input would cause the model to update more aggressively towards the frequent item); it also produced an opposite effect (i.e., higher frequency words elicited a larger error).

Methods: As described under Architecture, the orthographic neighborhood size for each of our 512 input items was implicitly specified in the hand-specified connectionist weights between their corresponding lexical unit and the orthographic units (e.g., *ball* and *gall* are orthographic neighbors, see Fig 2).

Results: Consistent with empirical studies, orthographic neighborhood size predicted the magnitude of lexicosemantic prediction error (OLD20: b = -35.06, t = -43.07, p <.001). High orthographic neighborhood words produced a larger lexicosemantic prediction error than low orthographic neighborhood words (Fig. 4) and this effect was continuous and graded (Fig 5).

Concreteness

Background: The magnitude of the N400 produced by concrete words is larger than that produced by abstract words (Kounios & Holcomb, 1994; Holcomb, Kounios, Anderson & West, 1999).

Methods: Following previous computational models of the N400 (Cheyette & Plaut, 2017; Rabovsky & McRae, 2014), we operationalized concreteness as the number of semantic features associated with each critical word (18 vs. 9).

Results: Concrete words produced a larger lexicosemantic prediction error overall than abstract words, consistent with empirical findings (Concreteness: b = 11.11, t = 13.63, p < .001; Fig. 6).

Word-pair Priming Effects

In a typical priming paradigm (e.g., Bentin, McCarthy &Wood, 1985), pairs of "prime" and "target" words are presented sequentially, and the target N400 response is assessed depending on whether these words are related or unrelated along a particular dimension. Here we simulated both repetition (Rugg, 1985; Misra & Holcomb, 2003) and semantic priming (Bentin, McCarthy &Wood, 1985; Rugg, 1985). On each trial the model was presented with a prime word for 20 iterations, followed by two blank iterations – in which orthographic state units were clamped to zero – and then a target word for an additional 20 iterations. To visualize the priming effects, we constructed time courses that averaged the lexicosemantic prediction error produced by all 512 inputs when preceded by either related or unrelated primes. In lmer analyses, Relatedness served as a categorical within-item predictor.

Repetition Priming

Background: In repetition priming paradigms, the amplitude of the N400 evoked by repeated targets is smaller than that evoked by unrelated targets (e.g. Rugg, 1985; Misra & Holcomb, 2003).

Methods: In these simulations, in the repeated condition, we presented prime words that were the same as each of our target 512 critical words, and in the unrelated condition, we presented another randomly selected word.

Results: Consistent with the empirical findings on the N400, we found that repeated targets elicited significantly smaller semantic bottom-up prediction error than unrelated targets (main effect of Relatedness: $b = -247.29$, $t = -190.78$, $p < .001$), see Fig. 7.

Semantic Priming

Background: The amplitude of the N400 evoked by semantically related targets is smaller than that evoked by unrelated targets (e.g. Bentin, McCarthy &Wood, 1985; Rugg, 1985; Holcomb, 1988; Holcomb & Neville, 1990).

Methods: Following previous computational models of the N400 (Cheyette & Plaut, 2017; Rabovsky & McRae, 2014; Rabovsky, Hansen & McClelland, 2018), we operationalized semantic relatedness as of the number of semantic features shared between the prime and the target. For each of the 512 targets in our stimulus set, there was one word that shared eight semantic features with it. This word was selected to serve as that target's semantically related prime. Each of these targets also had 256 words that shared zero semantic features with it, and we randomly sampled one prime from this set to serve as that target's unrelated prime.

Results: Consistent with empirical findings, targets preceded by a semantically related prime elicited significantly smaller semantic bottom-up prediction error than those preceded by an unrelated prime (main effect of Relatedness: b = -218.77, t = -129.57, $p < .001$, see Fig. 8.

Effects of Cloze & Constraint

In a naturalistic setting, words are rarely presented in isolation, and instead are often surrounded by informative sentence or discourse contexts. It is well established that amplitude of the N400 is highly sensitive to lexico-semantic predictability – i.e. the match/mismatch between a word and the lexico-semantic features that have been pre-activated by the preceding context (e.g. Kutas & Hillyard, 1984; DeLong, Urbach & Kutas, 2005; Federmeier & Kutas, 1999). To simulate the effects of top-down pre-activation, we clamped "contextual state units" at the highest level of the model with activations equal to the contextual predictability of each word (see below). After 20 iterations, we unclamped these contextual state units and clamped one of

the 512 bottom-up inputs for an addition 20 iterations. In these simulations, we examined the effects of both top-down predictability and contextual constraint on the magnitude of lexico-semantic errors.

Effect of cloze probability

Background: There is strong empirical evidence that the amplitude of the N400 evoked by a given word is inversely proportional to its lexical probability in context (e.g. Kutas & Hillyard, 1984; DeLong, Urbach & Kutas, 2005; Wlotko & Federmeier, 2012). Lexical probability is typically operationalized as the proportion of participants who produce the presented word following each sentence context during an offline task known as the cloze procedure (Taylor, 1953).

Methods: To simulate these effects of lexical predictability, we presented 512 critical words, at four different levels of probability: 99%, 50%, 25% and uniform (1/[total words] = 1/1579 = 0.06%). On each run, we clamped the contextual unit that corresponded to the input, assigning it a percentage of total activation at the contextual state layer, which always summed to 2. The remaining activation was always assigned uniformly across all other contextual state units.

Results: Consistent with prior N400 data, we observed increasingly smaller magnitude of bottom-up error unit activation as the bottom-up input became more predictable (Cloze: b = -86.58, t = -59.57, p < .001), see Figure 9.

Effect of lexical constraint, irrespective of cloze probability

Background: The amplitude of the N400 evoked by an incoming word during sentence comprehension is primarily sensitive to its predictability, rather than to whether or not it violates

a prior prediction (Kutas & Hillyard, 1984; Federmeier, Wlotko, De Ochoa-Dewald & Kutas, 2007; Kuperberg, Brothers, & Wlotko, 2020). For example, equally large N400 responses are produced by unpredictable words in non-constraining contexts (e.g. "*Helen reached up to dust the dresser*.") and unpredictable words that violate a strong lexical prediction (e.g. "*The groom took the bride's hand and placed the ring on her dresser*.", where the word "*finger*" was the expected completion.). Although, in these two sentences, the word *dresser* is matched in cloze probability (0%), the preceding contexts differ in their *contextual constraint* — operationalized as the percentage participants in a cloze task who produce the same word "best completion".

Methods: To simulate the effect of lexical constraint for unpredictable words, we presented each critical word at two different levels of contextual constraint (high constraint = 99%, vs. low-constraint = 0.06%, i.e. uniform pre-activation). We also compared these two unexpected conditions to simulations in which the input was correctly predicted (*high constraint expected*, 99% cloze).

Results: Consistent with N400 data, lexico-semantic errors were significantly reduced for predictable words (HC_unexp vs. HC_exp: b = -252.85, t = -173.85, p < .001), but there were no significant differences as a function of lexical constraint (HC_unexp vs. LC_unexp: b = -0.20, t = -0.14, p = .89; see Fig 10)

Semantic prediction overlap effect

Background: In addition to its sensitivity to lexical predictability, the amplitude of the N400 is also sensitive to the semantic relationship between a previously predicted word and the bottom-up input (Kutas & Hillyard, 1984; Federmeier & Kutas, 1999). For example, Federmeier & Kutas (1999) asked participants to read contexts like "*They wanted to make the hotel look*

more like a tropical resort. So along the driveway, they planted rows of…", which constrained for an expected word like *palms*. The amplitude of the N400 evoked by low cloze (<1%) words like *pines*, which shared several semantic features with the expected word, was smaller than that evoked by low cloze words that shared fewer features with expected word (e.g. *tulips*). This effect interacted with contextual constraint, with a larger attenuation of the N400 in high constraint compared to medium constraint contexts. This was taken as evidence that the N400 attenuation for unexpected inputs (*pines*) was driven by the pre-activation of semantic features that overlapped with the anticipated word (*palms*), with the assumption that semantic prediction was strongest following highly constraining contexts.

Methods: To simulate this semantic prediction overlap effect for each critical word, we pre-activated a contextual state unit for a word that shared either 8 semantic features or zero semantic features with the upcoming bottom-up input. Unlike our previous simulations, the presented and pre-activated words were always different, both lexically and orthographically. Contextual state units were assigned a probability of either 99% or 50% (high vs. medium constraint) with the remaining activation distributed equally across the other contextual state units. This gave rise to a 2 (Overlap: overlapping, non-overlapping) x 2 (Constraint: high, medium) design. We ran four simulations for each of our 512 critical words.

Results: The time course of the total lexicosemantic bottom-up error for each of the four conditions is presented in Figure 11. The bottom-up input elicited a significantly smaller semantic prediction error when the context pre-activated a word that was semantically related (Relatedness: $b = -24.04$, $t = -41.70$, $p < .001$), and this difference was larger in high constraint compared to medium constraint contexts (Relatedness x Cloze: $b = -5.50$, $t = -16.50$, $p < .001$; see Fig 11), consistent with the empirical results (Federmeier & Kutas, 1999).

Interactions

Background: It is well established that lexical processing is not encapsulated, and that top-down context can influence the impact of low-level lexical factors on the N400 response. In particular, the effects of both frequency and concreteness are smaller when a word has been repeated (repetition x frequency: Rugg, 1990; repetition x concreteness: Kounios & Holcomb, 1994) or presented in a predictable sentence context (cloze x frequency: Dambacher, Kliegl, Hofmann, & Jacobs, 2006; cloze x concreteness: Holcomb, Kounios, Anderson, & West, 1999). On the other hand, no studies have conclusively demonstrated interactions between orthographic neighborhood size and these contextual factors. In an ERP study where words were presented in isolation, Laszlo & Federmeier (2011) noted that orthographic neighborhood size effect was slightly reduced on second presentation, but no statistical tests were reported. In two ERP studies where words appear in a sentence context (Payne & Federmeier, 2018; Payne, Lee & Federmeier, 2015), the N400 effect of orthographic neighborhood size was not affected by the critical word's position in the sentence or by whether it appeared in a congruous sentence (versus syntactic prose and jumbled sentences). However, these latter studies did not explicitly manipulate cloze probability, and thus could not assess its interaction with orthographic neighborhood size.

Interactions with Repetition

Method: We re-analyzed the error values that were generated in the original repetition simulations, examining how the lexicosemantic bottom-up error varied as a function of orthographic neighborhood density (continuous), lexical frequency (continuous), and concreteness (Abstract, Concrete) across the two repetition conditions (Repeated, Unrepeated).

We fit a linear mixed effects model with three interaction terms: Repetition x OLD20, Repetition x Frequency and Repetition x Concreteness.

Results: Consistent with N400 data, we observed overall smaller error activations for words with sparser orthographic neighborhood density (OLD20: b = -33.31, t = -35.31, p <.001) and higher lexical frequency (Frequency: b = -4.17, t = -4.42, p < .001); and we observed greater error activation overall by more concrete words (Concreteness: b = 9.35, t = 9.90, p < .001). Each of these factors interacted with repetition (see Fig. 12), such that their effects were always weaker for repeated words (Repetition x OLD20: b = 32.54, t = 25.05, p < .001; Repetition x Frequency: b = 3.85, t = 2.96, p < .001; Repetition x Concreteness: b = -6.06, t = -4.66, p < .001).

Interactions with Cloze

Method: Similar to our repetition analyses, we also examined how the lexicosemantic bottom-up error varied as a function of orthographic neighborhood density (continuous), lexical frequency (continuous), and concreteness (Abstract, Concrete) at different extremes of cloze probability (99%, 0.06%). We fit a linear mixed effects model with three interaction terms: Cloze x OLD20, Cloze x Frequency and Cloze x Concreteness.

Results: Consistent with N400 data, we observed overall smaller error activations for words with sparser orthographic neighborhood density (OLD20: b = -18.33, t = -44.65, p <.001) and higher lexical frequency (Frequency: b = -2.60, t = -6.32, p < .001); and we observed greater error activation overall by more concrete words (Concreteness: b = 5.69, t = 13.86, p < .001). Each of these factors interacted with cloze (see Fig. 13), such that their effects were significantly

weaker in the high cloze condition (Cloze x OLD20: $b = 16.66$, $t = 42.18$, $p < .001$; Cloze x Frequency: $b = 1.36$, $t = 3.44$, $p < .001$; Cloze x Concreteness: $b = -5.31$, $t = -13.44$, $p < .001$)

Discussion

In this study, we successfully implemented a predictive coding model of lexico-semantic processing. We were able to show that the time course of the error activity in this model could capture (i) the morphology of the N400, (ii) the effects of lexical (orthographic neighborhood size, lexical frequency and concreteness) and contextual (cloze, constraint, semantic prediction overlap, semantic priming, repetition priming) factors on its amplitude, in addition to (iii) interactions between them (repetition/cloze x frequency/concreteness). There have been several previous computational models that successfully simulated subsets of these N400 effects (Brouwer, Crocker, Venhuizen & Hoeks, 2017; Cheyette & Plaut, 2017; Fitz & Chang, 2019; Laszlo & Armstrong, 2014; Laszlo & Plaut, 2012; Rabovsky, Hansen & McClelland, 2018; Rabovsky & McRae, 2014), and much qualitative discussion has linked the N400 to effects of probabilistic top-down prediction on lexico-semantic processing during language comprehension (Fitz & Chang, 2019; Rabovsky, Hansen & McClelland, 2018; Rabovsky & McRae, 2014). In this work, we introduced a computational model that actually uses a predictive coding architecture and algorithm to simulate the N400 during language comprehension.

In predictive coding theory, prediction error is the primary mechanism by which new (unpredicted) information is fed forward in the model. That is, prediction error plays a fundamental role in perceptual processing; there is no sense in which the system would be more effective if error were avoided. The system is in a situation where states at each level must infer the activity of lower-level states sitting behind a wall of error units. They have no direct access to their target; they can only make predictions and wait until the next iteration to “hear back” from

the error units. And so the system being described is one where bottom-up information fundamentally serves a prescriptive function (encoded as error), not a descriptive one.

For example, if the semantic state has relatively high activity at the units corresponding to the <organism>, <roar> and <croak> features, it might predict that the lexical state that is ambiguous between *lion* and *frog* (activating both units halfway). This prediction takes the form of a reconstruction: a vector that will be directly compared (at the next iteration) to the true lexical state vector at the error units. Suppose the lexical state unit corresponding to *lion* was in fact much more active than *frog*. The semantic state units do not have direct access to this information but instead rely on the bottom-up lexical error units to send back any unpredicted information: in this case, the information is that the semantic units corresponding to *lion* (i.e., <roar>) are less active than they should be. An analogous interaction happens between the semantic error units and the contextual level: the semantic prediction error constitutes the semantic information that was unpredicted by the contextual units.

Crucially, in predictive coding, the iterative interaction of prediction errors with state units at each level of representation serves to optimize a cost function (the generalized KL divergence between reconstructions and states; see Supplementary Materials for details) whose minimization allows higher level representation to converge on a state that best explains the bottom-up input.

This means that, as new unpredicted lexical and semantic bottom-up information initially becomes available to error units at these levels, the magnitude of lexical and semantic prediction error will at first rise because states at the higher level have not generated accurate reconstructions of these inputs. However, on each iteration of algorithm, the error will update the state, which will in turn generate a more accurate reconstruction. This means that, over the

course of multiple iterations, the magnitude of prediction error will fall again as it is "switched off" by higher level state units converging on the correct representation. This trajectory of prediction error over the course of predictive coding algorithm mirrors the characteristic rise-and-fall morphology of negative-going event-related potentials — a point noted by Friston (2005) in relation to the mismatch negativity. In all our simulations, the lexical and semantic prediction errors trace out the morphology of the N400 event-related component.

These simulations begin to address the high-level question expressed most clearly by Kutas & Federmeier (2011): "... the question becomes not where is the N400 generator localized, or whether there are multiple N400s, but rather what are the functions of the dynamic neural system of which scalp N400s are reflections?". To the extent that the total lexicosemantic error in our model shows both the time course and functional sensitivity profile that characterizes the N400, our simulations suggest an answer: a neural system implementing a predictive coding scheme for perceptual inference will require units whose collective activity turns out to closely resemble the N400. We summarize these simulations below.

Relationship to Empirical Findings

We successfully simulated effects of multiple lexical factors on the N400 amplitude elicited by single words. First, consistent with empirical findings (Holcomb, Grainger & O'Rourke, 2002; Laszlo & Federmeier, 2009, 2011; Midgley, Holcomb, van Heuven, & Grainger, 2008; Müller, Duñabeitia, & Carreiras, 2010), the magnitude of the elicited lexical and semantic PE was systematically larger when the input word had a denser orthographic neighborhood within the model (i.e., when the input corresponded to lexical representations that were connected to the same lower-level letter units as many other words in our model). This is because the bottom-up (error) information elicited by dense-neighbor inputs is far less specific

than that elicited by sparse-neighbor inputs, resulting in noisier updates of states, which will result in (i) a less accurate reconstruction being generated and (ii) a less specific error information being passed up the hierarchy at the next level. Therefore more error is required to converge on the correct state for dense- compared to sparse-neighborhood items.

Second, consistent with empirical findings that show larger N400 to concrete vs abstract words (Kounios & Holcomb, 1994; Holcomb, Kounios, Anderson & West, 1999), we observed that inputs with lexical representations that were connected to more vs fewer semantic features evoked larger lexicosemantic prediction errors. This follows from the simple fact that error is elicited elementwise – when more semantic units are involved (with the same amount of unpredicted information per unit), the total prediction error will be greater. Note that this effect was seen primarily at the semantic level, rather than the lexical level — although the lexical error elicited by a concrete (vs abstract) input will affect a larger number of semantic units (i.e., via more feedforward connections), the lexical error itself was not larger for concrete compared to abstract inputs. In principle, this finding is consistent with the observation that the concreteness effect on the N400 has a more anterior scalp distribution than other lexical effects.

Finally, we replicated the finding that words with a higher lexical frequency elicit smaller N400 amplitudes: we observed that inputs that corresponded to units that have stronger between-layer feedback connections elicited smaller lexicosemantic errors overall. This is because a unit with stronger feedback weights (i.e., associated with a high frequency word, see Fig 2) will have a larger influence on the reconstruction in response to a given amount of error compared to a unit whose feedback weights are weaker. Therefore, in situations where the error is ambiguous and non-specific, the model is biased in favor of the higher frequency word; thus frequency-biased feedback weights implement a form of stable, prior belief in our model (Norris, 2006). On the

other hand, in situations where the error is unambiguous, the processing of frequent (vs infrequent) words is facilitated in the sense that the model requires less error to converge on the correct representations. Note that although frequency information is permanently encoded in the model, its online influence on the dynamics of the model is eliminated when the state units associated with stronger feedback weights are inhibited (because the stronger weights will be multiplied by zero), consistent with the notion that accumulating evidence can override this prior.

We were also able to simulate the well-established effects of repetition (Laszlo & Federmeier, 2007, 2011; Misra & Holcomb, 2003; Nagy & Rugg, 1989; Rugg, 1985; Sim & Kiefer, 2005) and semantic priming (Bentin, McCarthy, & Wood, 1985; Federmeier & Kutas, 1999; Kutas, 1993; Kutas & Iragui, 1998; Rugg, 1985) on the N400. We presented prime-target pairs for 20 iterations each (interrupted with 2 blank iterations), and found that targets preceded by related primes elicited smaller lexical and semantic prediction errors than targets preceded by unrelated primes. This is because in both cases, the presentation of the related prime allowed the model to partially (in the case of semantic priming) or fully (repetition priming) converge on lexical and semantic state units that corresponded to the target's representation. Therefore, the amount of information that was not already predicted by the state units – i.e., error – upon the target's onset was significantly smaller for inputs that followed related versus unrelated primes.

Our model also captures several classic effects of prior context on the N400. In more general predictive processing frameworks, effects of context are usually attributed to top-down predictive pre-activation; the attenuation of the N400 to incoming lexical and semantic inputs that match these prior top-down predictions is taken to reflect the relative ease of accessing

predicted compared to unpredicted information (DeLong, Urbach & Kutas, 2005; Kuperberg, Brothers & Wlotko, 2020).

The predictive coding framework maps very intuitively on to these assumptions. To simulate pre-activation, we provided the highest "contextual" layer with top-down input for 20 iterations. During this pre-activation phase, the activated contextual state produced semantic reconstructions that were used to computed a top-down semantic bias (within the top-down semantic error units), which pre-activated the semantic state units, leading them to generate their own lexical reconstructions. In this way, the contextual information was propagated down the hierarchy in a top-down fashion. This meant that, when new expected lexical and semantic input became available (via bottom-up input), it produced smaller lexical and semantic prediction errors than unexpected input.

Importantly, we were able to show that the effects of context on lexical and semantic prediction error were graded in nature. In other words, like the N400 itself, the more probable the incoming word, the smaller the prediction error. When we summed this prediction error across the lexical and semantic levels of representation, this effect appeared roughly linear, in keeping with empirical findings on the N400, which likely reflects both lexical and semantic prediction error at scalp surface (Brothers, Noriega, Kuperberg, 2020).

Again consistent with empirical findings on the N400, the magnitude of lexical and semantic prediction error produced by unpredicted inputs was not different when they violated a strong prediction of a different word compared to when they were not predicted at all (simulated by pre-activating the model's entire lexicon uniformly). This finding is important because it speaks to a potential misunderstanding about the nature of the prediction error computed during predictive coding: it does not directly correspond to the colloquial sense of the term of "error" as

referring to a violation of prior expectations. Instead it simply refers to residual information within new bottom-up input that hasn't already been reconstructed by higher level of representation.

Finally, we were able to simulate the semantic overlap effect (Federmeier & Kutas, 1999; Holcomb & Neville, 1991), in which the lexicosemantic prediction error evoked by lexically unexpected words is smaller when these words share semantic features with the expected word. In our simulations, this occurred because our model assumed distributed representations that sometimes converge on individual lexical items, but that were shared across words. Therefore, even when we strongly pre-activated the model with a specific lexical item (e.g. *lion*, simulating a high constraint context, like *At the zoo, we heard the roaring of the* ___), this resulted in the pre-activation of distributed semantic features (e.g. <organism>, <animate>), resulting in an attenuation of the lexicosemantic prediction error produced by incoming words like *frog*, that shared some of these pre-activated features. Moreover, again consistent with empirical findings, which show that the semantic overlap effect is greater in high constraint than medium constraint contexts (Federmeier & Kutas, 1999), the influence of semantic overlap on the elicited errors was larger with greater pre-activation (see Fig 11).

In addition to the main effects of lexical and contextual factors on the N400, our simulations also captured the suppression of lexical main effects by contextual factors. In particular, it is well established that the effects of frequency and concreteness are attenuated by repetition (repetition x frequency: Rugg, 1990; repetition x concreteness: Kounios & Holcomb, 1994) and when the word is predictable in sentence context (cloze x frequency: Dambacher, Kliegl, Hofmann, & Jacobs, 2006; cloze x concreteness: Holcomb, Kounios, Anderson, & West, 1999), and these patterns were observed in our simulations.

As emphasized above regarding the frequency effect, weight-encoded biases are only expected to influence processing when state unit activity is ambiguous. On the other hand, when a given unit is sufficiently active relative to others – e.g., due to repetition or contextual predictability – the processing advantage (or disadvantage) conferred by frequency is expected to be undetectable. We believe that this is the basis for the interaction of frequency with repetition and cloze in our model.

Regarding the concreteness effect, we noted that the activation of more semantic state units (which correspond to a concrete bottom-up input) will generate a larger *total* semantic error for a given amount of unpredicted information per unit. For example, suppose u is the amount of error per unit, and suppose it is constant across different units (which is roughly true): a concrete word with 18 semantic features will elicit a total error of $18u$, and an abstract word with 9 features will elicit $9u$ of error: the difference is $18u - 9u = 9u$. However, when the semantic state units are sufficiently active – as might happen during repetition or contextual predictability – the amount of unpredicted information per unit will drop to u'. The total difference in error for a concrete vs abstract input is then given by $18u' - 9u' = 9u'$, which will be smaller than the original difference $9u$. This explains why the concreteness effect is expected to be suppressed when a word is predictable or repeated.

In addition to the interactions described above, we also observed clear reductions in the orthographic neighborhood effect for predictable words. Previously, some ERP experiments have suggested that orthographic neighborhood effects are unaffected by top-down context (Payne & Federmeier, 2018; Payne, Lee & Federmeier, 2015), but these studies investigated the role of word position (earlier vs. later in a sentence) rather than directly manipulating cloze probability. In contrast, another study (Molinaro et al., 2010) orthogonally manipulated both cloze and the

relative frequency of orthographic neighbors, which is also known to influence lexical competition (Grainger, 1990). This ERP study found an attenuation of the N400 neighborhood frequency effect for highly predictable words, which is generally in line with our current simulations. In order to test the predictions of this computational model more directly, additional N400 evidence will be needed exploring the interactions between orthographic competition and top-down contextual constraint. Our simulations found that the orthographic neighborhood size effect on the N400 was also suppressed when the word was repeated, which was somewhat supported by the only study that investigated this relationship (Laszlo & Federmeier, 2011). They found that the neighborhood size effect was slightly reduced on second presentation of the word, but no statistical tests were reported.

Relation to Other Models

Although there have been no previous attempts to model lexico-semantic activity within the predictive coding framework, other types of neural network models have been used to simulate various different lexical and contextual effects on the N400. We next discuss commonalities and differences between these previous models and the approach we took here.

One class of models largely focuses on modeling the sensitivity of the N400 to various lexical factors, as well as in minimal contexts (such as in semantic or repetition priming). In one approach, pioneered by Laszlo and Plaut (2012) (and subsequently extended by Cheyette & Plaut, 2017; Laszlo & Armstrong, 2014), the overall goal was to demonstrate how the mean semantic activity within a neurally plausible architecture could track the temporal dynamics of the N400 and its sensitivity to various bottom-up and top-down influences. In particular, they were committed to an architecture that separated excitatory from inhibitory connections and limited the number and distribution of inhibitory connections relative to the excitatory

connections. Under this design, a new bottom-up input causes a transient over-activation in the semantic units before it is inhibited, accounting for the characteristic rise and fall of the N400 response over time. Because inputs with more orthographic neighbors partially activate the semantic features of a larger number of other words, Laszlo & Plaut (2012) were able to simulate the effect of orthographic neighborhood size on the N400. Laszlo & Armstrong (2014) modified the model to account for repetition priming by simulating unit fatigue in response to repeated stimulation. Using a different fatigue function, Cheyette & Plaut (2017) set up a modified version of this model to simulate a set of contextual and lexical effects and their interactions (see Table 1). In addition to the simulations above, they simulated the concreteness effect (concrete words activate more features than abstract words), the frequency effect (low frequency words were less effective than over-trained high frequency words at suppressing features of their orthographic neighbors), and the interaction of both of these effects with repetition (because a low frequency/high concreteness prime elicits greater activation, it will cause larger fatigue at the target units, dulling subsequent effects of frequency and concreteness).

In the present model, we observed the above lexical effects for the same high-level reasons. The effect of orthographic neighborhood on the N400 was driven by the co-activation of semantic features of lexical neighbors; concrete (vs abstract) inputs elicited a larger N400 because they activated a larger number of semantic units (which required a larger total error); and the frequency effect was driven by the superior ability of high (vs low) frequency inputs to reconstruct the input with the same amount of activation, essentially outcompeting co-activated units. In contrast, the present model replicated the contextual patterns described above for different (high-level) reasons. The repetition and semantic priming effects were observed due to state-unit driven *suppression* – not decay – of the error units. This highlights a fundamental

distinction between the set of models described above and our predictive coding model: the N400 was operationalized as the difference between expected and observed lexicosemantic activity, i.e. prediction error. The activation and suppression of error units by state units explains the rise-and-fall morphology of the time course. As noted below, the functional segregation of units into states and errors is biologically plausible, independently motivated, and is an empirically testable claim to the extent that state and error activity is dissociable.

The idea that the N400 reflects prediction error – the difference between predictions that were implicitly encoded within the brain and inputs from the environment – inspired the approach proposed by Rabovsky & McRae (2014). These authors demonstrated that the N400 is best operationalized within their attractor network as an "implicit prediction error": the cross-entropy between semantic activity induced by bottom-up input (the prediction) and an "ideal" target word, in the sense of supervised learning. They showed that the magnitude of this error patterned with the N400 under many of the experimental conditions presented above (see Table 1). However, the fact that the "ideal target" can only be externally provided by the modeler poses a challenge for neurobiological plausibility - it is not clear even in principle how the correct target might enter the system in order for the learning signal to be computed. Although this is a general problem that all models relying on supervised learning (e.g., Cheyette & Plaut, 2017) must address, it is particularly salient for this model given that the cross-entropy error is the main outcome measure. Our model bypasses this issue by providing only low-level "sensory" input at the bottom-most level, and generating the correct targets for higher layers dynamically over time[3].

[3] While this depended on hand-coded weights, these can be trained without externally-provided correct targets in principle (as described in Lee & Seung, 2001). However, training the model comes at the expense of having uninterpretable weights, which we wanted to avoid here (see McCloskey, 1991 for an elaboration on this point).

A second class of models have focused on simulating the effects of sentence context on the N400 (Brouwer et al., 2017; Fitz & Chang, 2019; Rabovsky et al., 2018). Their recurrent network architecture allows words in a sentence to be presented to the neural network in a sequence, with recurrent connections capturing task-dependent temporal dependencies between the word inputs (cf. Elman, 1990). By training on a task that often depended on the full sentence meaning (next-word prediction: Fitz & Chang, 2019; thematic-role assignment: Brouwer et al., 2017; Rabovsky et al., 2018), the internal state within each model learns to track an overall "event" representation from individual word inputs. In all of these models, the N400 in response to a given input was quantified as the change in state from before to after this input is observed, although this change is conceptualized differently for each model. In Rabovsky et al., (2018), it is quantified as the change induced by a new input on the layer implicitly encoding the full event representation (the sentence gestalt layer); in Brouwer et al. (2017), it is quantified as the change induced by a new input on an intermediate layer that maps from a lexical input to its semantic representation; and in Fitz & Chang (2019), it is quantified as the difference between a "top-down" lexical prediction generated by the model and the subsequent lexical input, and it is viewed as a learning signal. Although these models were focused on simulating sentence context effects, some could also simulate minimal context effects (word repetition and semantic priming: Fitz & Chang, 2019; Rabovsky et al., 2018) and even lexical frequency (Rabovsky et al., 2018).

The contextual representations in these models were more natural and principled than the dummy level we employed to simulate effects of context in the present model, and we return to this important difference later when discussing how the current model can be extended. Setting this point aside, however, what distinguishes these previous approaches from the present one is that the N400 signal was computed in a post-hoc manner by the modeler, and it therefore played

no on-line role in comprehension itself - the process of inferring meaning from the input. In fact, in Brouwer et al. (2017), the signal is never used by the model; and in Rabovsky et al. (2018), the signal is only used for one simulation (temporal difference learning as a mechanism of repetition x incongruity interaction). This is in stark contrast with predictive coding in which prediction error is computed as a necessary step directly in the service of probabilistic inference. As such, prediction error in our model (collectively measured as the N400[4]) plays a much more central role in comprehension. It constitutes the lexico-semantic information that is actually passed from lower to higher levels of cortical representation — the primary medium through which bottom-up information flows up the model.

To sum up, while previous computational models have primarily simulated either higher-level contextual effects (Brouwer et al., 2017; Fitz & Chang, 2019; Rabovsky et al., 2018) or lower-level lexical effects (Cheyette & Plaut, 2017; Laszlo & Armstrong, 2014; Laszlo & Plaut, 2012; Rabovsky & McRae, 2014) on the N400, we were able to capture a variety of both kinds of effects in our approach. Further, the rise-and-fall morphology of the N400 – captured in Cheyette & Plaut (2017), Laszlo & Armstrong (2014), and Laszlo & Plaut (2012) by constraining the architecture – fell out of the processing dynamics in our model with no additional constraints imposed. Most importantly, by explicitly representing unpredicted information in error units, we were able to formalize the qualitative intuition expressed in general predictive processing frameworks that the N400 reflects the amount of unpredicted lexico-semantic information encoded in an incoming input.

[4] While it can be argued that our model doesn't use the *lexicosemantic* prediction error explicitly, most of our simulated effects were manifest both at the lexical and semantic levels, whose errors the model does compute and use.

There are also several additional strengths of the approach taken here to model N400. It connects basic computational principles of language processing with a general algorithm that has successfully explained a wide range of neural and perceptual phenomena in visual neuroscience, including end-stopping (Rao & Ballard, 1999), repetition suppression (Auksztulewicz & Friston, 2016) attentional modulations of neural responses (Spratling, 2008; Spratling, 2014), bistable perception (Weilnhammer et al., 2017) and motion illusions (Lotter, Kreiman & Cox, 2020), as well as some aspects of high-level cognition (Spratling, 2016). While the representations on which these mechanisms operate will differ a lot between different domains, it is remarkable that we were able to model a key neural signature produced during language comprehension – the N400 – by minimally adapting this general approach that was not developed for language processing. This suggests that the basic principles of predictive coding may be used in the service of a wide range of functions.

Predictive coding also yields insights about the value and roles of prediction more generally. During language comprehension, we must infer the meaning of each incoming word as it becomes available in real time, and prediction has long been thought to play an important role in increasing both the speed and efficiency of this inferential process (Kuperberg & Jaeger, 2016). By using a prior context to generate top-down predictions of incoming inputs, comprehenders can gain a head start on processing. But the fact of the matter is that most sentences are novel and most words are unpredictable – why should comprehenders bother predicting if the predictions will nearly always be incorrect? Predictive coding offers a novel perspective on this issue: the iterative computation of predictions and prediction errors essentially enables a *self-supervised* mechanism that the brain can exploit to learn a richly structured internal model of its environment, addressing the problem of routing a "high-level"

learning signal into the brain. On this view, accuracy and speed are advantageous side effects in a system that is primarily exploiting predictions and errors in the service of short-term inference and long-term learning.

While our goal here is not to provide evidence that the brain implements a predictive coding scheme, the current model makes several predictions that can be tested in future empirical studies, which would provide that evidence. For example, the model characterizes the N400 as prediction error at both lexical and semantic levels of representation. Although it is difficult to distinguish these at the surface of the scalp (especially with electrophysiological methods), it is possible that the lexical and semantic errors stem from different underlying neural sources. For example, lexical activity has been more closely linked to activity within mid-lateral and ventral regions of the left temporal cortex (Lau et al., 2008), while semantic activity has been linked to activity within more anterior ventral and medial temporal regions (Ralph et al., 2017). In our simulations, the modulation of lexical and semantic prediction error mostly patterned together. However, there were some simulations where the two diverged, making clear predictions that can be tested in future MEG studies that have the spatial and temporal resolution to localize these effects. For example, we found that the concreteness effect was largely localized to the semantic level (see Fig. 6 Inset); the semantic priming effect was localized to the lexical and semantic level (see Fig. 8 Inset); and the effect of repetition was also pronounced at the orthographic level (see Fig. 7 Inset).

We have also emphasized that predictive coding posits the functional segregation of state and error units (Rao & Ballard, 1999; Friston, 2005; Spratling, 2017). As such, we expect to find neural evidence for the distinction between state activity and error activity in the brain. In particular, we argued that phase-locked evoked responses are primarily sensitive to prediction

error (Friston, 2005), and so it may be possible to dissociate this from representational activity (detected using other multivariate methods) that may not necessarily increase in amplitude with new inputs. For example, studies of low-level visual perception suggest that, post-stimulus onset, multivariate methods can detect sharpened neural activity to predicted inputs, even when they produce reduced overall activity (Kok, Jehee & DeLange, 2012; Bell et al., 2016).

Limitations & Future Directions

Our implementation of the predictive coding model was highly simplified: we used a limited vocabulary with artificial distributed semantic representations, provided contextual information in the form of external pre-activation (as opposed to RNNs which can handle context naturally), and opted to hand-code the weights between levels of representation. These choices were motivated by a need for maximal transparency and interpretability, since our goal was to illustrate *why* our model accounts for the relevant neural phenomena (see McClelland, 2009; McCloskey, 1991).

While the scope of the model was limited in a number of ways, none of these limitations are fundamental. Future versions of the model may be able to incorporate pretrained word embeddings (e.g., word2vec; Mikolov et al., 2013) as a proxy for semantic features; the weights can be trained (Lee & Seung, 2001); and the model's inability to form sentence representations from word sequences can begin to be addressed with a version of the model that can handle time-varying stimuli. Indeed, a dynamic predictive coding model with this property has been developed in the domain of vision (Rao, 1999; Jiang et al., 2021). There has also been work showing that predictive coding can be used to approximate back-propagation along arbitrary computation graphs (Millidge et al., 2020), allowing models that are already highly successful at natural language processing tasks like long-short term memory (LSTM; Hochreiter &

Schmidhuber, 1997) networks and potentially even transformers (Vaswani et al., 2017) to be implemented exclusively using predictive coding principles. This potentially allows us to test the predictive coding framework with more realistic stimuli.

Another important direction for future work is to simulate psycholinguistic behavioral findings with this framework (cf. Cheyette & Plaut, 2017; Laszlo & Plaut, 2012). This can readily be addressed with a behavioral response system that is a function of neural activity or prediction error. For example, for a lexical decision task, we might use a threshold on the steady state value of lexical error as a decision criterion. Finally, in order to provide an objective benchmark of the success of this model in explaining human neurophysiological responses, in future work we hope to directly map moment-by-moment prediction errors generated by the model to human subject single-item EEG data. This dataset could provide a valuable quantitative benchmark for comparing the merits of different N400 modeling approaches.

Appendix

Algorithm 1 Model Dynamics for Cloze, Constraint & Semantic Overlap Simulations

Initialize uniformly: $\mathbf{st}^0_{\text{orth}}$, $\mathbf{st}^0_{\text{lex}}$, $\mathbf{st}^0_{\text{sem}}$, $\mathbf{r}^0_{\text{orth}}$, $\mathbf{r}^0_{\text{lex}}$, $\mathbf{r}^0_{\text{sem}}$, $\mathbf{b}^0_{\text{orth}}$

Initialize to zero: $\mathbf{b}^0_{\text{lex}}, \mathbf{b}^0_{\text{sem}}$

$\epsilon_1 = 0.01, \epsilon_2 = 0.0001$

Preactivation Iterations $= 20$

for $t = 1$ **to** T **do**

  if $t \leq$ Preactivation Iterations **then**

    $\mathbf{st}^t_{\text{orth}} \leftarrow \max(\epsilon_2, \mathbf{st}^{t-1}_{\text{orth}}) \odot (\mathbf{b}^{t-1}_{\text{orth}})$

  else

    $\mathbf{st}^t_{\text{orth}} \leftarrow$ Bottom-Up Input

  end if

  $\mathbf{e}^t_{\text{orth}} \leftarrow \mathbf{st}^t_{\text{orth}} \oslash \max\left(\epsilon_1, \mathbf{r}^{t-1}_{\text{orth}}\right)$

  $\mathbf{b}^t_{\text{orth}} \leftarrow \mathbf{r}^{t-1}_{\text{orth}} \oslash \max\left(\epsilon_1, \mathbf{st}^t_{\text{orth}}\right)$

  $\mathbf{st}^t_{\text{lex}} \leftarrow \max(\epsilon_2, \mathbf{st}^{t-1}_{\text{lex}}) \odot (\mathbf{W_1}\mathbf{e}^t_{\text{orth}} + \mathbf{I_1}\mathbf{b}^{t-1}_{\text{lex}})$

  $\mathbf{e}^t_{\text{lex}} \leftarrow \mathbf{st}^t_{\text{lex}} \oslash \max\left(\epsilon_1, \mathbf{r}^{t-1}_{\text{lex}}\right)$

  $\mathbf{b}^t_{\text{lex}} \leftarrow \mathbf{r}^{t-1}_{\text{lex}} \oslash \max\left(\epsilon_1, \mathbf{st}^t_{\text{lex}}\right)$

  $\mathbf{r}^t_{\text{orth}} \leftarrow \mathbf{V_1}\mathbf{st}^t_{\text{lex}}$

  $\mathbf{st}^t_{\text{sem}} \leftarrow \max(\epsilon_2, \mathbf{st}^{t-1}_{\text{sem}}) \odot (\mathbf{W_2}\mathbf{e}^t_{\text{lex}} + \mathbf{I_2}\mathbf{b}^{t-1}_{\text{sem}})$

  $\mathbf{e}^t_{\text{sem}} \leftarrow \mathbf{st}^t_{\text{sem}} \oslash \max\left(\epsilon_1, \mathbf{r}^{t-1}_{\text{sem}}\right)$

  $\mathbf{b}^t_{\text{sem}} \leftarrow \mathbf{r}^{t-1}_{\text{sem}} \oslash \max\left(\epsilon_1, \mathbf{st}^t_{\text{sem}}\right)$

  $\mathbf{r}^t_{\text{lex}} \leftarrow \mathbf{V_2}\mathbf{st}^t_{\text{sem}}$

  if $t \leq$ Preactivation Iterations **then**

    $\mathbf{st}^t_{\text{ctx}} \leftarrow$ Top-Down Input

  else

    $\mathbf{st}^t_{\text{ctx}} \leftarrow \max(\epsilon_2, \mathbf{st}^{t-1}_{\text{ctx}}) \odot (\mathbf{W_3}\mathbf{e}^t_{\text{sem}})$

  end if

  $\mathbf{r}^t_{\text{sem}} \leftarrow \mathbf{V_3}\mathbf{st}^t_{\text{ctx}}$

end for

$\mathbf{st}$, $\mathbf{e}$, $\mathbf{b}$, $\mathbf{r}$ refer to states, bottom-up prediction errors, top-down biases, and reconstructions respectively. $\mathbf{W}$ is a matrix mapping from a lower to higher level of representation ($\mathbf{W_1}$: orthographic to lexical; $\mathbf{W_2}$: lexical to semantic; $\mathbf{W_3}$: semantic to contextual); $\mathbf{V}$ is a matrix mapping from a higher to a lower level of representation, and it is the transpose of $\mathbf{W}$. Superscripts denote the iteration during which the variable was computed, subscripts denote its level of representation, one of orthographic (orth), lexical (lex), semantic (sem) or contextual (ctx). $\mathbf{I_1}$ and $\mathbf{I_2}$ are identity matrices that were divided by the same normalization factors as the corresponding $\mathbf{W}$ and $\mathbf{V}$ matrices.

Note on uniform initialization: orthographic, lexical and semantic units were set to an activation of $\frac{1}{26}$, $\frac{1}{1579}$ and $\frac{1}{12929}$ respectively. Initializing $\mathbf{b}^0_{\text{lex}}$ and $\mathbf{b}^0_{\text{sem}}$ uniformly (instead of zero) did not influence our results.

References

Auksztulewicz, R., & Friston, K. (2016). Repetition suppression and its contextual determinants in predictive coding. cortex, 80, 125-140.

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.

Ballard, D. H., & Jehee, J. (2012). Dynamic coding of signed quantities in cortical feedback circuits. Front Psychol, 3, 254. doi:10.3389/fpsyg.2012.00254

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. Journal of memory and language, 68(3), 255-278.

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.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. arXiv preprint arXiv:1406.5823.

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

Bentin, S., McCarthy, G., & Wood, C. C. (1985). Event-related potentials, lexical decision and semantic priming. Electroencephalography and clinical Neurophysiology, 60(4), 343-355.

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

Brothers T, Noriega S, Kuperberg GR. "Proportional semantic pre-activation during sentence comprehension: Evidence from ERPs." Talk presented at the 33rd Annual CUNY Conference on Human Sentence Processing, University of Massachusetts, Amherst, MA. March 2020.

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

Brown, C., & Hagoort, P. (1993). The processing nature of the N400: Evidence from masked priming. Journal of cognitive neuroscience, 5(1), 34-44.

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. Behavioral Research Methods.;41:977-90.

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.

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

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.

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.

Denison, R. N., Piazza, E. A., & Silver, M. A. (2011). Predictive context influences perceptual selection during binocular rivalry. Frontiers in human neuroscience, 5, 166.

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

Elman, J. L., & McClelland, J. L. (1984). Speech perception as a cognitive process: The interactive activation model. In Speech and language (Vol. 10, pp. 337-374). Elsevier.

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

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

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.

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

Friston, K. (2005). A theory of cortical responses. Philosophical transactions of the Royal Society B: Biological sciences, 360(1456), 815-836.

Garrido, M. I., Kilner, J. M., Kiebel, S. J., & Friston, K. J. (2009). Dynamic causal modeling of the response to frequency deviants. Journal of Neurophysiology, 101(5), 2620-2631.

Grainger, J. (1990). Word frequency and neighborhood frequency effects in lexical decision and naming. Journal of memory and language, 29(2), 228-244.

Hagoort, P., Baggio, G., & Willems, R. M. (2009). Semantic unification. In The cognitive neurosciences, 4th ed. (pp. 819-836). MIT press.

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.

Hauk, O., Davis, M. H., Ford, M., Pulvermüller, F., & Marslen-Wilson, W. D. (2006). The time course of visual word recognition as revealed by linear regression analysis of ERP data. Neuroimage, 30(4), 1383-1400.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8), 1735-1780.

Hohwy, J., Roepstorff, A., & Friston, K. (2008). Predictive coding explains binocular rivalry: An epistemological review. Cognition, 108(3), 687-701.

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 PJ, Kounios J, Anderson JE, West WC. (1999). Dual-coding, context-availability, and concreteness effects in sentence comprehension: An electrophysiological investigation. Journal of Experimental Psychology: Learning, Memory, and Cognition.;25(3):721-42.

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.

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.

Jiang, L. P., Gklezakos, D. C., & Rao, R. P. (2021). Dynamic Predictive Coding with Hypernetworks. bioRxiv.

Jordan, M. I. (1986). Attractor dynamics and parallellism in a connectionist sequential machine. In Proceedings of the Eighth Annual Conference of the Cognitive Science Society (pp. 531-546s). Hillsdale, NJ: Erlbaum.

Keller, G. B., & Mrsic-Flogel, T. D. (2018). Predictive processing: a canonical cortical computation. Neuron, 100(2), 424-435.

Kok, P., Jehee, J. F., & De Lange, F. P. (2012). Less is more: expectation sharpens representations in the primary visual cortex. Neuron, 75(2), 265-270.

Kounios J, Holcomb PJ. (1994) Concreteness effects in semantic processing: ERP evidence supporting dual-coding theory. Journal of Experimental Psychology: Learning, Memory, and Cognition;20(4):804-23.

Kuperberg, G. R. (2016). Separate streams or probabilistic inference? What the N400 can tell us about the comprehension of events. Language, cognition and neuroscience, 31(5), 602-616.

Kuperberg, G. R., & Jaeger, T. F. (2016). What do we mean by prediction in language comprehension?. Language, cognition and neuroscience, 31(1), 32-59.

Kuperberg, G. R., Brothers, T., & Wlotko, E. W. (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.

Kutas, M. (1993). In the company of other words: Electrophysiological evidence for single-word and sentence context effects. Language and cognitive processes, 8(4), 533-572.

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.

Kutas, M., & Iragui, V. (1998). The N400 in a semantic categorization task across 6 decades. Electroencephalography and Clinical Neurophysiology/Evoked Potentials Section, 108(5), 456-471.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2015). Tests for random and fixed effects for linear mixed effect models (lmer objects of lme4 package). R package version 2.0-33. http://cran.r-project.org/package=lmerTest.

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

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 and Language, 132, 22-27.

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

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. (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.

Lee, T. S., & Mumford, D. (2003). Hierarchical Bayesian inference in the visual cortex. JOSA A, 20(7), 1434-1448.

Lee, D. D., & Seung, H. S. (2001). Algorithms for non-negative matrix factorization. In Advances in Neural Information Processing Systems 13 - Proceedings of the 2000 Conference, NIPS 2000 (Advances in Neural Information Processing Systems). Neural information processing systems foundation.

Lotter, W., Kreiman, G., & Cox, D. (2020). A neural network trained for prediction mimics diverse features of biological neurons and perception. Nature Machine Intelligence, 2(4), 210-219.

McClelland, J. L. (1994). The interaction of nature and nurture in development: A parallel distributed processing perspective.

McClelland, J. L. (2009). The place of modeling in cognitive science. Top Cogn Sci, 1(1), 11-38. doi:10.1111/j.1756-8765.2008.01003.x

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.

McCloskey, M. (1991). Networks and theories: The place of connectionism in cognitive science. Psychological science, 2(6), 387-395.

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

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

Millidge, B., Tschantz, A., & Buckley, C. L. (2020). Predictive coding approximates backprop along arbitrary computation graphs. arXiv preprint arXiv:2006.04182.

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

Molinaro, N., Conrad, M., Barber, H. A., & Carreiras, M. (2010). On the functional nature of the N400: Contrasting effects related to visual word recognition and contextual semantic integration. Cognitive Neuroscience, 1(1), 1-7.

Müller, O., Duñabeitia, J. A., & Carreiras, M. (2010). Orthographic and associative neighborhood density effects: What is shared, what is different?. Psychophysiology, 47(3), 455-466.

Mumford, D. (1992). On the computational architecture of the neocortex. Biological cybernetics, 66(3), 241-251.

Nagy, M. E., & Rugg, M. D. (1989). Modulation of Event□Related potentials by word repetition: The effects of Inter□Item lag. Psychophysiology, 26(4), 431-436.

Norris, D. (2006). The Bayesian reader: explaining word recognition as an optimal Bayesian decision process. Psychological review, 113(2), 327.

Payne, B. R., & Federmeier, K. D. (2018). Contextual constraints on lexico-semantic processing in aging: Evidence from single-word event-related brain potentials. Brain research, 1687, 117-128.

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.

R Core Team (2016). R: A language and environment for statistical computing. Vienna, Austria, ISBN 3-900051-07-0.

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.

Ralph, M. A. L., Jefferies, E., Patterson, K., & Rogers, T. T. (2017). The neural and computational bases of semantic cognition. Nature Reviews Neuroscience, 18(1), 42-55.

Rao, R. P. (1999). An optimal estimation approach to visual perception and learning. Vision research, 39(11), 1963-1989.

Rao, R. P., & 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.

Rugg, M. D. (1985). The effects of semantic priming and word repetition on event□related potentials. Psychophysiology, 22(6), 642-647.

Rugg, M. D. (1990). Event-related brain potentials dissociate repetition effects of high-and low-frequency words. Memory & cognition, 18(4), 367-379.

Rumelhart, D. E., & McClelland, J. L. (1982). An interactive activation model of context effects in letter perception: II. The contextual enhancement effect and some tests and extensions of the model. Psychological review, 89(1), 60.

Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. Biometrics bulletin, 2(6), 110-114.

Shipp, S. (2016). Neural elements for predictive coding. Frontiers in psychology, 7, 1792.

Sim, E. J., & Kiefer, M. (2005). Category-related brain activity to natural categories is associated with the retrieval of visual features: Evidence from repetition effects during visual and functional judgments. Cognitive Brain Research, 24(2), 260-273.

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. (2012b). 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. (2016b). A neural implementation of Bayesian inference based on predictive coding. Connection Science, 28(4), 346-383.

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. Computational intelligence and neuroscience, 2009.

Van Berkum, J. J., Hagoort, P., & Brown, C. M. (1999). Semantic integration in sentences and discourse: Evidence from the N400. Journal of cognitive neuroscience, 11(6), 657-671.

Van Petten, C., & Kutas, M. (1990). Interactions between sentence context and word frequency in event-related brain potentials. Memory & cognition, 18(4), 380-393.

Vilares, I., & Kording, K. (2011). Bayesian models: the structure of the world, uncertainty, behavior, and the brain. Annals of the New York Academy of Sciences, 1224(1), 22.

Wacongne, C., Changeux, J. P., & Dehaene, S. (2012). A neuronal model of predictive coding accounting for the mismatch negativity. Journal of Neuroscience, 32(11), 3665-3678.

Weilnhammer, V., Stuke, H., Hesselmann, G., Sterzer, P., & Schmack, K. (2017). A predictive coding account of bistable perception-a model-based fMRI study. PLoS computational biology, 13(5), e1005536.

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.

Taylor, W. L. (1953). “Cloze procedure”: A new tool for measuring readability. Journalism quarterly, 30(4), 415-433.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. arXiv preprint arXiv:1706.03762.

Yarkoni, T., Balota, D., & Yap, M. (2008). Moving beyond Coltheart's N: a new measure of orthographic similarity. Psychonomic Bulletin & Review, 15(5), 971-979. doi:10.3758/PBR.15.5.971

Table 1

Simulated N400 Effects

			N400 models							
			Word-based models					Sentence-based models		
			Present Work	Laszlo & Plaut (2012)	Laszlo & Armstrong (2014)	Cheyette & Plaut (2017)	Rabovsky & McRae (2014)	Brouwer, Crocker, Venhuizen & Hoeks (2017)	Rabovsky, Hansen & McClelland (2018)	Fitz & Chang (2019)
Simulated N400 effects	**Lexical-level**	Orthographic Neighborhood Size	✓	✓	×	✓	✓	×	×	×
		Lexical Frequency	✓	×	×	✓	✓	×	✓	×
		Concreteness/Semantic Richness	✓	×	×	✓	✓	×	×	×
	Contextual-level	Semantic Priming	✓	×	×	✓	✓	×	✓	×
		Associative Priming	×	×	×	✓	×	×	✓	×
		Repetition Priming	✓	×	✓	✓	✓	×	✓	✓
		Cloze probability	✓	×	×	×	×	×	✓	✓
		Related Cloze Anomaly	✓	×	×	×	×	×	✓	×
		Role Reversal Anomaly	×	×	×	×	×	✓	✓	×
		Semantic Incongruity	×	×	×	×	×	✓	✓	×
		Position in Sentence	×	×	×	×	×	×	✓	✓
		Word Order Violation	×	×	×	×	×	×	✓	×
		Constraint for Unexpected Endings	✓	×	×	×	×	×	✓	✓
		Linguistic Adaptation	×	×	×	×	×	×	✓	✓
	Interactions	Repetition x Frequency	✓	×	×	✓	✓	×	×	×
		Repetition x Concreteness	✓	×	×	✓	✓	×	×	×
		Repetition x Incongruity	×	×	×	×	×	×	✓	×
		Cloze x Frequency	✓	×	×	×	×	×	×	×
		Cloze x Concreteness	✓	×	×	×	×	×	×	×

Note. Summary for all simulations that were and were not carried out by word-based and sentence-based models of the N400.

Table 2

Semantic Features for Miniature Lexicon

		Lexical Item			
		lion	*frog*	*lime*	*corn*
Shared Semantic Features	SF1	<lion-frog>	<lion-frog>	<lime-corn>	<lime-corn>
	SF2	<lion-frog-lime-corn>	<lion-frog-lime-corn>	<lion-frog-lime-corn>	<lion-frog-lime-corn>

Note. Semantic features are defined for a miniature four-item lexicon using the following procedure. The first shared semantic feature (SF1) of a given word is shared by exactly one other word (e.g., the feature <lion-frog> is shared by the words lion and frog, and there is no other word that has that feature). Note that each semantic feature is distinct - <lion-frog> is not the feature <lion> combined with that of <frog>; rather, <lion-frog> might denote a more general feature like <animal>. The second shared semantic feature of a given word (SF2) is a distinct feature shared by exactly four words. Again, note that <lion-frog-lime-corn> is not the combination of the individual features <lion-frog>, <lime-corn> etc., but rather a distinct feature which might denote a more general semantic feature like <organism>. In general, the nth shared semantic feature is a distinct feature that is designed to be shared by exactly 2^n words.

Table 3

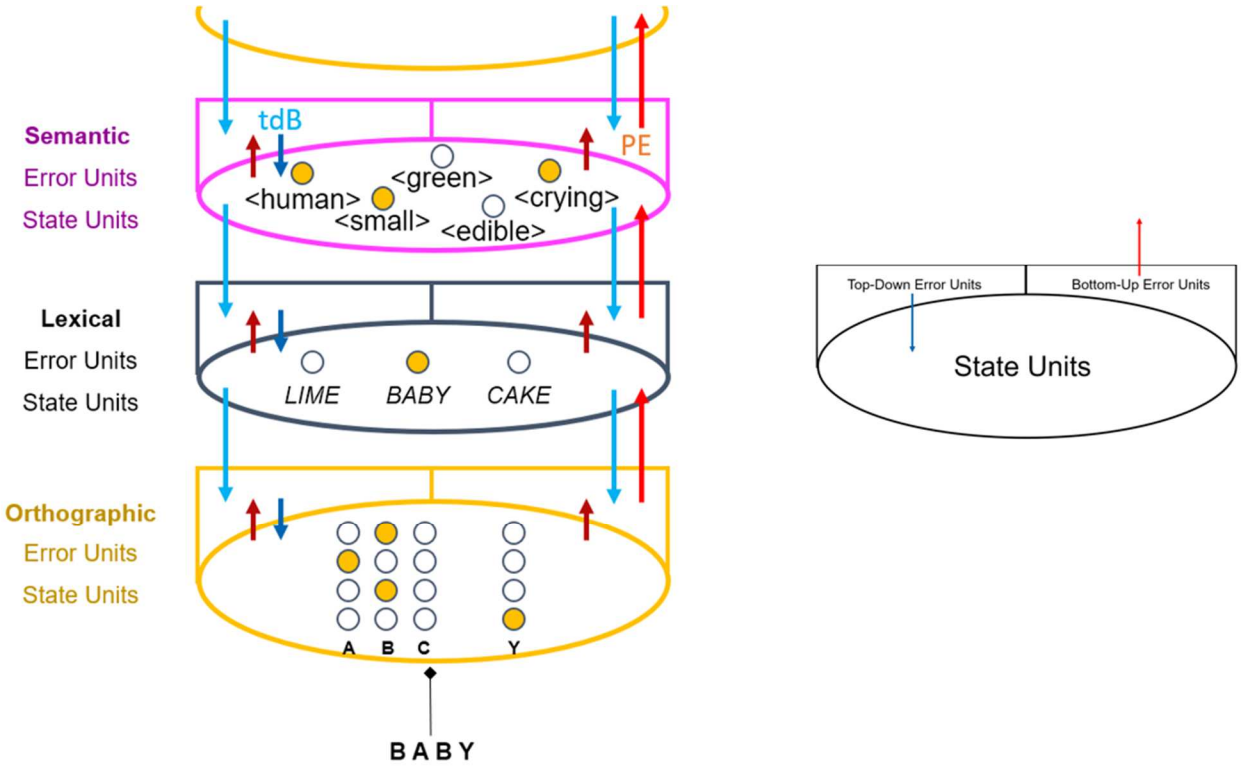
Cut-off Values for Frequency and Neighborhood Size

Factor	Cutoff
Orthographic Neighborhood Density (OLD 20)	OLD20 ≤1.7 is high neighborhood density, else low
Lexical Frequency	Lg10WF > 2.30 is high frequency, else low

Note. Summary of definitions of high versus low orthographic neighborhood density and lexical frequency.

Figure 1A

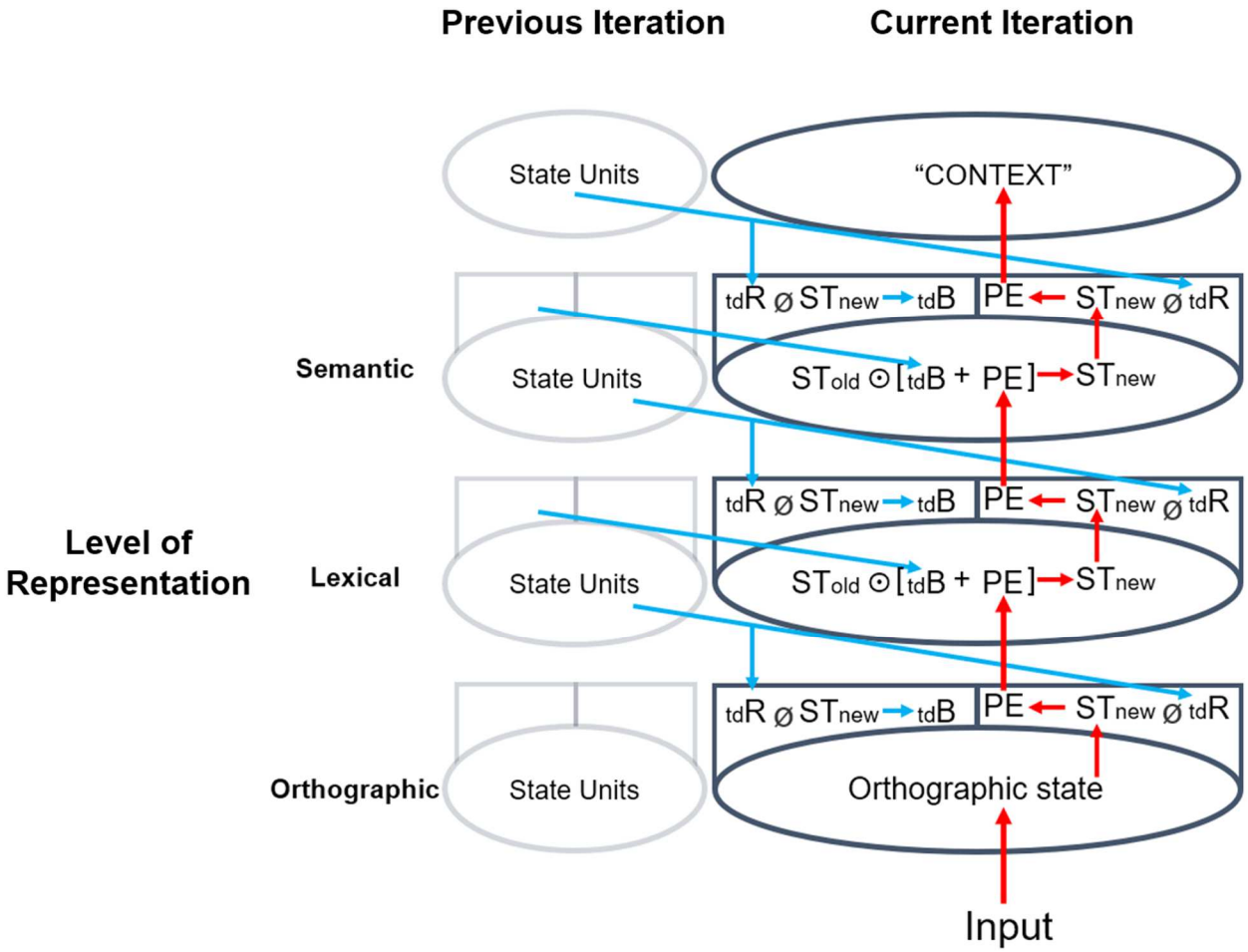
High-Level Model Schematic



Note. Schematic of the predictive coding model of lexicosemantic processing. State units (ovals) at a given level of cortical representation continuously generate top-down predictions or "reconstructions" (light blue arrow) that are actively propagated to the error units of the level below (the left & right halves of the "cap" on top of the oval). The prediction error (i.e., unpredicted bottom-up information) computed by the bottom-up error units is then passed back up (light red arrow) and used to update the higher-level state unit representations, thereby allowing them to generate more accurate top-down reconstructions on the next iteration of the algorithm; the top-down bias (unfulfilled predictions of the level above) computed by the top-down error units is copied to the state units at the same level (dark blue arrow) to bias their update at the next iteration.

Figure 1B

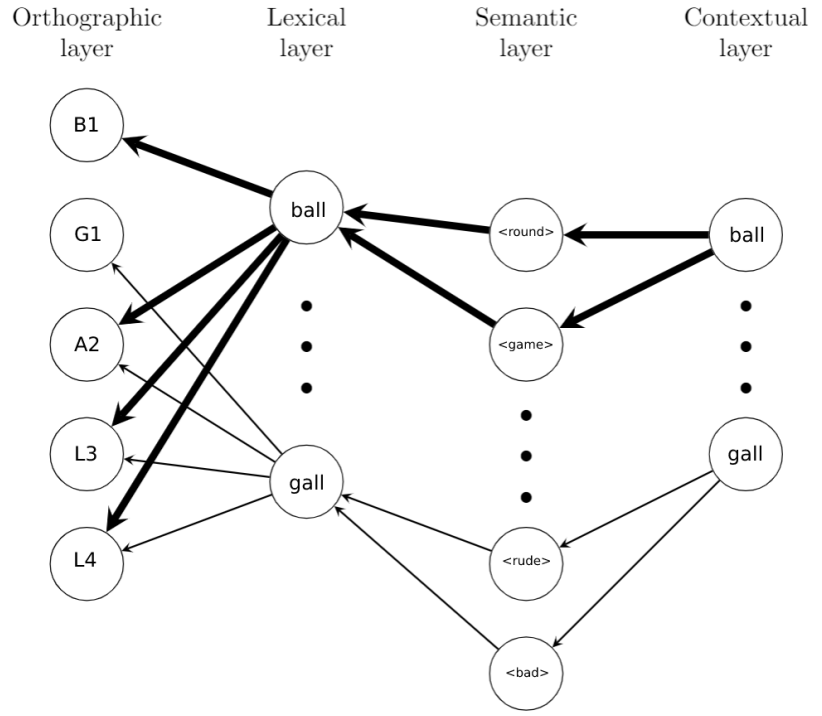
Model Architecture



Note. Schematic of the predictive coding architecture at two consecutive iterations. State units at different levels of representation interface at error units. Updated states (STnew) and prediction errors (PE) are passed up; top-down biases (tdB) and reconstructions (tdR) are passed down. This process repeats iteratively, and reconstructions improve with every iteration as the model settles on the beliefs most consistent with the bottom-up orthographic input.

Figure 2

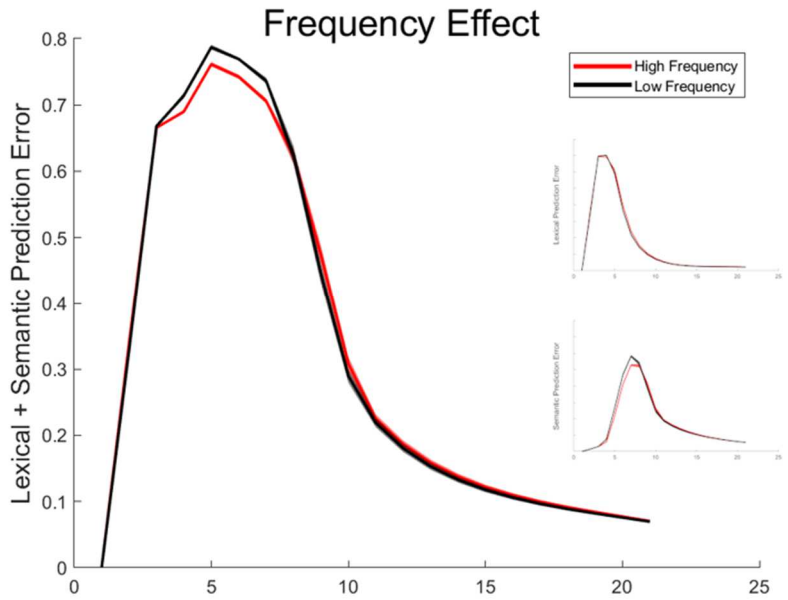
Feedback Weights were Frequency Sensitive



Note. Schematic emphasizing how the feedback weights associated with a frequent word (*ball*) were strengthened relative to an infrequent word (*gall*). Weight strength is indicated by arrow thickness.

Figure 3

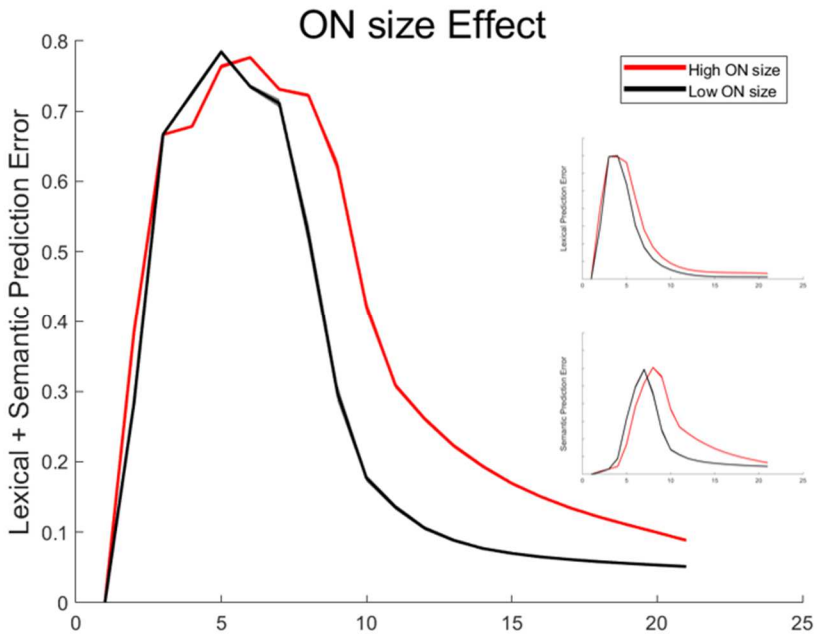
Frequency Effect



Note. Effect of frequency on the lexicosemantic prediction error, averaged over items from each category. X axis shows the number of iterations since stimulus onset. High frequency items elicited smaller lexicosemantic prediction error than low frequency items. Inset figure shows separate lexical (up) and semantic (down) errors for high and low frequency items. Units of error are arbitrary.

Figure 4

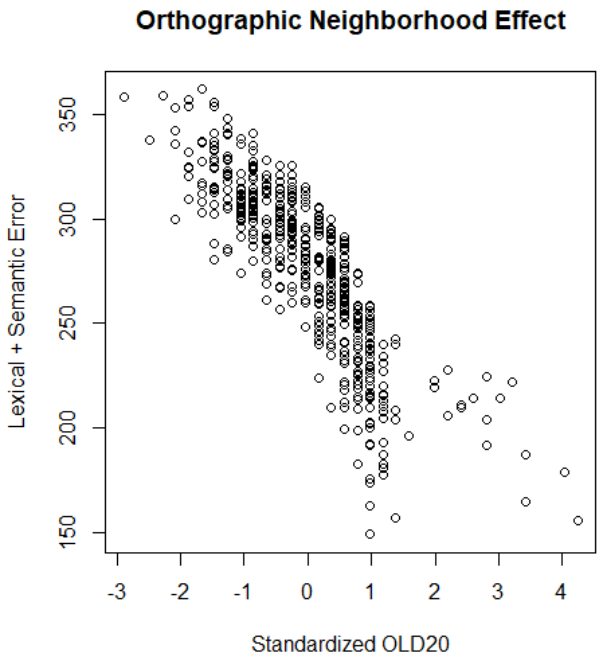
Effect of Orthographic Neighborhood Size



Note. Effect of orthographic neighborhood (ON) size on the lexicosemantic prediction error, averaged over items from each category. X axis shows the number of iterations since stimulus onset. Items with many orthographic neighbors (high ON size) elicited a larger lexicosemantic prediction error than items with fewer neighbors (low ON size). Inset figure shows separate lexical (up) and semantic (down) errors for high and low ON size items.

Figure 5

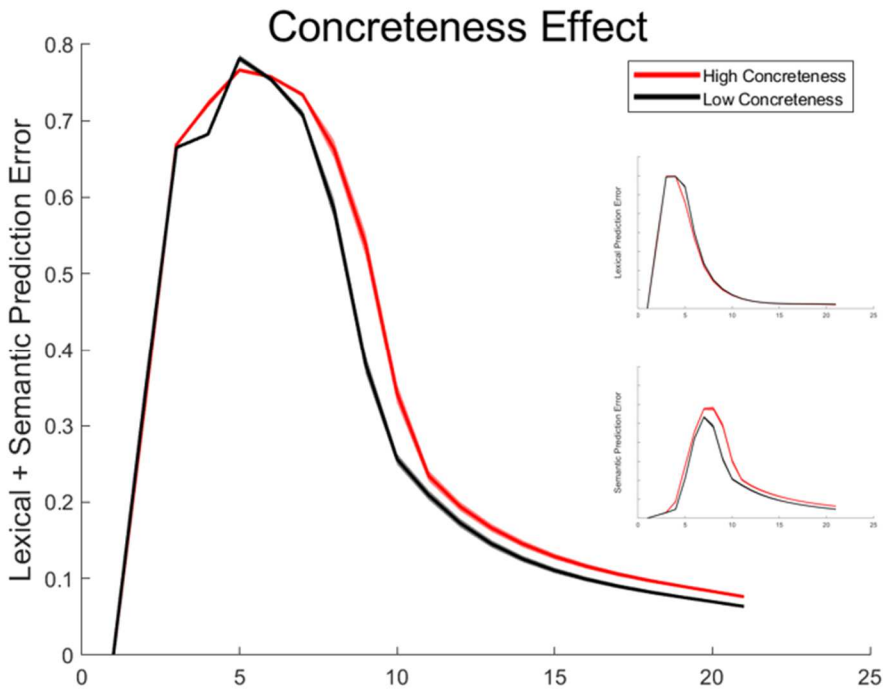
Continuous Effect of Orthographic Neighborhood Size



Note. Lexicosemantic error decreased continuously over the range of OLD20 values.

Figure 6

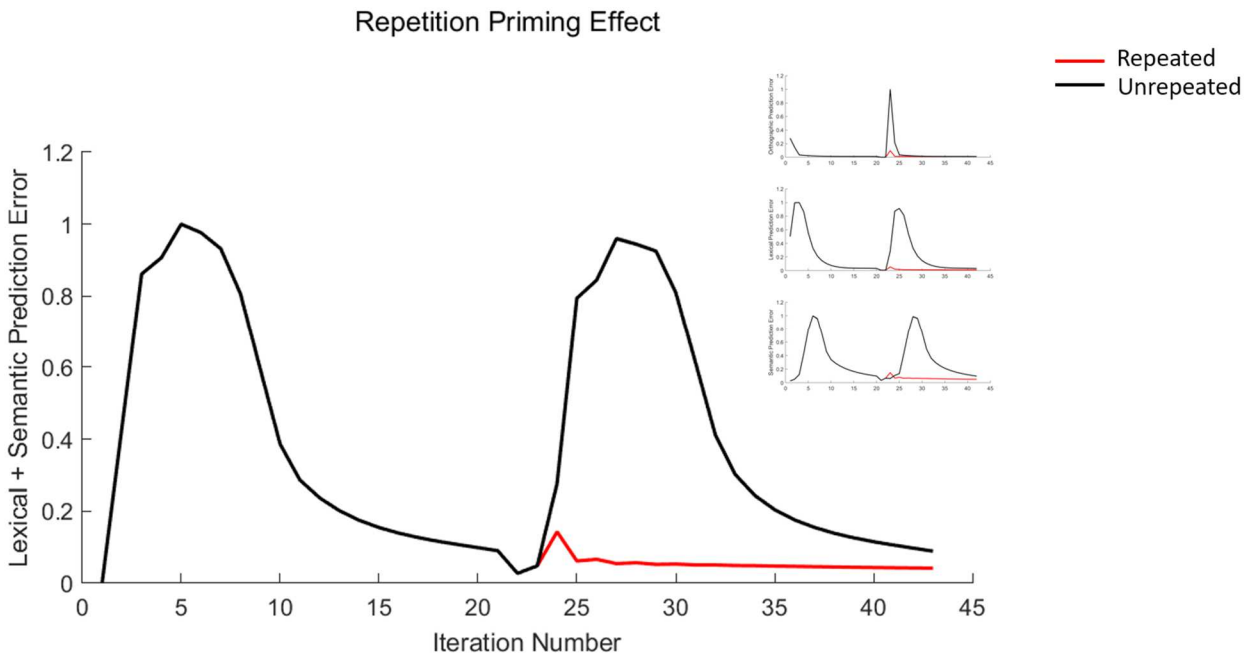
Concreteness Effect



Note. Effect of concreteness on the lexicosemantic prediction error, averaged over items from each category. X axis shows the number of iterations since stimulus onset. High concreteness items (those associated with a larger number of semantic features) elicited larger lexicosemantic prediction error than low concreteness items. Inset figure shows separate lexical (up) and semantic (down) errors for items with high and low concreteness. Note that the concreteness effect was limited to the semantic level.

Figure 7

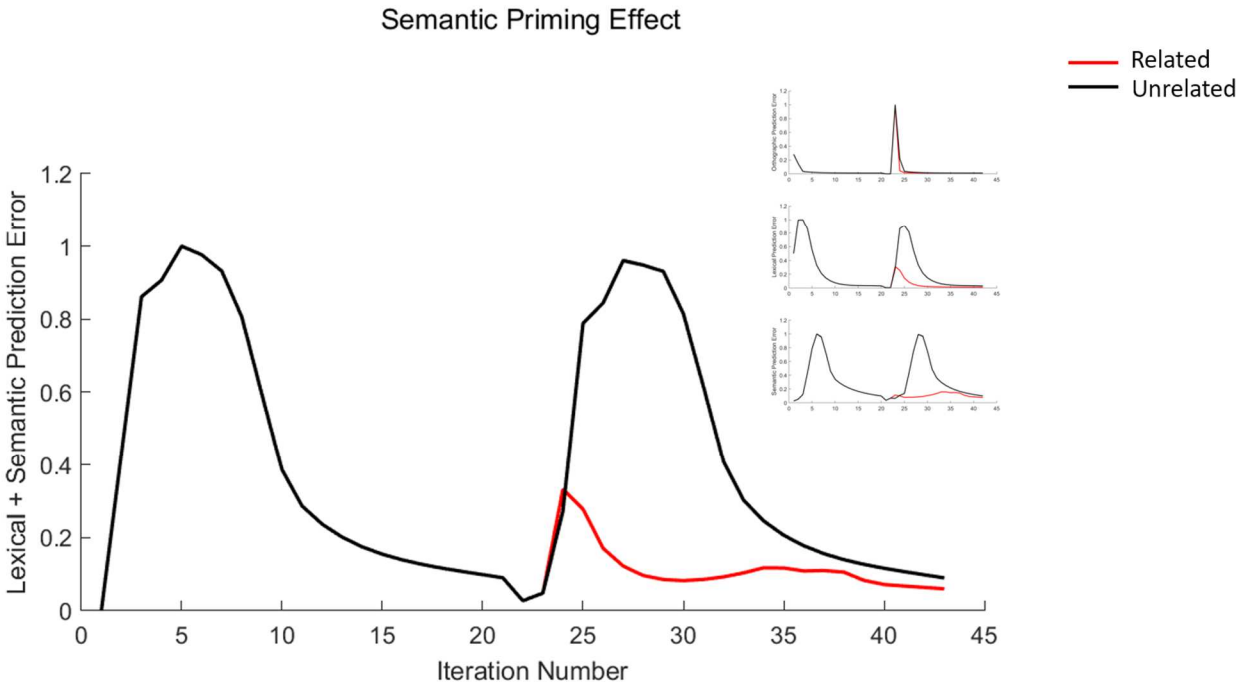
Repetition Priming Effect



Note. Effect of repetition priming on the lexicosemantic prediction error, averaged over items from each category. X axis shows the number of iterations since the onset of the prime stimulus. The target stimulus was presented at iteration #22. Repeated targets elicited a smaller lexicosemantic prediction error than unrepeated targets. Inset figure shows separate orthographic (top), lexical (middle) and semantic (bottom) errors. Note that the effect of repetition priming was pronounced at all levels.

Figure 8

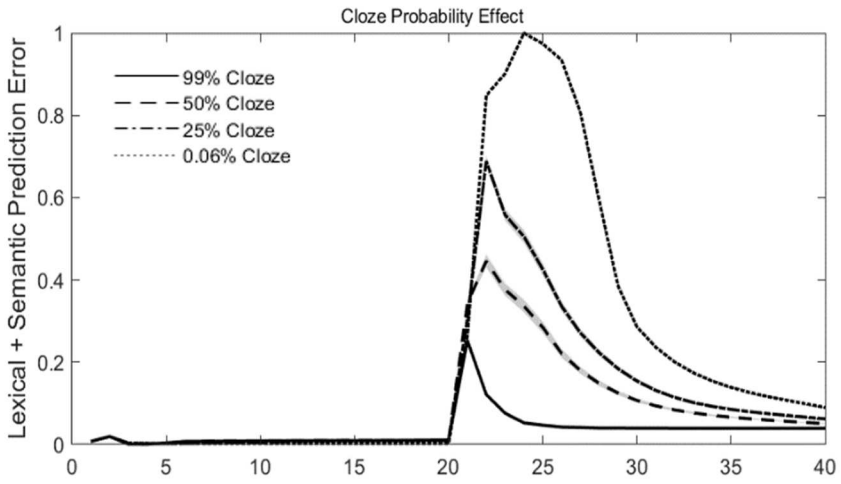
Semantic Priming Effect



Note. Effect of semantic priming on the lexicosemantic prediction error, averaged over items from each category. X axis shows the number of iterations since the onset of the prime stimulus. The target stimulus was presented at iteration #22. Related targets elicited a smaller lexicosemantic prediction error than unrelated targets. Inset figure shows separate orthographic (top), lexical (middle) and semantic (bottom) errors. Note that the effect of semantic priming was limited to the lexical and semantic levels.

Figure 9

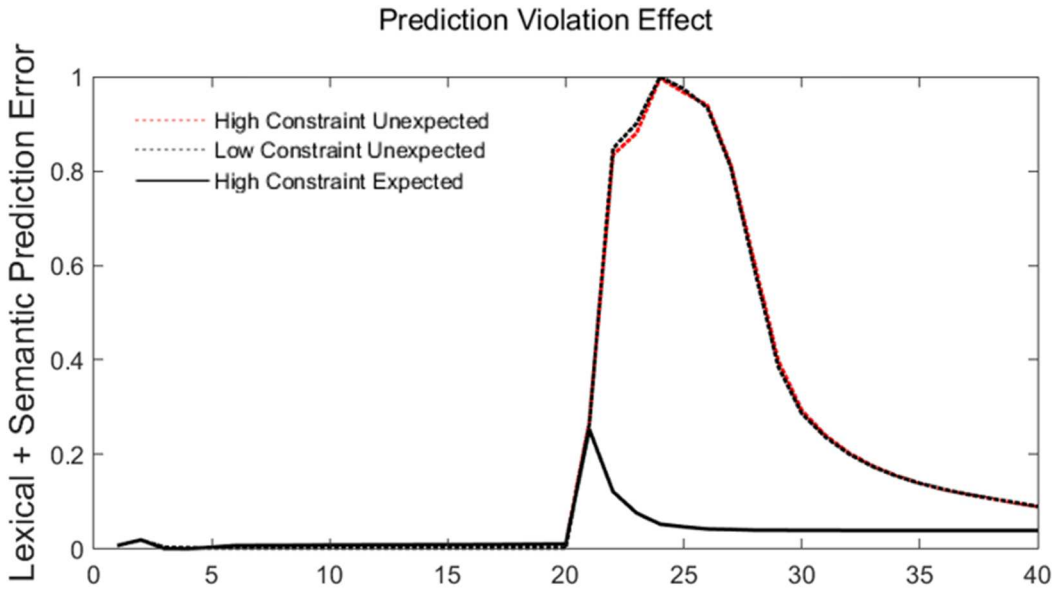
Cloze Probability Effect



Note. Effect of cloze probability on the lexicosemantic prediction error, averaged over all items from each cloze category. X axis shows the number of iterations since the onset of pre-activation. The bottom-up stimulus was presented at iteration #21. Higher cloze items elicited a smaller lexicosemantic prediction error than lower cloze items.

Figure 10

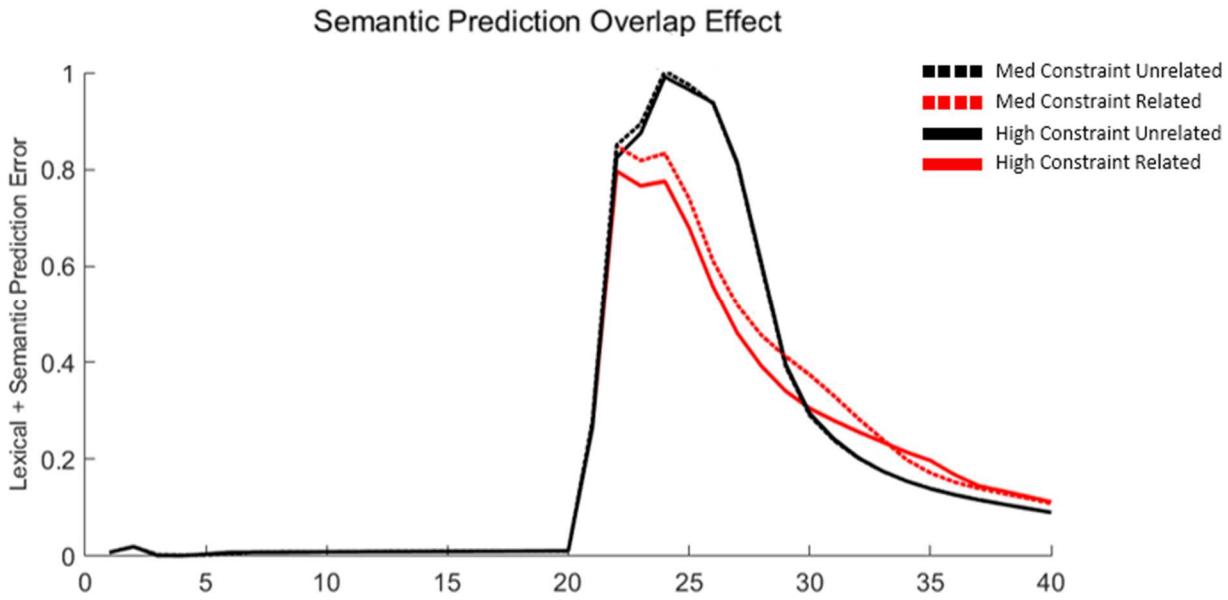
Effect of Constraint



Note. Effect of constraint on the lexicosemantic prediction error, averaged over all items from each category. X axis shows the number of iterations since the onset of pre-activation. The bottom-up stimulus was presented at iteration #21. High constraint (99%) unexpected items (*The groom took the bride's hand and placed the ring on her dresser*) elicited the same amount of lexicosemantic prediction error as low constraint unexpected items (*Helen reached up to dust the dresser*). Both unexpected conditions elicited a larger error than the high constraint expected condition (*The groom took the bride's hand and placed the ring on her finger*).

Figure 11

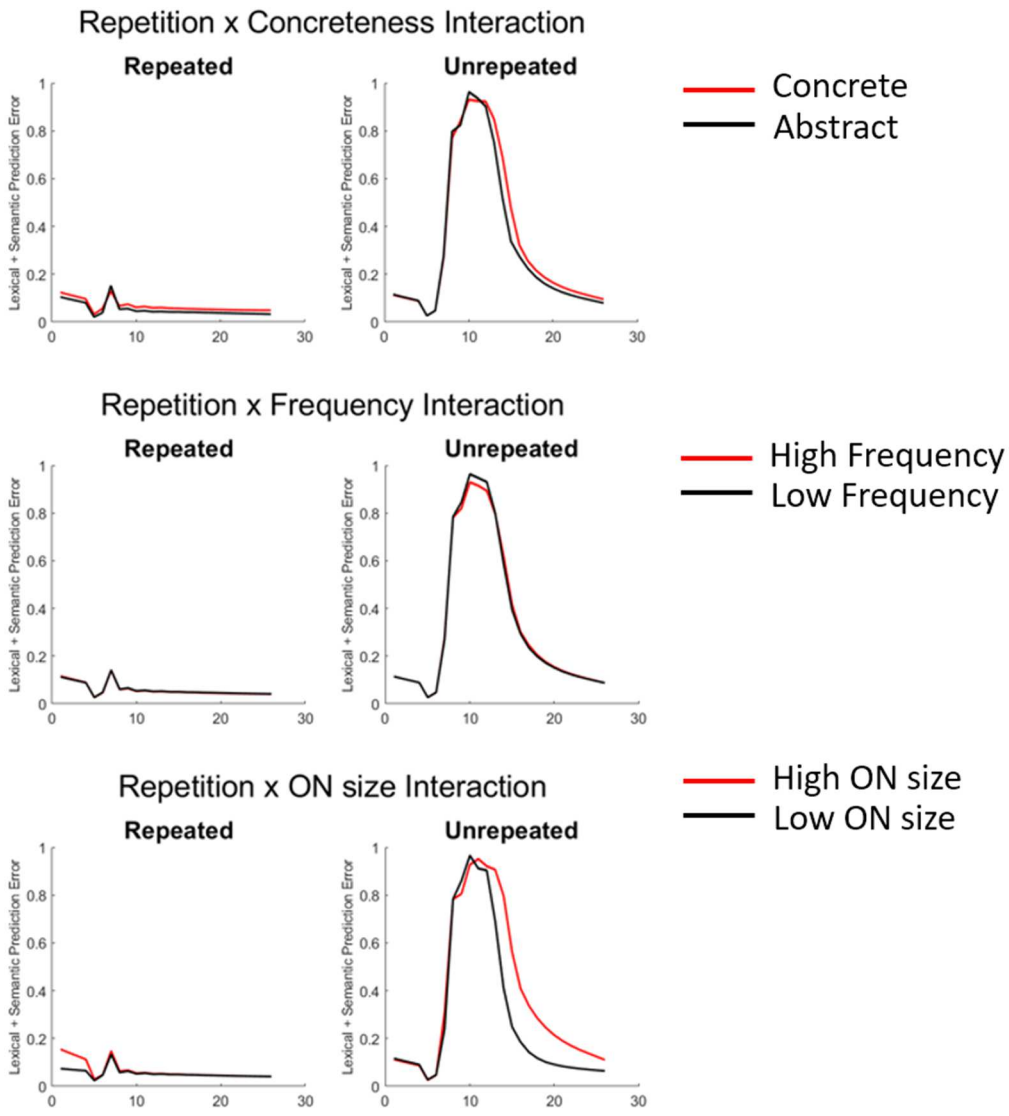
Semantic Prediction Overlap Effect



Note. When a particular item is predicted and an unexpected item is presented, the amount of error elicited by the latter is a function of its semantic overlap with the expected item. X axis shows the number of iterations since the onset of pre-activation. The bottom-up stimulus was presented at iteration #21. All bottom-up inputs were unexpected. When the input shared semantic features with the expected item (in high/99% and moderately/50% constraining contexts), it elicited smaller lexicosemantic prediction error than when it did not. This difference was larger in high (vs medium) constraint contexts.

Figure 12

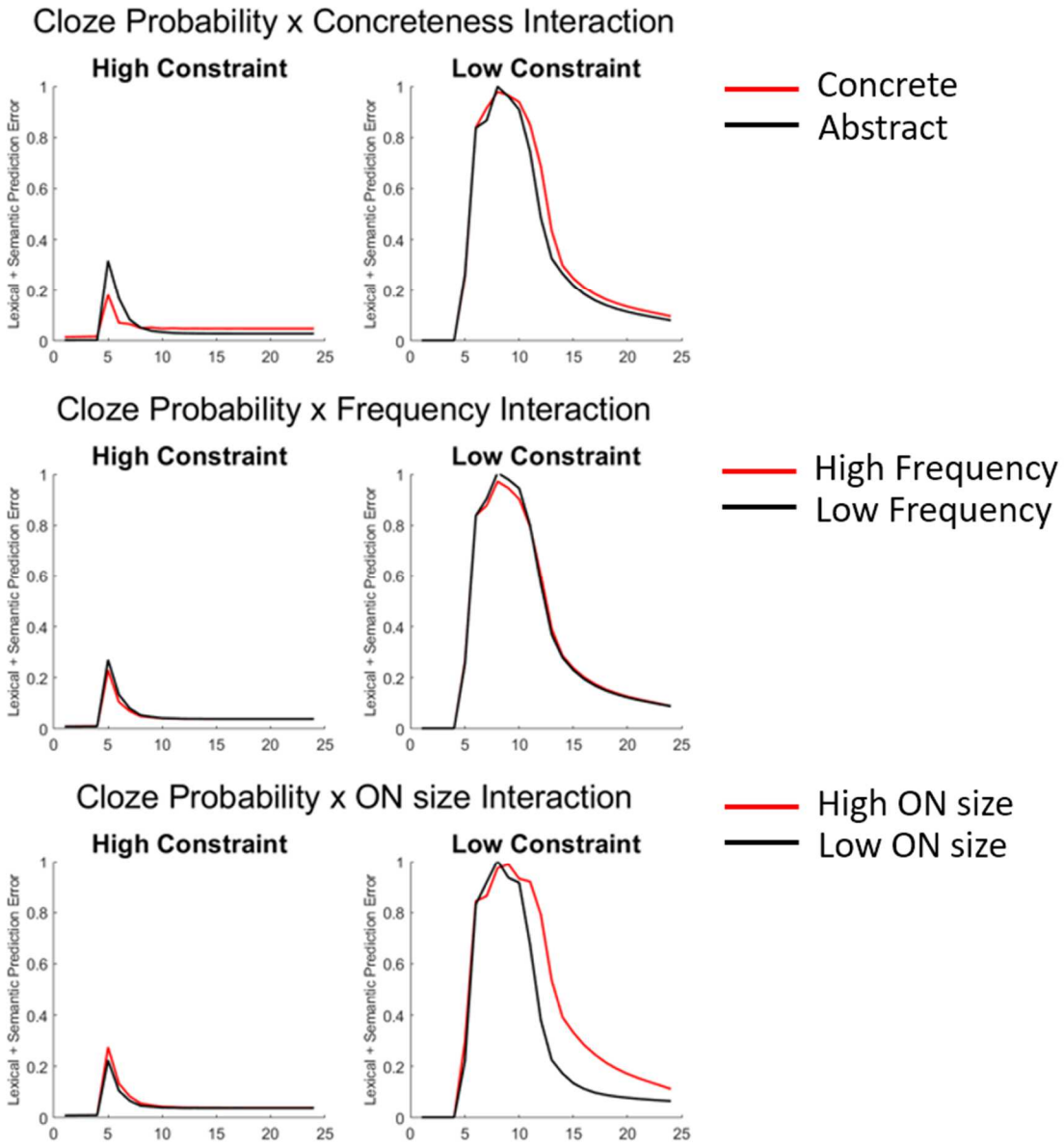
Interaction of Repetition with Lexical Factors



Note. The influence of all lexical factors (concreteness, frequency and orthographic neighborhood size) on the lexicosemantic prediction error was smaller for repeated versus unrepeated items. X axis shows five iterations prior to the onset of the repeated/unrepeated target.

Figure 13

Interaction of Cloze with Lexical Factors



Note. The influence of all lexical factors (concreteness, frequency and orthographic neighborhood size) on the lexicosemantic prediction error was smaller for high versus low cloze items. X axis shows five iterations prior to the onset of the high/low cloze bottom-up stimulus.