Cognition 246 (2024) 105755

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

A predictive coding model of the N400

Samer Nour Eddine [a,*], Trevor Brothers [a,c], Lin Wang [a,b], Michael Spratling [d], Gina R. Kuperberg [a,b]

[a] Department of Psychology and Center for Cognitive Science, Tufts University, United States of America
[b] Department of Psychiatry and the Athinoula A. Martinos Center for Biomedical Imaging, Massachusetts General Hospital, Harvard Medical School, United States of America
[c] Department of Psychology, North Carolina A&T, United States of America
[d] Department of Informatics, King's College London, United Kingdom

ARTICLE INFO

Keywords:
Language comprehension
Prediction
Prediction error
Orthographic
Semantic
Bayesian inference

ABSTRACT

The N400 event-related component has been widely used to investigate the neural mechanisms underlying real-time language comprehension. However, despite decades of research, there is still no unifying theory that can explain both its temporal dynamics and functional properties. In this work, we show that predictive coding – a biologically plausible algorithm for approximating Bayesian inference – offers a promising framework for characterizing the N400. Using an implemented predictive coding computational model, we demonstrate how the N400 can be formalized as the lexico-semantic prediction error produced as the brain infers meaning from the linguistic form of incoming words. We show that the magnitude of lexico-semantic prediction error mirrors the functional sensitivity of the N400 to various lexical variables, priming, contextual effects, as well as their higher-order interactions. We further show that the dynamics of the predictive coding algorithm provides a natural explanation for the temporal dynamics of the N400, and a biologically plausible link to neural activity. Together, these findings directly situate the N400 within the broader context of predictive coding research. More generally, they raise the possibility that the brain may use the same computational mechanism for inference across linguistic and non-linguistic domains.

1. Introduction

A key discovery in the history of psycholinguistics was the presence of a neural signature of online language processing — the N400 event-related potential (ERP; Kutas & Hillyard, 1980, 1984). Decades of research have established that the N400 indexes neural processes at the heart of semantic processing (Kutas & Federmeier, 2011). There has therefore been considerable interest in developing a theoretical framework for understanding the role it plays in language comprehension. This, however, has proved a formidable challenge. Several theories and computational models have provided compelling explanations for its functional properties. However, a unifying, biologically plausible account remains elusive. *Predictive coding* is a computational algorithm that has been proposed to carry out perceptual inference in the brain (Friston, 2005; Mumford, 1992; Rao & Ballard, 1999; Spratling, 2016b). Using an implemented predictive coding model of lexico-semantic processing, we show that the magnitude of lexico-semantic prediction error tracks the temporal dynamics of the N400 as well as its functional sensitivity to both lexical and contextual information. Together, these findings raise the possibility that the brain employs predictive coding to infer meaning from form, with the N400 playing a central role in this inferential process.

The N400 ERP is a negative-going waveform that is detected at the scalp surface using both electroencephalography (EEG) and magnetoencephalography (MEG) between 300 and 500 ms following the onset of any meaningful stimulus, such as a word or a picture (see Kutas & Federmeier, 2011, for a review). During language processing, the N400 is highly sensitive to the relationship between a word and its prior context, regardless of whether this context is a single word (in semantic and repetition priming paradigms, e.g. Bentin, McCarthy, & Wood, 1985; Rugg, 1985), or a more extended sentence or discourse context (e.g., DeLong, Urbach, & Kutas, 2005; Kutas & Hillyard, 1984; Van Berkum, Hagoort, & Brown, 1999). The N400 is also elicited by words presented out of context where its amplitude is sensitive to several lexical variables, including orthographic neighborhood size (e.g. *core* > *kiwi*; Holcomb, Grainger, & O'Rourke, 2002; Laszlo & Federmeier,

* Corresponding author at: Department of Psychology, Tufts University, 490 Boston Avenue, Medford, MA 02155, United States of America.
E-mail address: Samer.Nour_Eddine@tufts.edu (S. Nour Eddine).

https://doi.org/10.1016/j.cognition.2024.105755
Received 22 March 2023; Received in revised form 14 February 2024; Accepted 19 February 2024
Available online 29 February 2024

2011), lexical frequency (e.g. *wart* < *cold*; Rugg, 1990; Van Petten & Kutas, 1990), and concreteness/semantic richness (e.g. *lime* > *know*; Kounios & Holcomb, 1994; Holcomb, Kounios, Anderson, & West, 1999; Rabovsky, Sommer, & Abdel Rahman, 2012b).

Despite extensive work on the N400, there is still no general consensus on its functional significance. For many years, two competing theories dominated the debate: a lexical access and an integration account. Briefly, the lexical access account 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 links the fully accessed item with its prior context (Brown & Hagoort, 1993; Hagoort, Baggio, & Willems, 2009). However, as several researchers pointed out, this type of dichotomy between "access" and "integration" has difficulty in explaining the sensitivity of the N400 to *both* lexical and contextual factors (Baggio & Hagoort, 2011; Kuperberg, 2016; Kutas & Federmeier, 2011). More generally, this dichotomy 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 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 linguistic form of an individual word. This theory therefore provided some intuition for why the N400 is sensitive to both lexical variables and contextual predictability. For example, as a new bottom-up input activates its overlapping orthographic neighbors, the co-activation of their semantic features would result in an enhanced N400 response (see Laszlo & Federmeier, 2011). And if a prior context pre-activates expected upcoming semantic features, then the amplitude of the N400 to an incoming word that encodes the same features should be attenuated, even if that word is lexically unexpected (Federmeier & Kutas, 1999).

On the other hand, the theory's flexibility leaves a number of cognitive mechanisms 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 neighbors 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 (Brouwer, Crocker, Venhuizen, & Hoeks, 2017; Cheyette & Plaut, 2017; Fitz & Chang, 2019; Laszlo & Armstrong, 2014; Laszlo & Plaut, 2012; Rabovsky, 2020; Rabovsky, Hansen, & McClelland, 2018; Rabovsky & McRae, 2014); see Nour Eddine, Brothers, & Kuperberg, 2022 for a comprehensive review). Broadly, these models of the N400 fall into two classes: "word-level" and "sentence-level".

The word-level models were trained to map a single word-form input (e.g., a letter-string), clamped at the input layer, on to a pattern of activation that represented the word's meaning at the top (output) layer (Cheyette & Plaut, 2017; Laszlo & Armstrong, 2014; Laszlo & Plaut, 2012; Rabovsky & McRae, 2014). In one set of studies, Laszlo, Plaut, Armstrong and Cheyette used a biologically motivated Semantic Activation architecture to simulate the N400 as the total activity produced within its semantic (output) layer (Cheyette & Plaut, 2017; Laszlo & Armstrong, 2014; Laszlo & Plaut, 2012). This model was able to simulate the effects of several lexical variables on the N400 produced by words presented in isolation, including the effects of orthographic neighborhood size (Laszlo & Plaut, 2012), lexical frequency (Cheyette & Plaut, 2017), and semantic richness (Cheyette & Plaut, 2017). It was also able to simulate the attenuation of the N400 to target inputs in both repetition (Laszlo & Armstrong, 2014) and semantic priming paradigms (Cheyette & Plaut, 2017), see Table 1.

In another word-level model, Rabovsky and McRae (2014) simulated the N400 as the difference (cross-entropy error) between the activity produced by the model's semantic (output) layer and an ideal "correct" semantic target presented to the model. The authors showed that this operationalization of the N400 could account for a similar range of findings as above (see Table 1).

The sentence-level models were trained to map a sequence of word inputs onto a higher event-level representation (Brouwer et al., 2017; Rabovsky et al., 2018), or onto the model's prediction of a subsequent word (Fitz & Chang, 2019). These training goals required the model to retain a representation of the full sequence of prior inputs as well as to implicitly predict upcoming information. This was achieved by including a recurrent element in the network (cf. Elman, 1990; Elman & McClelland, 1984). The N400 was modeled either as the amount of change that the input induced within a particular hidden layer within the network (Brouwer et al., 2017; Rabovsky, 2020; Rabovsky et al., 2018), or as the difference between a next-word prediction that was explicitly generated by the model, and the word that was subsequently presented (Fitz & Chang, 2019). Together, these models were able to simulate multiple effects of a prior context on the N400 evoked by incoming words (see Table 1, and Nour Eddine et al., 2022 for a detailed review).

The architectures and assumptions of these different computational models of the N400 are quite different from one another. However, it is worth emphasizing that in all except the Semantic Activation model (Cheyette & Plaut, 2017; Laszlo & Armstrong, 2014; Laszlo & Plaut, 2012), the N400 was operationalized as a *difference* value that was calculated by the modeler outside the model's architecture. This difference value was conceptualized either as a "prediction error" (Rabovsky & McRae, 2014; Fitz & Chang, 2019, or as a "change-in-state" (Brouwer et al., 2017; Rabovsky et al., 2018). It was assumed to emerge either as a byproduct of other computations (Brouwer et al., 2017) and/or to serve as a signal for downstream learning (Fitz & Chang, 2019; Rabovsky et al., 2018; Rabovsky & McRae, 2014). In no case, however, did it play a direct functional role in comprehension itself. This is in contrast with predictive coding, which proposes that prediction error, computed locally at each level of representation, plays an integral role in the optimization algorithm that the brain uses to approximate inference, i.e., the process of inferring meaning from an input's linguistic form.[1]

Predictive coding refers to a biologically plausible computational architecture, with a particular arrangement of feed-forward and feedback connections, that implements an optimization algorithm approximating Bayesian inference.[2] It was initially proposed to explain extra-classical receptive field effects in the visual cortex (Rao & Ballard, 1999; see also Mumford, 1992), and was later expanded into a more general account of perceptual inference in the brain (Friston, 2005; Clark, 2013; see also Spratling, 2016b). In predictive coding, *prediction error* is defined as the residual information observed at a given level of the cortical hierarchy that cannot be explained by top-down predictions (or "reconstructions') that are generated by the level above. This error is encoded within "error units", and is passed up to the level above where it is used to modify representations encoded within functionally distinct "state units". As a result, these higher-level state units generate more

[1] As we elaborate further in the Discussion section, in predictive coding, prediction error can also, in principle, be used for downstream learning (Whittington & Bogacz, 2017; Millidge et al., 2020; Song et al., 2020).

[2] This is in contrast to a more general and non-specific use of the term "predictive coding" that has sometimes been used in the language comprehension literature to refer to any type of prediction or "prediction error" in the brain.

Table 1
Phenomena simulated by computational models of the N400. An overview of the range of N400 phenomena that have been modeled in the literature.

		Word-level models				Sentence-level models			
		Laszlo and Plaut (2012)	Laszlo and Armstrong (2014)	Cheyette and Plaut (2017)	Rabovsky and McRae (2014)	Brouwer et al. (2017)	Rabovsky et al. (2018); Rabovsky (2020)	Fitz and Chang (2019)	Lexico-semantic Predictive Coding^
Lexical variables	Orthographic neighborhood size	✓	—	✓	✓	—	—	—	✓
	Lexical frequency	—	—	✓	✓	—	✓	—	✓
	Concreteness/ Semantic richness	—	—	✓	✓	—	—	—	✓
	Lexical status	—	—	—	—	—	—	—	✓
Word-pair priming	Repetition priming	—	✓	✓	✓	—	✓	✓	✓
	Semantic priming	—	—	✓	✓	—	✓	—	✓
	Associative priming	—	—	✓	—	—	✓	—	—
Contextual effects	Lexical probability	—	—	—	—	—	✓	✓	✓
	Constraint for unexpected endings	—	—	—	—	—	✓	✓	✓
	Anticipatory semantic overlap	—	—	—	—	—	✓	—	✓
	Anticipatory orthographic overlap	—	—	—	—	—	—	—	✓
	Role reversal anomaly	—	—	—	—	✓	✓	—	—
	Semantic incongruity	—	—	—	—	✓	✓	—	—
	Position in sentence	—	—	—	—	—	✓	✓	—
	Word order violation	—	—	—	—	—	✓	—	—
	Linguistic adaptation	—	—	—	—	—	✓	✓	—
	Article prediction violation (a/an)	—	—	—	—	—	✓	—	—
Interactions	Constr. x Anticipatory sem. Overlap	—	—	—	—	—	—	—	✓
	Repetition x Frequency	—	—	✓	✓	—	—	—	✓
	Repetition x Semantic Richness	—	—	✓	✓	—	—	—	✓
	Repetition x Incongruity	—	—	—	—	—	✓	—	—
	Cloze x Frequency	—	—	—	—	—	—	—	✓
	Cloze x Semantic Richness	—	—	—	—	—	—	—	✓
Learning effects	Changes across development	—	—	—	—	—	✓	—	—
	L2 priming with min. L2 knowledge	—	—	—	—	—	✓	—	—

^We include the current Lexico-Semantic Predictive Coding model in this Table for completion. However the effects we simulated will only be reported in the Results section.

accurate top-down predictions on the next iteration of the algorithm, which suppress the lower-level prediction error. This process takes place at each level of the hierarchy such that, over multiple iterations, the magnitude of prediction error — the total activity produced by the error units — gradually decreases as the state units converge upon the representation that best explains the bottom-up input.

The idea that higher levels of a representational hierarchy generate top-down predictions that facilitate the processing of inputs at lower levels is largely consistent with how prediction is typically framed in more general psycholinguistic and neurobiological frameworks of language processing. According to many of these accounts, during incremental language comprehension, the brain continually generates top-down predictions, based on a high-level interpretation of the prior context, which facilitate the processing of incoming words whose semantic features match these predictions (DeLong et al., 2005; Federmeier, 2007; Kuperberg & Jaeger, 2016, Section 3.5). The amplitude of the N400 is conceptualized as reflecting the ease of accessing the semantic features of these incoming words (i.e., the ease of lexico-semantic access or retrieval), or as the amount of unpredicted lexico-semantic *information* encoded within the bottom-up input (see Kuperberg, 2016 for discussion). Indeed, some researchers have explicitly appealed to the principles of predictive coding to explain the functional role of the N400 (e.g. Bornkessel-Schlesewsky & Schlesewsky, 2019; Kuperberg et al., 2020; Rabovsky & McRae, 2014; Xiang & Kuperberg, 2015). To date, however, there have been no attempts to simulate the N400 (or any other language ERP component) using an implemented predictive coding model. Developing such an implementation is important not only in formalizing our intuitions about the role of prediction in comprehension, but also for linking the language system to more general computational mechanisms implicated in other perceptual and cognitive domains.

In the present work, we built a computational model of lexico-semantic processing that was based on exactly 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 operationalized the N400 as *lexico-semantic prediction error* — the total activity produced by error units at the semantic and lexical levels on each iteration of the algorithm, as the model inferred the meaning of orthographic inputs. We carried out a series of simulations to determine whether the principles of predictive coding can

account for the temporal dynamics of the N400, as well as its functional sensitivity to (1) various lexical variables, (2) priming, (3) contextual effects, and (4) their higher-order interactions.

2. Methods

2.1. Model architecture

The basic structure of the predictive coding model used in all simulations is shown in Fig. 1. It consisted three levels of linguistic representation (orthographic, lexical, semantic) and a fourth layer at the top that represented individual concepts that corresponded to each lexical item, and that allowed the modeler to provide probabilistic pre-activation to simulate the effects of context (see Simulations). Similar to classic connectionist architectures like Interactive Activation and Competition (IAC) (Chen & Mirman, 2012; McClelland & Rumelhart, 1981) and TRACE (McClelland & Elman, 1986), instead of training the model to learn its own internal representations, we incorporated interpretable psycholinguistic representations at each level of the hierarchy and hand-coded the weights that described the mappings between representations at successive levels. This enabled us to examine and describe the mechanics of the predictive coding scheme in terms of constructs that we were able to interpret directly (although we note that it is possible for the model to learn these parameters without any modification to the architecture — a point we return to in the Discussion).

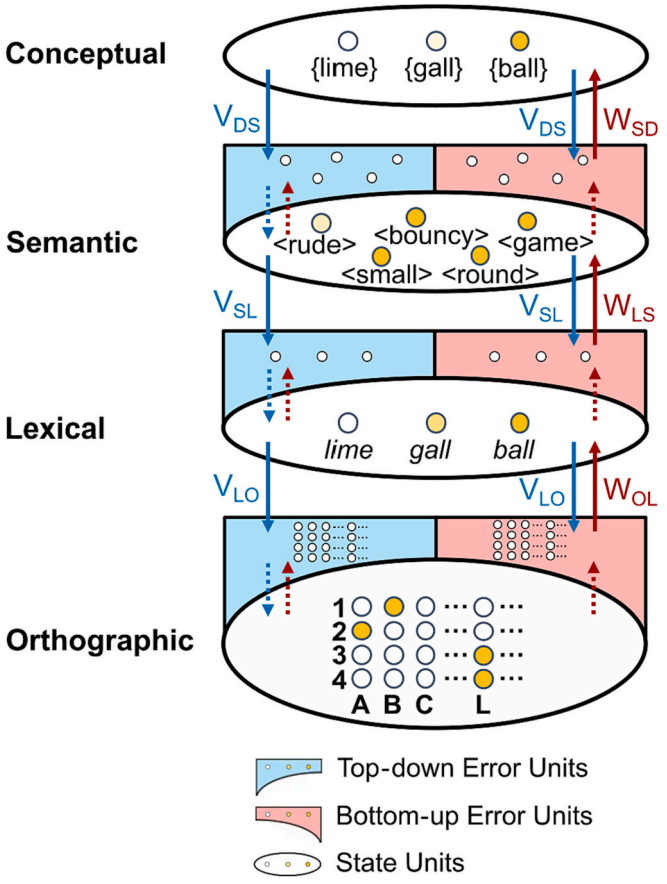

Fig. 1. Predictive coding model architecture. State units at three levels of linguistic representation (Orthographic, Lexical and Semantic) and at the highest conceptual layer are depicted as small circles within the large ovals. Error units at each of the three levels of linguistic representation are depicted as small circles within the half arcs. Dotted arrows indicate one-to-one connections between error and state units at the same level of representation. Solid arrows indicate many-to-many connections between error and state units across levels of representation. These many-to-many connections were specified using hand-coded weight matrices: W (feedforward) and V (feedback). V_{LO}/W_{OL}: Connections between the lexical and orthographic level; V_{SL}/W_{LS}: Connections between the semantic and lexical level; V_{DS}/W_{SD}: Connections between conceptual and semantic level. We schematically depict the activity pattern of the model's state units after it has settled on the representation of the item, *ball*. Different shades of yellow are used to indicate each state unit's strength of activity. At the Orthographic level, four state units are activated: B in the first position, A in the second position, and L in the final two positions. At the Lexical level, the unit corresponding to *ball* is mostly strongly activated, and its orthographic neighbor *gall* is partly activated because it shares three letters with *ball*. At the Semantic level, the units corresponding to the semantic features of *ball* (<bouncy>, etc.) are shown with different levels of activation. At the highest Conceptual layer, the unit corresponding to the representation of *ball* is most strongly activated. Because the model has settled, activity within error units at all levels is minimal. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

A key claim of predictive coding is that each layer has two types of connectionist units — *state units*, which encode the internal representations being inferred, and *error units,* which encode the *difference* in information (i.e., residual information) between that represented by state units at the same level and the information represented in state units at the level above. Across successive levels of the hierarchy, state units communicate exclusively through error units, which pass residual information between layers.

In our architecture, we implemented these basic principles by incorporating state units and two types of error units ("bottom-up" error units encoding "prediction error" and "top-down" error units encoding "top-down" error or bias, see Algorithm below) at each of the three linguistic layers of the architecture (the topmost conceptual layer only contained state units). Within each linguistic level, the state and error units shared one-to-one connections. Across successive levels, each higher-level state unit was connected to lower-level error units via many-to-many connections that we hand-coded using two matrices, V and W.

The V matrix coded the *feedback* connections that specified the *generative parameters* of the model; that is, each column of the V matrix specified how a given higher-level "latent cause" generated an idealized noise-free pattern of observations at the lower level. In our generative model, each individual concept at the highest level layer (e.g. {ball}) served as a latent cause of a distinct combination of semantic features (e.g. the combination of <round>, <game>, <small> and < bouncy>). In turn, each unique combination of semantic features served as a latent cause of a specific lexical representation (e.g. *ball*), which itself served as the latent cause for a distinct set of orthographic features at the lowest level of the model (e.g. "B-A-L-L"), see Fig. 2. The W matrix was simply the transpose of matrix V, and encoded the *feedforward* connections.

The lowest *orthographic* level included 104 sets of state/error units, each encoding one of 26 letter identities (A-Z) at one of four possible spatial positions (following McClelland & Rumelhart, 1981).

The middle *lexical* level included 1579 sets of state/error units, each representing a four-letter word in the model's lexicon (e.g., *baby, lime*). In most of our simulations, we used orthographic inputs that corresponded to 512 of the 1579 lexical units in the model's architecture. As discussed below, for each these *critical words,* orthographic neighborhood size, semantic richness and frequency (each operationalized as described below) were uncorrelated. The remaining words in the model's lexicon primarily served to increase the range and variability of orthographic neighborhood size for inputs presented during our simulations, as described below. The third *semantic* level included 12,929 state/error units, each representing a unique semantic feature (e.g., <small>, <human>; following Cheyette & Plaut, 2017; Rabovsky & McRae, 2014).

The *orthographic-lexical* matrices connected lower-level orthographic error units with higher-level lexical state units. These matrices specified the *spelling* of each word; that is, each column in the V matrix specified the mapping between a particular lexical item and the correct position of

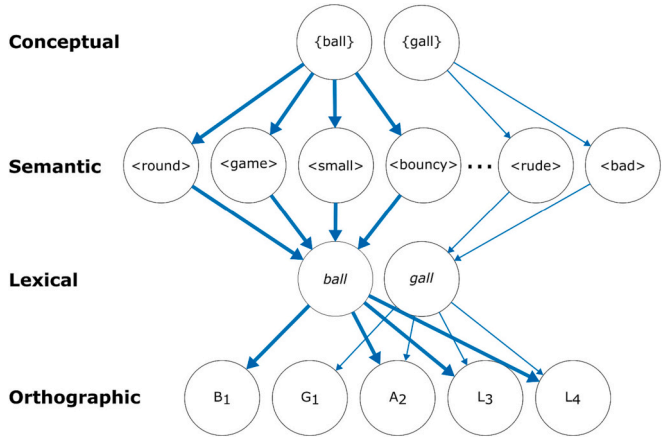

Fig. 2. Schematic illustration of the generative feedback connections for two words in the model's lexicon. Each circle indicates a representational node, and the blue arrows indicate feedback connections between layers. Note that, for simplification, this diagram does not distinguish between state and error units. In the model itself, however, the feedback connections linked higher-level state units with lower-level error units, see Fig. 1. To specify the frequency of each lexical item, we modified the connection strengths of its unique set of feedback connections. This is depicted schematically using arrow thickness. For example, the arrows are thicker for *ball* than *gall* because *ball* is more frequent. Although each lexical item has its own unique set of connections, these connections can terminate on shared nodes. For example, the lexical-orthographic feedback connections for *ball* and *gall* both terminate on the same A_2, L_3, and L_4 nodes, and the semantic-lexical feedback connections for the semantic features, <round>, <game>, <small> and < bouncy>, all terminate on the same lexical node, *ball*. In the model itself, this resulted in each lexical item having a particular "orthographic neighborhood size" and a particular "semantic richness". For example, *ball* and *gall* are orthographic neighbors, and the semantic richness of the word *ball* is greater than *gall* because the former lexical item is connected to more semantic features (4 vs. 2). For the purpose of our simulations, we defined each lexical item's orthographic neighborhood size as the number of lexical units with which it shared 3 letters. We defined "semantically rich" items as those that were connected to 18 features, and "non-rich" items as those that were connected to 9 features. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

each letter (e.g., *ball* → B in position 1, A in position 2, L in position 3, and L in position 4; see Fig. 2). This resulted in each lexical item having a particular orthographic neighborhood size, which, for the purpose of our simulations, was defined as the number of lexical units with which it shared 3 letters. Because the lexicon included >1500 words, the lexical items in the model took on a wide range of orthographic neighborhood sizes, ranging from 0 to 21.

The *lexico-semantic* matrices connected lower-level lexical error units with higher-level semantic state units. These matrices specified the *meaning* of each word; that is, each row in the V matrix specified the mappings between each lexical unit and its particular set of semantic features (e.g., *ball* → < round>, <bouncy>, etc.). To define the semantic richness of each of the 512 critical lexical units, we assigned half of these items 18 semantic features (semantically *rich* items) and the other half 9 semantic features (*non-rich* items), based on a median split on their concreteness ratings taken from Brysbaert, Warriner, and Kuperman (2014). Each of these 512 words shared between 0 and 8 semantic features with at least one other lexical unit, allowing us to simulate effects of semantic relatedness on the N400. For simplicity, each of the remaining 1067 lexical units was assigned 9 unique semantic features.

In order to set the frequency of lexical items within the model, we modified the strength of each item's unique set of feedback connections (see Fig. 2). In principle, there are multiple ways to encode lexical frequency into the model. Because knowledge of lexical frequency is acquired over a long period of time and reflects a stable prior (Norris, 2006), we chose to encode this bias into the top-down (feedback) weights of the network (see Spratling, 2016a), rather than as a more transient bias in the form of "resting level" activations in the state units (cf. McClelland & Rumelhart, 1981). Specifically, we increased the value of each word's entries in the V matrices with a value that was proportional to its SUBTLEX-US frequency (Brysbaert & New, 2009)). This score was obtained by shifting and scaling the log frequencies of all 1579 words in the model's lexicon so that they fell into the [0, 0.1] range. This range was selected so that the addition of frequency scores did not dramatically alter the mean and maximum value of the connections in any matrix.

2.2. The predictive coding algorithm

The architecture described above implemented the predictive coding algorithm — an optimization algorithm that approximates Bayesian inference, i.e. the process of inferring higher-level latent causes (i.e., columns of the V matrices) from particular patterns of observations (see Spratling, 2016a; Spratling, 2017). *State units* at each level represent the information that is being inferred, regardless of its predictability. The observed pattern encoded by state units at each level can be conceptualized as a dynamically changing "target" that higher-level state units are trying to reconstruct. *Error units* encode the *difference* (i.e. residual information) between the observed state patterns and the top-down predictions or reconstructions generated by the level above.

As noted under Architecture, we incorporated two types of error units that encode two types of errors (cf. Rao & Ballard, 1999, see Supplementary Materials). First, "bottom-up error units" represented *prediction error* — residual information encoded within the observed state patterns that was not present within the top-down reconstructions generated by state units at the level above. This prediction error served as an update signal of the state units at the level above so that they generated better reconstructions on the next iteration of the algorithm. Second, "top-down error units" computed residual information encoded within the top-down reconstructions that was not present within the observed state patterns. This served as a *top-down bias*[3] within state units for modifying the state patterns at the same level, so that they served as a better target for state units at the level above.

Over multiple iterations of the algorithm, as states units at each level are updated, both the magnitude of the prediction error and the top-down bias decrease, and the model reaches a global, internally consistent state that can accurately explain the bottom-up input at multiple levels of representation.

The specific predictive coding algorithm we used in the present study was a minimally modified version of the Predictive Coding/Biased Competition-Divisive Input Modulation algorithm (Spratling, 2008; Spratling, 2016b; see Supplementary Materials for details on how the original version was modified). This algorithm shares many processing principles with the influential predictive coding approach developed by Rao and Ballard (1999). However, the error units compute the residual information via element-wise division rather than element-wise subtraction (see Spratling, 2008). This ensures rapid convergence of the algorithm and guarantees that the activity across all units remains non-negative, similar to biological neurons.

As illustrated schematically in Fig. 3, at each iteration, *n*, of the predictive coding algorithm, the following processes occur in sequence at each level of the hierarchy.

1) The updating of state units

At each level, state units are updated based on (a) the bottom-up

[3] Note that this quantity is sometimes referred to as "top down error" (Rao & Ballard, 1999). Here, we refer to it as "top-down bias" to highlight its functional role as acting as a bias for updating state units.

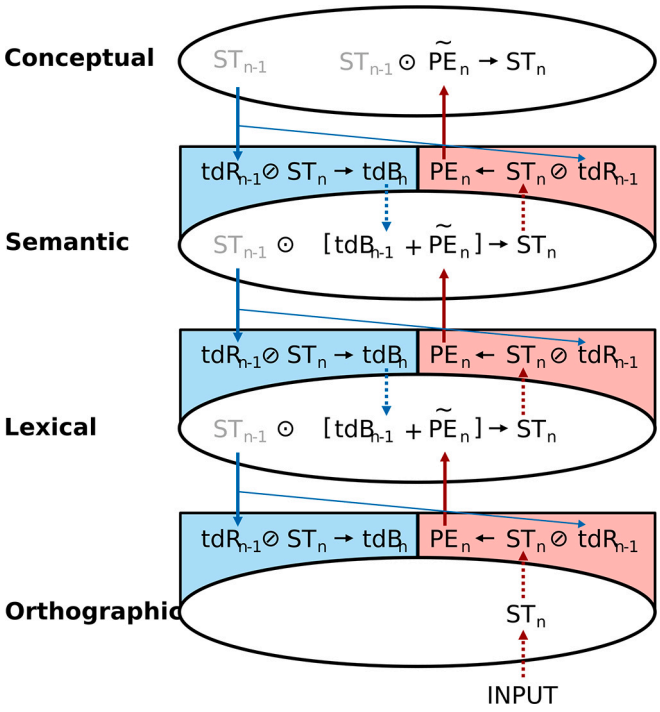

Fig. 3. Predictive coding algorithm. Schematic illustration of the predictive coding algorithm operating on the n^{th} iteration, following the presentation of bottom-up orthographic input. As in Fig. 1, at each layer, the large ovals contain state units, the red half-arcs contain bottom-up error units, and the blue half-arcs contain top-down error units. Each variable's subscript indicates the iteration on which it was computed. Solid arrows indicate the linear transformation of a variable through the V and W matrices. Dotted arrows indicate the copying of a variable. The same three steps occur in sequence at each level of representation: (1) State units are updated, based on (a) the top-down bias computed at the same level on the previous iteration, and (b) the prediction error computed at the level below on the same iteration ($ST_n \leftarrow ST_{n-1} \odot [tdB_{n-1} + \widetilde{PE}_n]$), and their values are copied to the top-down and bottom-up error units at the same level. (2) Bottom-up error units compute prediction error (PE_n) through elementwise division (ST_n Ø tdR_{n-1}) and pass this prediction error up to state units at the level above by transforming its dimensionality ($\widetilde{PE}_n = W \cdot PE_n$), and top-down error units compute top-down bias (tdB_n) and copy this top-down bias to state units at the same level so that it is ready to update the state units on the subsequent $[n+1]^{th}$ iteration); (3) State units generate top-down reconstructions of activity at the level below via linear transformation by the V (generative) matrix, i.e., V·ST = tdR, and pass these reconstructions down to the error units at the level below. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

input on the current iteration (n) and (b) a top-down bias (tdB) that was computed on the previous iteration (n-1). At the orthographic level, the bottom-up input is the orthographic vector provided by the modeler; at higher levels, the input is the prediction error (PE) that is computed at the level below on the current iteration of the algorithm. The updated state is computed through element-wise multiplication: $ST_n = ST_{n-1} \odot (tdB_{n-1} + \widetilde{PE}_n)$

Thus, at a given level of the hierarchy, state units are updated based on prediction error computed at the level below on the same iteration of the algorithm, so that these state units will generate more accurate top-down reconstructions of activity at the level below on the subsequent iteration of the algorithm. In contrast, the top-down bias, which was computed at the same level on the previous iteration of the algorithm, modifies the state pattern such that it is brought closer to the top-down reconstruction that was generated by the level above on the previous iteration, i.e., so that this state pattern serves as a better "target" for these higher-level reconstructions/predictions on the current iteration of the algorithm.

2) The computation of residual information: Prediction error and Top-down bias

The error units at each level represent the residual *difference* in information between the top-down reconstruction that was computed on the previous iteration of the algorithm (tdR_{n-1}) and the updated state pattern at the same level on the current iteration (ST_n). Within the bottom-up error units, the updated state pattern (at the same level) is divided elementwise by the top-down reconstruction (tdR) from the level above (which was computed on the previous iteration of the algorithm), yielding prediction error: $PE_n = ST_n$ Ø tdR_{n-1}. This prediction error encodes the residual information within the state units that is not present within the top-down reconstruction. It is multiplied by the W matrix ($\widetilde{PE} = W \cdot PE$), w hich transforms its dimensionality so that it can serve as the bottom-up input for updating state units at the level above on the current iteration of the algorithm.

Within the top-down error units, the top-down reconstruction, which was computed on the previous iteration of the algorithm, is divided elementwise by the current state through element-wise division, yielding a top-down Bias (tdB): $tdB_n = tdR_{n-1}$ Ø ST_n. This top-down bias encoded residual information within the top-down reconstructions that was not present within the state units. It was passed to state units at the same level through one-to-one connections, to be used to bias state updates on the subsequent iteration of the algorithm ($n+1$).

3) The generation of top-down reconstructions (predictions) to be used on the subsequent iteration

After the state units are updated at all levels, they generate a top-down reconstruction (tdR) of the pattern of activity at the level below by multiplying the current higher-level state vector (ST_n) by the V (generative) matrix, i.e., V· $ST_n = tdR_n$. This top-down reconstruction will be used to compute the prediction error and the top-down bias on the subsequent iteration of the algorithm (as described under step 2).

2.3. Simulations

We simulated a wide range of benchmark phenomena in the N400 literature. These included: (1) Lexical effects (effects of orthographic neighborhood size on words and pseudowords, and effects of lexical frequency and semantic richness on words); (2) Priming effects (repetition priming and semantic priming); (3) Contextual effects (lexical probability, contextual constraint, effects of anticipatory semantic overlap on words, and effects of anticipatory orthographic overlap on words and pseudowords), and (4) Interactions between each of the lexical variables with repetition priming and lexical probability. Details of how we carried out each specific simulation are given in the Results section.

For simulations involving only real words, we used orthographic input vectors that corresponded to 512 four-letter "critical words" in the model's lexicon. As noted above, these words were selected to have a wide range of lexical frequencies, orthographic neighborhood sizes, and semantic richness values (half rich, half non-rich), and, by design, these three variables were all uncorrelated ($|r| < 0.07$) across our 512 critical words. For simulations involving pseudowords, we used 400 words and 400 pseudowords that were matched on orthographic neighborhood size.

In all simulations, after initializing the model, we presented the bottom-up orthographic input by clamping an orthographic vector that corresponded to the stimulus of interest. We then ran the predictive coding algorithm for 20 iterations.

2.4. Operationalization of the N400 and visualizations

As explained above, in predictive coding, bottom-up error units

encodes "prediction error" — a vector-valued quantity that represents the residual information encoded by state units that is not already present within top-down reconstructions. Thus, the sum of all elements in each prediction error vector simply corresponds to the total activity produced by the bottom-up error units. Based on the large psycholinguistic literature linking the N400 to stimulus-induced processing at the lexical and semantic levels of representation, we elected, a priori, to operationalize the N400 as the total activity produced by these bottom-up error units at the lexical and semantic levels of representation at each iteration of the algorithm following stimulus presentation. We refer to this scalar value simply as *lexico-semantic prediction error*.

For each simulation, we constructed a full time course of the simulated N400 by averaging lexico-semantic prediction error across all critical items on each of the 20 iterations of the algorithm following stimulus presentation. For completeness, in Supplementary Materials we also show the time course of prediction error produced at the semantic, lexical, and orthographic levels separately (Supplementary Figs. 1, 2 and 3). Except where noted otherwise, variables that modulated lexico-semantic prediction error affected the prediction error produced at the lexical and semantic levels separately.

2.5. Statistical analyses

For all statistical analyses, our dependent measure was the mean magnitude of the total lexico-semantic prediction error produced by each item, averaged across iterations 2 to 11 following stimulus presentation. In all cases, this constituted a 10-iteration time window that surrounded the peak of the error response. Although this window was chosen somewhat arbitrarily by visual inspection, we note that the pattern of effects we describe does not vary with the choice of time window.

To examine the effects of lexical variables, which varied between items, we carried out simple regression analyses. To examine the effects of priming and contextual variables, which varied within items, we carried out linear mixed effects regressions using lme4 package version 1.1–31 (Bates, Mächler, Bolker, & Walker, 2015) in R version 4.2.2 (R Core Team, 2022). In these models, we first attempted to fit the maximal random effects structure. In the case of convergence failures, we simplified the random effects structure following the recommendations of Barr, Levy, Scheepers, and Tily (2013). Statistical significance was assessed using a type-III sums of squares estimation, with *p*-values estimated using the Satterthwaite approximation (Satterthwaite, 1946) using lmerTest version 3.1–3 (Kuznetsova, Brockhoff, & Christensen, 2017).

3. Results

3.1. Time course of the simulated N400

In all simulations, the time course of the simulated N400 (lexico-semantic prediction error) showed a rise-and-fall waveform-like morphology, similar to the empirical N400. After stimulus onset, the magnitude of the lexico-semantic prediction error rose to a peak at around iteration 5 before steadily decreasing to a minimum by iteration 20. This fall in lexico-semantic prediction error occurred when the model had successfully settled on the particular set of conceptual, semantic and lexical state units that correctly encoded the bottom-up orthographic input. At this point, the state units were producing top-down predictions/reconstructions that accurately predicted activity at the level below, thereby suppressing activity produced within these error units (i.e. prediction error).

Although the fall of lexico-semantic prediction error provides evidence for successful lexico-semantic access, we also wanted to make contact with previous word recognition models that operationalized successful lexical access as the selection of one lexical item out of a pool of possibilities. Therefore, in each simulation, we set a threshold of 3.0 on the activity accumulating within the lexical state units, and compared the identity of the first lexical state unit to cross this threshold with the identity of the true input: If these matched, then we took this as evidence that the model had successfully identified the orthographic input (see Supplementary Fig. 5). With the exception of one condition in one simulation,[4] within 20 iterations of stimulus onset, the model correctly identified the input on >99% of trials in all conditions (i.e. inaccurate performance on at most two trials per condition).

3.2. Effects of lexical variables

3.2.1. Effect of orthographic neighborhood size on words

Empirically, words with more orthographic neighbors (e.g. *ball*: *bull*, *call*, *bail*) produce a larger N400 response than words with fewer neighbors (e.g., *kiwi*; Holcomb et al., 2002; Laszlo & Federmeier, 2011).

In our model, orthographic neighborhoods were determined by the pattern of weights (specified by the V and W matrices) that connected the lexical and orthographic units (see Fig. 2 in Methods). For our simulations, we operationalized the *orthographic neighborhood size* (ONsize) of each of our 512 critical lexical items as the number of words in the model's lexicon with which it overlapped in three letter positions. For example, *ball* and *gall* are orthographic neighbors because they share A_2, L_3, and L_4.

Mirroring the empirical findings, we found that lexico-semantic prediction error was larger on words with a larger versus smaller ONsize ($b = 32.63$, $t = 43.82$, $p < .001$, see Fig. 4A).

3.2.2. Effect of orthographic neighborhood size on pseudowords and effect of lexical status

Empirical studies have shown that orthographic neighborhood size not only modulates the amplitude of the N400 produced by real words (WISH), but also by pseudowords (*WUSH, Laszlo & Federmeier, 2011; Holcomb et al., 2002). Moreover, the magnitude of this effect is the same on words and pseudowords (Laszlo & Federmeier, 2011). On the other hand, several studies have reported that pseudowords elicit larger N400s than words (*BAVE > GAVE; Bentin, 1987), even when controlling for orthographic neighborhood size (Holcomb et al., 2002; Meade, Midgley, Dijkstra, & Holcomb, 2018; Braun et al., 2006; although see Laszlo & Federmeier, 2011).

To determine whether our model could explain these effects of orthographic neighborhood size on the N400 produced by pseudowords, and to examine the effect of lexical status (words vs. pseudowords) on lexico-semantic prediction error, we carried out another set of simulations, this time using a new set of 400 words and 400 pseudowords that were designed to have identical neighborhood sizes (see Section 3.6 below for details of how these stimuli were developed). Consistent with the empirical findings, we saw an effect of orthographic neighborhood size on pseudowords (see Fig. 4B), and this effect did not differ between the words and pseudowords (Main effect of ONsize: $b = 24.78$, $t = 27.03$, $p < .001$; no interaction between ONsize and Lexical Status: $b = -0.09$, $t = -0.09$, $p = .93$). Also consistent with many of the empirical findings (Braun et al., 2006; Holcomb et al., 2002; Meade et al., 2018), there was also a main effect of Lexical Status due to a larger lexico-semantic prediction error on pseudowords than words ($b = -33.69$, $t = -36.77$, $p < .001$).

These effects of orthographic neighborhood size on lexico-semantic

[4] For reasons we will return to in the Discussion, the (related) orthographically overlapping inputs in the anticipatory orthographic overlap simulation (see Section 3.5) required significantly more time to cross threshold. Indeed, for nearly half the items (48%), no lexical state unit crossed the threshold within the first 20 iterations after stimulus onset. However, amongst the half that crossed the threshold, 83% matched the target. Moreover, when the model was given more time to process the input (i.e. up to 40 iterations), all items crossed the threshold, and accuracy improved from 83% to 91%.

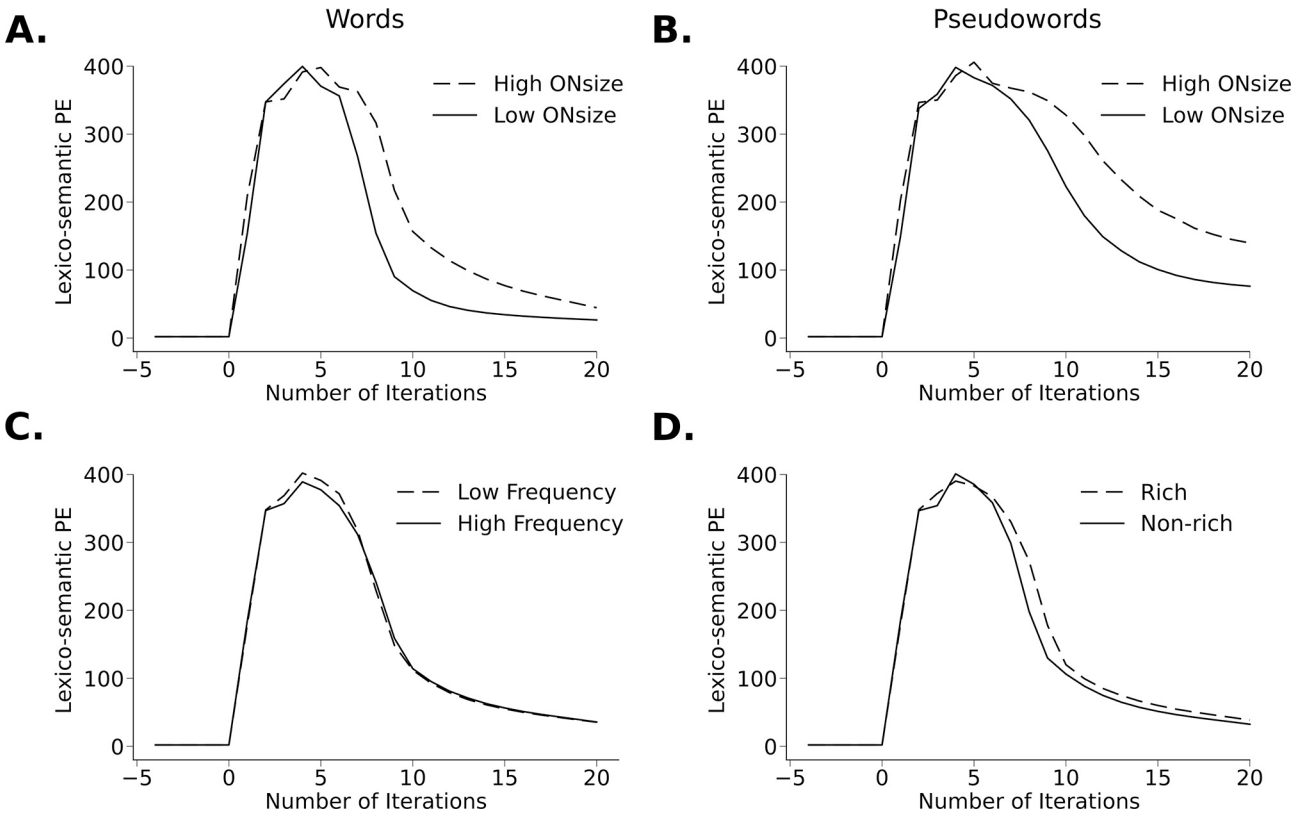

Fig. 4. Effects of lexical variables on the time course of lexico-semantic prediction error. In this and subsequent figures, in each plot, the x-axis shows the number of iterations after stimulus onset, and the y-axis shows the total lexico-semantic prediction error (PE) (arbitrary units), averaged across items within each condition. Because the standard errors are very small and thus barely visible, we opted not to include them in the plots. A. High vs. Low Orthographic Neighborhood size (ONsize), based on a median split across 512 critical words. High ONsize words elicited a significantly larger lexico-semantic prediction error than Low ONsize words. B. High vs. Low ONsize, based on a median split across 400 pseudoword items. High ONsize pseudowords elicited a significantly larger lexico-semantic prediction error than Low ONsize pseudowords. C. High vs. Low Frequency, based on a median split across 512 critical words. Low frequency items elicited a significantly larger lexico-semantic prediction error than high frequency items. D. Rich vs. Non-rich (lexical items connected to 18 vs. 9 semantic features). Rich items elicited a significantly larger lexico-semantic prediction error than Non-rich items.

prediction error arose because when the orthographic inputs (both words and pseudowords) with large neighborhood sizes arrived at the lexical level, they partially activated a large number of closely overlapping lexical state units. This produced a larger lexical prediction error, which, in turn, activated semantic state units of all these neighbors, and therefore a larger semantic prediction error.

In addition to being larger in magnitude, the lexico-semantic prediction error produced by inputs with larger orthographic neighborhood sizes took slightly longer to fall than the prediction error produced by inputs with smaller neighborhood sizes, i.e. a delayed downslope in the simulated N400, see Fig. 4A and B. This slight delay in minimizing lexico-semantic prediction error occurred because it took the state units slightly longer to settle on the correct conceptual and semantic representations. Intriguingly, a visual examination of empirical data for these contrasts appears to show a similar delayed downslope in the N400 in some cases (e.g. Laszlo and Federmeier (2011) Fig. 3), although this has not been systematically examined in the literature.

The reason why pseudowords (*WUSH) produced a larger lexico-semantic prediction error than real words, even when controlling for orthographic neighborhood size, is that, after activating multiple lexical candidates with varying degrees of orthographic overlap (WISH, BUSH, LUSH, etc.), the model was unable to settle on a single lexico-semantic state that could explain the bottom-up input.

3.2.3. *Effect of lexical frequency*

The amplitude of the N400 is smaller to words of higher frequency (e.g., *ball*) than to words of lower frequency (e.g. *gall*; Rugg, 1990; Van Petten & Kutas, 1990; Laszlo & Federmeier, 2014; Hauk, Davis, Ford, Pulvermuller, & Marslen-Wilson, 2006). Within a Bayesian framework, the effects of frequency can be conceptualized as a prior belief during perceptual inference (Delaney-Busch, Morgan, Lau, & Kuperberg, 2019; Norris, 2006; Spratling, 2016a). In order to bias the model's prior beliefs, we incorporate the frequency of each word in the model's top-down generative weights. Specifically, as described under Model Architecture (and see Fig. 2 for a schematic depiction), for each lexical item, we increased the strength of its unique set of feedback connections in proportion to its SUBTLEX-US frequency (Brysbaert & New, 2009). This ensured that, all else being equal, higher frequency items received stronger feedback activity (top-down reconstructions) from higher-level state units as the predictive coding algorithm approximated Bayesian inference. (Note that if we had encoded frequency in the bottom-up feedforward weights, this would not have not biased the model's reconstructions).

As expected, our simulations showed that higher frequency words produced a smaller lexico-semantic prediction error than lower frequency words ($b = -4.05$, $t = -5.43$, $p < .001$; see Fig. 4C). This is because the stronger feedback weights allowed higher levels of the network to generate predictions that better suppressed the production of prediction error. When we looked at each level separately, we found that this suppression appeared to be mostly limited to the semantic level (see Supplementary Fig. 1).

3.2.4. *Effect of semantic richness*

The N400 is generally larger to words with more concrete meanings

(Holcomb et al., 1999; Kounios & Holcomb, 1994; Lee & Federmeier, 2008), more semantic associates (Laszlo & Federmeier, 2011), and a larger number of semantic features (Amsel, 2011; Rabovsky et al., 2012b; but see Kounios et al., 2009).

Following previous models (Cheyette & Plaut, 2017; Rabovsky & McRae, 2014), to capture these effects, we operationalized the semantic richness of each word in the model's lexicon as the number of semantic features with which it was connected. As described under Model Architecture (see Methods), each of our 512 critical lexical items was connected to either 18 features (Rich) or 9 features (Non-rich).

Our simulations show that semantically rich words produced a larger lexico-semantic prediction error than the Non-rich words (Richness: $b = 9.03$, $t = 12.12$, $p < .001$; see Fig. 4D). This follows from the simple fact that prediction error is summed elementwise across all error units. Therefore, if a lexical unit is linked to a larger number of semantic units, its activation will produce a larger total prediction error.

Of note, the scalp distribution of the semantic richness/concreteness effect on the N400 is more frontal than the classic centroparietal N400 effect (Holcomb et al., 1999; Kounios & Holcomb, 1994; Lee & Federmeier, 2008). It has been hypothesized that this is because the effect originates primarily at the level of semantic features, rather than at the lexical level that links these features to linguistic form (e.g. Kounios & Holcomb, 1994). Consistent with this account, we note that when we looked at the effect of richness on prediction error generated the lexical and semantic levels separately, the effect did indeed appear to stem primarily from prediction error produced at the semantic rather than the lexical level (see Supplementary Materials Fig. 1).

3.3. Effects of word-pair priming

In a typical priming paradigm, pairs of "prime" and "target" words are presented sequentially, with the prime being either related or unrelated to the target along some dimension. Empirically, the amplitude of the N400 is smaller to repeated than non-repeated targets in repetition priming paradigms (e.g. *lime – lime* vs. *flow – lime*: Rugg, 1985; Misra & Holcomb, 2003) and smaller to semantically related than unrelated targets in semantic priming paradigms (e.g. *sour – lime* vs. *flow – lime:* Bentin et al., 1985; Rugg, 1985; Holcomb, 1988; Holcomb & Neville, 1990). Additionally, several studies have shown that the repetition priming effect is larger than the semantic priming effect (Deacon, Dynowska, Ritter, & Grose-Fifer, 2004; Rugg, 1985).

To simulate these priming effects, we clamped an orthographic "prime" input for 20 iterations, followed by two blank iterations (all input units clamped to zero), followed by either a related or unrelated "target" input for an additional 20 iterations.

3.3.1. Effect of repetition priming

In the repetition priming simulations, the prime was either identical to the target, or it was unrelated, sharing no semantic features with the target. Consistent with the empirical findings, we found that the repeated targets produced a smaller lexico-semantic prediction error than the non-repeated targets ($b = -125.52$, $t = -140.3$, $p < .001$; see Fig. 5A).

3.3.2. Effect of semantic priming

In the semantic priming simulations, the prime either shared eight semantic features (related condition) or no semantic features with the target (unrelated condition). Again consistent with the empirical findings, we found that the semantically related targets produced a smaller lexico-semantic prediction error than the unrelated targets ($b = -107.25$, $t = -87.15$, $p < .001$; see Fig. 5B).

Finally, consistent with the empirical findings, lexico-semantic prediction error was smaller to repeated than to semantically related targets ($b = -34.67$, $t = -20.46$, $p < .001$).

The reason why the simulated N400 was attenuated to primed (versus unprimed) targets is because the presentation of the prime led the model to fully converge (in the case of repetition priming) or partially converge (in the case of semantic priming) on the lexical and semantic state units that corresponded to the target. This resulted in more accurate top-down reconstructions and therefore a smaller lexico-semantic prediction error to primed versus unprimed targets.

We note that in addition to being smaller in amplitude, the simulated N400 evoked by primed targets peaked earlier than that evoked by unprimed targets (see Fig. 5A and B). Empirically, there is some evidence that repeated targets produce a waveform that diverges earlier from that produced by unrepeated targets, between 200 and 300 ms (Rugg, Doyle, & Melan, 1993; Rugg & Nieto-Vegas, 1999) possibly reflecting a distinct N250 effect (see Grainger & Holcomb, 2009; Holcomb & Grainger, 2006; Kiyonaga, Grainger, Midgley, & Holcomb, 2007). However, in general, the peak latency of the N400 is fairly stable (Federmeier & Laszlo, 2009). One possible reason for this discrepancy is that the hierarchical structure employed in the current predictive coding model was unrealistically shallow, resulting in an artificial proximity between the lexico-semantic states and the bottom-up input. We will return to this point in the Discussion under Limitations and Future Directions.

3.4. Contextual effects

During language comprehension, the amplitude of the N400 is strongly influenced by the broader sentence and discourse context. Predictive models of online language comprehension posit that this is

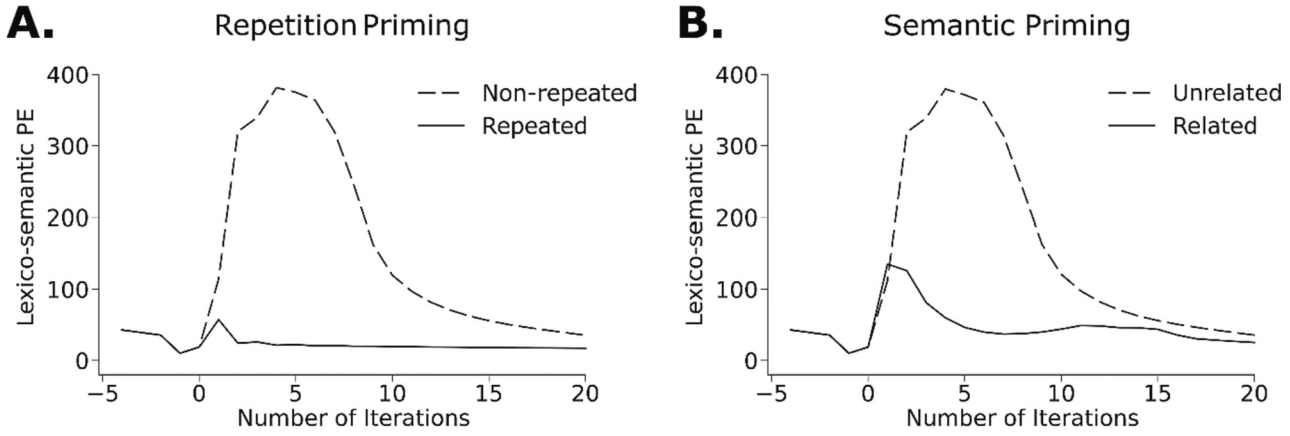

Fig. 5. Effects of word-pair priming on the time course of lexico-semantic prediction error. A. Effect of repetition priming. B. Effect of semantic priming: Unrelated (zero semantic features shared between prime and target) vs. Related (eight semantic features shared between prime and target).

because comprehenders incrementally use the prior context to infer a higher-level interpretation, which they use to generate top-down predictions that pre-activate lower lexical representations (e.g., Federmeier, 2007; Kuperberg & Jaeger, 2016, Section 3.5; Wang, Kuperberg, & Jensen, 2018).

In order to simulate this top-down predictive pre-activation, we clamped the expected state unit at the highest conceptual layer of the model with the desired percentage of pre-activation and allowed this predicted information to flow down the model for 20 iterations. We did this by fixing the total level of activation to a constant value (see Supplementary Materials) and distributing the remaining activation evenly across the remaining units. During this top-down pre-activation phase, the steps of the predictive coding algorithm remained unchanged: the conceptual layer generated semantic reconstructions, which induced a top-down bias in the model's semantic state units. These newly activated semantic states then produced reconstructions that led to the pre-activation of lexical state units, and so on down the network. After the 20 iterations of pre-activation, we unclamped the expected conceptual state units at the top of the network and presented a new bottom-up input at the orthographic level for an additional 20 iterations.

3.4.1. Effects of lexical probability

The amplitude of the N400 is strongly influenced by the lexical *probability* of each word, given its prior context. The probability of a word can be estimated either using the cloze procedure (the proportion of participants who produce that word during a sentence completion task (Taylor, 1953), or using large language models (e.g., GPT-3, Brown et al., 2020). Several studies have shown that the amplitude of the N400 is inversely proportional to these estimates of lexical probability in context (studies using cloze estimates: Kutas & Hillyard, 1984; DeLong et al., 2005; Wlotko & Federmeier, 2012; Brothers, Morgan, Yacovone, & Kuperberg, 2023; studies using estimates from large language models: Michaelov, Coulson, & Bergen; Szewczyk & Federmeier, 2022; Heilbron, Armeni, Schoffelen, Hagoort, & de Lange, 2022).

To simulate the effect of lexical probability on the N400, we presented each of our 512 critical words at four different levels of probability: 99%, 50%, 25% and uniform (1/[total words] = 1/1579 = 0.06%). In the linear mixed effects analyses, probability was standardized and served as a within-item predictor. As expected, we observed a graded reduction in lexico-semantic prediction error as lexical probability increased ($b = -87.42$, $t = -54.02$, $p < .001$; see Fig. 6A). This is because, with increasing predictability, the model was able to settle on an increasingly more accurate set of semantic and lexical states prior to the appearance of the target word. After stimulus onset, this resulted in more accurate top-down reconstructions and greater suppression of lexico-semantic prediction error.[5]

We note that, similar to the priming simulations, the smaller the simulated N400s, the earlier their peak latencies (see Fig. 6A). Again, this contrasts with the empirical N400: Although there is some evidence that the N400 peaks slightly earlier to highly predictable words in sentence contexts (Brothers, Swaab, & Traxler, 2015), and slightly later to more implausible words (Brothers et al., 2015; Nieuwland et al., 2020), for the most part, the peak of the N400 predictability effect is stable (Federmeier & Laszlo, 2009). As noted above in relation to priming, we think that this discrepancy may, in part, be due to the current model's relatively shallow hierarchical structure (see Discussion, Limitations and Future Directions).

3.4.2. No effect of Constraint on unpredicted words

The amplitude of the N400 is not sensitive to the lexical constraint of the prior context, when controlling for lexical probability. For example, the N400 response is equally large to unexpected but plausible words that violate strong lexical predictions in a high constraint contexts (e.g. "*Every morning he took his dog for a swim*", where the word "*walk*" was expected) and to unexpected but plausible words in low constraint contexts (e.g. "*Helen reached up to dust the dresser*") (Federmeier, Wlotko, De Ochoa-Dewald, & Kutas, 2007; Kuperberg et al., 2020; Kutas & Hillyard, 1984).

To simulate the null effect of Constraint (over and above predictability), we presented unexpected critical words in either a *high constraint* condition, in which we strongly pre-activated a different randomly selected word (99%), or in a *low constraint* condition in which all words were given uniform pre-activation (0.06%). In the linear mixed effects analysis, with Constraint serving as a categorical within-items predictor, we found that there was no difference in the magnitude of lexico-semantic prediction error produced by the *high constraint unexpected* and the *low constraint unexpected* inputs (Constraint: $b = 0.23$, $t = 0.85$, $p = .40$; see Fig. 6B). This is because lexico-semantic prediction error is only sensitive to the residual lexico-semantic information encoded within the bottom-up input that is *not* predicted by the level above. The amount of residual lexico-semantic information encoded within an unexpected input is the same, regardless of whether prior incorrect predictions were concentrated over one specific set of semantic features/lexical candidate, or whether they were spread diffusely over multiple semantic features/lexical candidates.

3.4.3. Effect of anticipatory semantic overlap

In addition to its sensitivity to lexical probability, the amplitude of the N400 is also sensitive to the semantic relationship between a predicted word and the observed bottom-up input (Federmeier & Kutas, 1999; Kutas & Hillyard, 1984). For example, Federmeier and Kutas (1999) presented participants with highly constraining contexts, e.g. "*They wanted to make the hotel look more like a tropical resort. So along the driveway, they planted rows of…*", followed by critical words that were highly predictable (e.g., *palms*), lexically unpredictable (<1% probability) but with semantic features that overlapped with the expected continuation (e.g. *pines*), or lexically unpredictable (<1% probability) but with fewer overlapping expected features (e.g. *tulips*). They observed a graded reduction in the N400 response across the three conditions (*palms* < *pines* < *tulips*). Although this *anticipatory semantic overlap effect* on the N400 was originally described on unexpected implausible targets (Kutas & Hillyard, 1984; Federmeier & Kutas, 1999; for more recent replications, see DeLong, Chan, & Kutas, 2019; Ito, Corley, Pickering, Martin, & Nieuwland, 2016), it is also seen on *plausible* continuations (both zero-cloze: Thornhill & Van Petten, 2012; DeLong & Kutas, 2020, and non-zero cloze: Brothers et al., 2023).

To simulate this effect, we pre-activated the model with each of our 512 words, assigning the conceptual states a probability of 99% and clamping them for 20 iterations. We then presented the model with (a) the same word that was pre-activated (*expected*), (b) a different word that shared eight semantic features with the expected word (*unexpected semantically overlapping*), or (c) a different word that shared no semantic features with the expected word (*unexpected unrelated)*. We also ensured that the *unexpected semantically overlapping* and the *unexpected unrelated* words had minimal orthographic overlap with the *expected* words (0.38 and 0.37 characters respectively) and that the extent of this overlap did not differ between the two conditions ($t < 1$, $p = .73$).

As shown in Fig. 7A, consistent with the empirical findings, we saw a graded reduction of lexico-semantic prediction error across the three conditions, with the *unexpected unrelated* words producing a

[5] While it is well established that behavioral measures of processing difficulty and N400 amplitude decrease with increasing lexical probability, there has been some debate about whether this linking function is linear (Brothers & Kuperberg, 2021), logarithmic (Smith & Levy, 2013) or both, depending on the range of probabilities examined (Szewczyk & Federmeier, 2022). The predictive coding model used in these simulations is not well suited for addressing this question because we implemented top-down predictability manually. In addition, on informal inspection, the relationship between lexical probability and lexico-semantic prediction error (and indeed state activity) varied non-systematically with different hyperparameter choices.

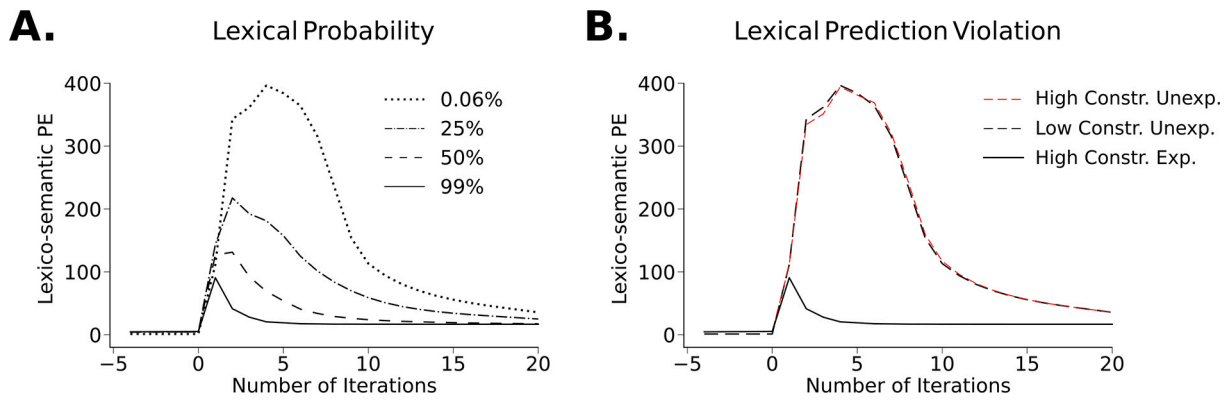

Fig. 6. Effects of Lexical Probability and Constraint on the time course of lexico-semantic prediction error. A. Effect of lexical probability: Lexico-semantic prediction error decreased with increasing lexical probability. B. Effect of Constraint. Lexico-semantic prediction error was equally large to *high constraint unexpected* (High-Constr. Unexp.) and *low constraint unexpected* (Low Constr. Unexp.) inputs, relative to the *expected* inputs (High Constr. Exp).

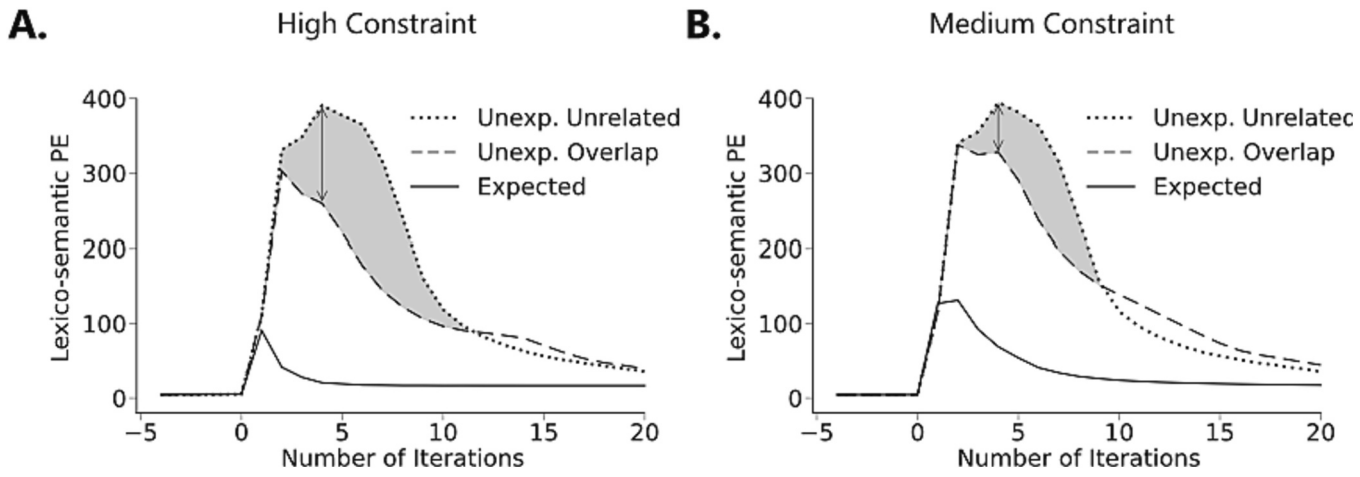

Fig. 7. Effects of anticipatory semantic overlap on the time course of lexico-semantic prediction error. A. In the *high constraint* condition (in which the model was pre-activated with 99% probability), lexico-semantic prediction error was largest to the *unexpected unrelated* words (Unexp. Unrelated), smaller to the *unexpected semantically overlapping* words (Unexp. Overlap) and smallest to the *expected* words. B. In the *medium constraint* condition (in which the model was pre-activated with 50% probability), lexico-semantic prediction error also decreased across the three conditions. However, as indicated using arrows/shading, in this *medium constraint* condition, the *difference* in prediction error produced by the *unexpected unrelated* and the *unexpected semantically overlapping* words was smaller than this difference in the *high constraint* condition.

significantly larger prediction error than the *unexpected semantically overlapping* words ($b = 158.07$; $t = 61.42$, $p < .001$), which, in turn, produced a significantly larger prediction error than the *expected* words ($b = 95.27$; $t = 37.02$, $p < .001$).

The reason why lexico-semantic prediction error to the *unexpected semantically overlapping* words was attenuated is because the pre-activated expected conceptual representations produced re-constructions that pre-activated shared semantic features, which, in turn, produced reconstructions that suppressed both lexical and semantic prediction error.

3.4.4. *Effect of contextual constraint on the anticipatory semantic overlap effect*

In addition to showing an effect of anticipatory semantic overlap on the N400 produced by unexpected words in high constraint contexts, Federmeier and Kutas (1999) also showed that this effect interacted with contextual constraint (see also Ito et al., 2016 for a recent replication using a different set of materials); that is, the degree to which the N400 was reduced in response to the *unexpected semantically overlapping* words (*pines*), relative to the *unexpected unrelated* words (*tulips*), was *greater* in the *high constraint* contexts than in the *medium constraint* contexts.

To simulate this interaction between Semantic Overlap and Constraint, we carried out the same simulations as described above, except that instead of assigning the conceptual states a probability of 99%, we assigned them a probability of 50%. We then compared the magnitude of prediction error produced by the *unexpected unrelated* and the *unexpected semantically overlapping* inputs across the *high constraint* and *medium constraint* conditions. As shown in Fig. 7B, there appeared to be a smaller difference in lexico-semantic prediction error between the *unexpected unrelated* and *unexpected semantically overlapping* words in the *medium constraint* condition (17% reduction) than in the *high constraint* condition (35% reduction). This was confirmed by a linear mixed-effects model that crossed Semantic Overlap (*semantically overlapping*, *unrelated*) with Constraint (*high constraint*, *medium constraint*), which showed that, in addition to a main effect of Semantic Overlap ($b = -45.45$, $t = -22.83$, $p < .001$) and no main effect of Constraint ($p = .48$), there was a significant interaction between Semantic Overlap and Constraint ($b = -49.82$, $t = -17.69$, $p < .001$). This is because the stronger pre-activation of conceptual and semantic features in the *high constraint* (99%) than in the *medium constraint* (50%) condition, led them produce more accurate reconstructions that resulted in a greater suppression of lexical and semantic prediction error to the semantically overlapping unexpected inputs.

3.4.5. Effect of anticipatory orthographic overlap on words

The N400 is also sensitive to the *orthographic* overlap between a predicted and encountered input (DeLong et al., 2019; Ito et al., 2016; Laszlo & Federmeier, 2009). For example, Laszlo and Federmeier (2009) presented participants with highly constraining sentence contexts, e.g. "*The genie granted his third and final…*", followed by critical words that were *expected* (e.g., *wish*), unexpected (<1% probability) but with several letters in common with the expected word, i.e. *unexpected orthographically overlapping* (e.g. *dish*), or *unexpected unrelated* (<1% probability) with no letters in common with the expected word, but matched to the two other conditions on orthographic neighborhood size (e.g. *claw*). The authors observed a graded reduction of the N400 across the three conditions (*wish* < *dish* < *claw*).

To simulate this *anticipatory orthographic overlap* effect, we used 400 words from a set of stimuli that we developed for the purpose of these simulations (see Section 3.6 for details). We pre-activated the model with each of these 400 words, assigning the conceptual states a probability of 99% and clamping them for 20 iterations. We then presented the model with (a) the same word that was pre-activated (*expected*), (b) a different word that shared three letters (but no semantic features) with the expected word (*unexpected orthographically overlapping*), or (c) a different word that had minimal orthographic overlap (and no semantic overlap) with the expected word (*unexpected unrelated*).

Consistent with the empirical findings, we saw a graded reduction of lexico-semantic prediction error across the three conditions over the period we used to simulate the N400, with the *unexpected unrelated* words producing a significantly larger prediction error than the *unexpected orthographically overlapping* words ($b = 134.70$; $t = 115.47$, $p < .001$), which, in turn, produced a significantly larger prediction error than the *expected* words ($b = 119.65$; $t = 97.77$, $p < .001$).

The reason why the *unexpected orthographically overlapping* words (DISH) and non-words (*WUSH) produced a smaller lexico-semantic prediction error than the *unexpected unrelated* words is because pre-activating the model with the expected word (*wish*) resulted in the generation of orthographic reconstructions (W-I-S-H) that partially suppressed the orthographic prediction error produced by the bottom-up input (D-**I**-**S**-**H**). Therefore, less orthographic prediction error flowed up the hierarchy, inducing only small updates in state units, and therefore smaller prediction errors at the lexical and semantic levels throughout the time period we used to operationalize the N400.

We note that, although smaller in amplitude, the simulated N400 produced by *unexpected orthographically overlapping* words had a more prolonged time course than that produced by the *unexpected unrelated* inputs, see Fig. 8A. We will return to the reasons for this in the Discussion section. In the empirical literature, this type of prolongation of the N400 to *unexpected orthographically overlapping* words has not been described. However, this may be because following the N400 time window, *unexpected orthographically overlapping* words produce a posteriorly distributed positive-going ERP component (the P600) (DeLong et al., 2019; Ito et al., 2016; Laszlo & Federmeier, 2009; Vissers, Chwilla, & Kolk, 2006), which would have masked any prolongation of the N400 at the scalp surface (i.e. component overlap, see Kuperberg, Kreher, Sitnikova, Caplan, & Holcomb, 2007; Brouwer & Crocker, 2017). One way of exploring this possibility would be to carry out the same study using MEG where this type of spatiotemporal component overlap is less of an issue (for detailed discussion, see Wang & Kuperberg, 2023, Study 2, Discussion).

3.4.6. Effect of anticipatory orthographic overlap on pseudowords

In addition to showing an anticipatory orthographic overlap effect on real words, Laszlo and Federmeier (2009) demonstrated the same effect on pseudowords; that is, the N400 was smaller in response to pseudowords that shared letters with an expected word (e.g. *WUSH) than to unrelated pseudowords that were matched on orthographic neighborhood size (e.g. *CLAF). To demonstrate this effect, the authors developed a set of stimuli that allowed them to cross Orthographic Overlap (*orthographically overlapping* vs. *unrelated*) and Lexical Status (*word*, *pseudoword*), and match all four conditions on orthographic neighborhood size.

For our simulations, following Laszlo and Federmeier (2009), we developed a set of 400 real word and 400 matched pseudoword stimuli from our 1579-word lexicon. We began with a set of 400 "base-words", and, for each of these base words, we constructed a quadruplet of items, thereby setting up a 2 × 2 design that crossed Orthographic Overlap of the critical item with the expected base word (*orthographically overlapping*, *unrelated*) and Lexical Status (*word*, *pseudoword*), while matching items across the four conditions on orthographic neighborhood size. First, for each base-word (WISH), we selected a real word (DISH) and a pseudoword (*WUSH) that overlapped with it in three letter positions, and that was matched to it on orthographic neighborhood size (mean ONsize across conditions: 6.71, SD: 3.58). Second, to ensure that each item appeared in both the *orthographically overlapping* and the *unrelated* conditions, we symmetrically paired each base-word (WISH) with another base-word (CLAW) that was of the same orthographic neighborhood size, with the restriction that neither base-word overlapped at any letter position with the other's word or pseudoword orthographic neighbor (DISH/CLAW or *WUSH/*CLAF). This yielded 400 items in each of the four conditions: (a) *orthographically overlapping word* (e.g. DISH), (b) *unrelated word* (e.g. CLAW), (c) *orthographically overlapping pseudoword* (e.g. *WUSH), and (d) *unrelated pseudoword* (e.g. *CLAF).

After first pre-activating the model with the base word (WISH, 99% probability) for 20 iterations, we then presented the model with each of the four possible continuations (DISH, *WUSH, CLAW, *CLAF) in each quadruplet, each for 20 iterations. As shown in Fig. 8B, just as for the real word stimuli, *unexpected orthographically overlapping* pseudowords (*WUSH) produced a lexico-semantic prediction error that was smaller than that produced by the *unexpected unrelated* pseudowords (*CLAF) but larger than that produced by the *expected* words (WISH).

A linear mixed-effects model that crossed Orthographic Overlap (*orthographically overlapping*, *unrelated*) and Lexical Status (*word*, *pseudoword*) confirmed a main effect of Orthographic Overlap ($b = -87.75$, $t = -119.58$, $p < .001$). Indeed, the effect on pseudowords was even larger than that on words (Lexical Status x Orthographic Overlap: $b = 20.39$, $t = 27.80$, $p < .001$; note that Laszlo & Federmeier, 2009 found that the effect of Orthographic Overlap was the same on words and pseudowords). Similar to the effect on real words, the simulated N400 produced by *unexpected orthographically overlapping* pseudowords also had a more prolonged time course than that produced by the *unexpected unrelated* pseudowords, see Fig. 8B and Discussion section.

Finally, consistent with the simulations reported in Section 1.2, we saw a main effect of Lexical Status (*pseudowords* > *words*, $b = -13.42$, $t = 18.29$, $p < .001$) because the model was unable to settle on a single lexico-semantic state that could explain the bottom-up pseudoword input, resulting in a larger lexico-semantic prediction error overall.

3.5. Interactions between lexical variables and (a) repetition priming and (b) lexical probability

In Section 1, we described simulations of the effects of several lexical variables on the N400 produced by words presented in isolation, and in Sections 2 and 3, we described simulations of priming and contextual effects, respectively. There is also empirical evidence that the effects of lexical variables on the N400 can be modulated by both repetition priming and contextual predictability. For example, in priming paradigms, the effects of frequency and semantic richness/concreteness are reduced on repeated relative to non-repeated targets (Repetition x Frequency: Rugg, 1990; Repetition x Richness/Concreteness: Rabovsky, Sommer, & Abdel Rahman, 2012a; Kounios & Holcomb, 1994). Similarly, during sentence comprehension, the effects of frequency and semantic richness/concreteness are smaller on predictable words than on unpredictable words (Probability x Frequency: Dambacher, Kliegl, Hofmann, & Jacobs, 2006; Probability x Concreteness: Holcomb et al.,

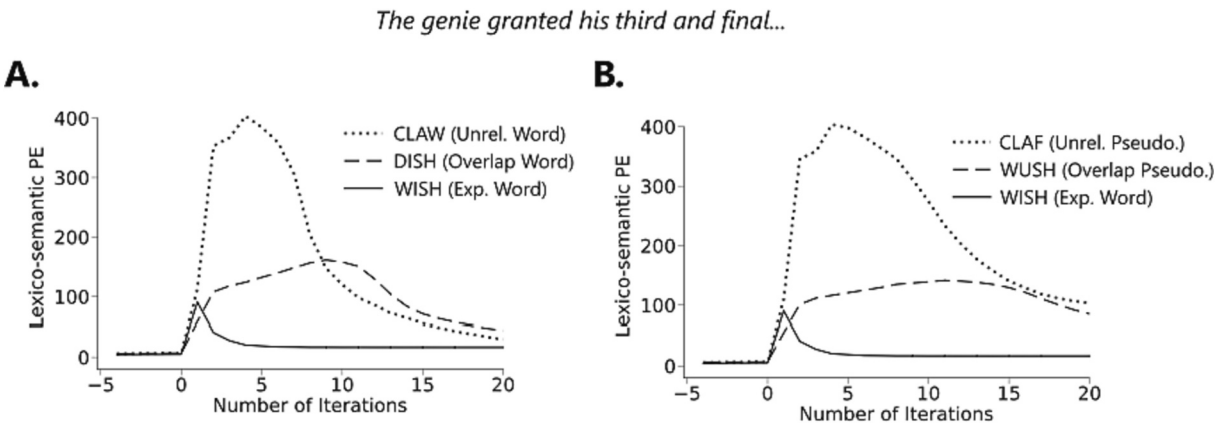

Fig. 8. Effect of anticipatory orthographic overlap on the time course of lexico-semantic prediction error. A. Effect of anticipatory orthographic overlap on words. Lexico-semantic prediction error was largest to the *Unexpected unrelated* words (CLAW, Unrel. Word), smaller to the *unexpected orthographically overlapping* words (DISH, Overlap. Word) and smallest to *expected* words (WISH, Exp. Word). B. Effect of anticipatory orthographic overlap on pseudowords. Lexico-semantic prediction error was largest to the *Unexpected unrelated* pseudowords (*CLAF, Unrel. Pseudo.), smaller to the *unexpected orthographically overlapping* pseudowords (*WUSH, Overlap Pseudo.) and smallest to *expected* words (WISH, Exp. Word).

1999).

Currently, it is less clear how contextual factors influence the magnitude of the orthographic neighborhood effect on the N400. In word lists, Laszlo and Federmeier (2011) described a slightly smaller effect of orthographic neighborhood size on repeated than non-repeated words, although no statistical tests were reported. However, during sentence comprehension, two ERP studies (Payne & Federmeier, 2018; Payne, Lee, & Federmeier, 2015) reported that the effect of orthographic neighborhood size was the same on words appearing early in sentences (less predictable) as on words appearing later in sentences (more predictable). This suggests that, in contrast to other lexical variables, the effect of orthographic neighborhood size may not be overridden by the effect of contextual predictability on the N400. However, no previous study has orthogonally manipulated lexical probability and orthographic neighborhood size on the same words in an experimental design.

Here, we simulated interactions between each of our lexical variables — ONsize, Frequency and Richness — with both repetition priming and lexical probability, by re-analyzing data from the simulations described above.

3.5.1. *Interactions between lexical variables and repetition priming*

In Fig. 9 (top row), we show the effects of ONsize, Frequency and Richness on the magnitude of lexico-semantic prediction error produced by *repeated* versus *non-repeated* target words in our repetition priming simulations. In all three cases, the effects of each of these lexical variables appeared to be smaller on the *repeated* than the *non-repeated* targets. A linear mixed effects model that included both main effects (Repetition, ONsize, Frequency and Richness) as well as three interaction terms (Repetition x ONsize, Repetition x Frequency, and Repetition x Richness) confirmed main effects of Repetition ($b = -125.52$, $t = -232.81$, $p < .001$) as well as each of the three lexical variables (ONsize: $b = 15.37$, $t = 28.47$, $p < .001$; Frequency: $b = -1.62$, $t = -3.00$, $p = .003$; Richness: $b = 6.46$, $t = 11.95$, $p < .001$). Critically, it also revealed interactions between Repetition and all three lexical variables (Repetition x ONsize: $b = -15.13$, $t = -28.02$, $p < .001$; Repetition x Frequency: $b = 1.29$, $t = 2.39$, $p = .017$; Repetition x Richness: $b = -3.22$, $t = -5.96$, $p < .001$), which, in all cases, were driven by smaller effects on repeated than non-repeated words (ONsize: 99% reduction, Frequency: 89% reduction, Richness: 67% reduction).

3.5.2. *Interactions between lexical variables and lexical probability*

In Fig. 9 (bottom row), we show the effects of the same three lexical variables on the magnitude of lexico-semantic prediction error produced by highly expected (99%) and unexpected (0.06%) words. The effects of each of these lexical variables appeared to be smaller on the *expected* than the *unexpected* critical words. Again, this was confirmed by a linear mixed effects model that included all four main effects (Probability, ONsize, Frequency, and Richness) and three interaction terms (Probability x ONsize, Probability x Frequency, and Probability x Richness). In addition to confirming main effects of Probability ($b = -94.35$, $t = -352.47$, $p < .001$) and each of the three lexical variables (ONsize: $b = 18.79$, $t = 49.58$, $p < .001$; Frequency: $b = -2.73$, $t = -7.19$, $p < .001$; Richness: $b = 5.25$, $t = 13.84$, $p < .001$), this analysis revealed interactions between Probability and each lexical variable (Probability x ONsize: $b = -11.53$, $t = -43.00$, $p < .001$; Probability x Frequency: $b = 1.08$, $t = 4.03$, $p < .001$; Probability x Richness: $b = -3.11$, $t = -11.58$, $p < .001$). In all cases, these interactions arose because the effect of each lexical variable was smaller when the word was predictable than when it was unpredictable (ONsize: 96% reduction, Frequency: 73% reduction, Richness: 94% reduction).

These under-additive interactions receive a straightforward interpretation within our predictive coding framework. Generally, an input that is repeated or expected will produce minimal prediction error, limiting the influence its lexical characteristics can have relative to the non-repeated or unexpected conditions, giving rise to an interaction.

Just as for words presented in isolation (see Section 1), in the non-repeated and unexpected conditions, higher frequency words produced a smaller lexico-semantic prediction error than lower frequency words because the stronger feedback weights allowed for the generation of more accurate predictions. In the case of repeated or highly predictable targets, however, this default prior was overridden, and so the magnitude of lexico-semantic prediction error was small, regardless of frequency. Similarly, the interactions with Richness arose because the additional unpredicted lexico-semantic information (i.e., lexico-semantic prediction error) carried by words with many (versus few) semantic features was reduced when these representations were pre-activated. Finally, the effect of orthographic neighborhood size was reduced for both repeated and expected words. This is because in these conditions, top-down reconstructions suppressed the prediction error produced by co-activated lexical state units and their associated semantic features.

4. Discussion

Language comprehension can be understood as probabilistic inference — the process of inferring internal states that underlie observed linguistic inputs (Chater, Crocker, & Pickering, 1998; Kuperberg & Jaeger, 2016; Levy, 2008; Narayanan & Jurafsky, 2002; Norris, 2006).

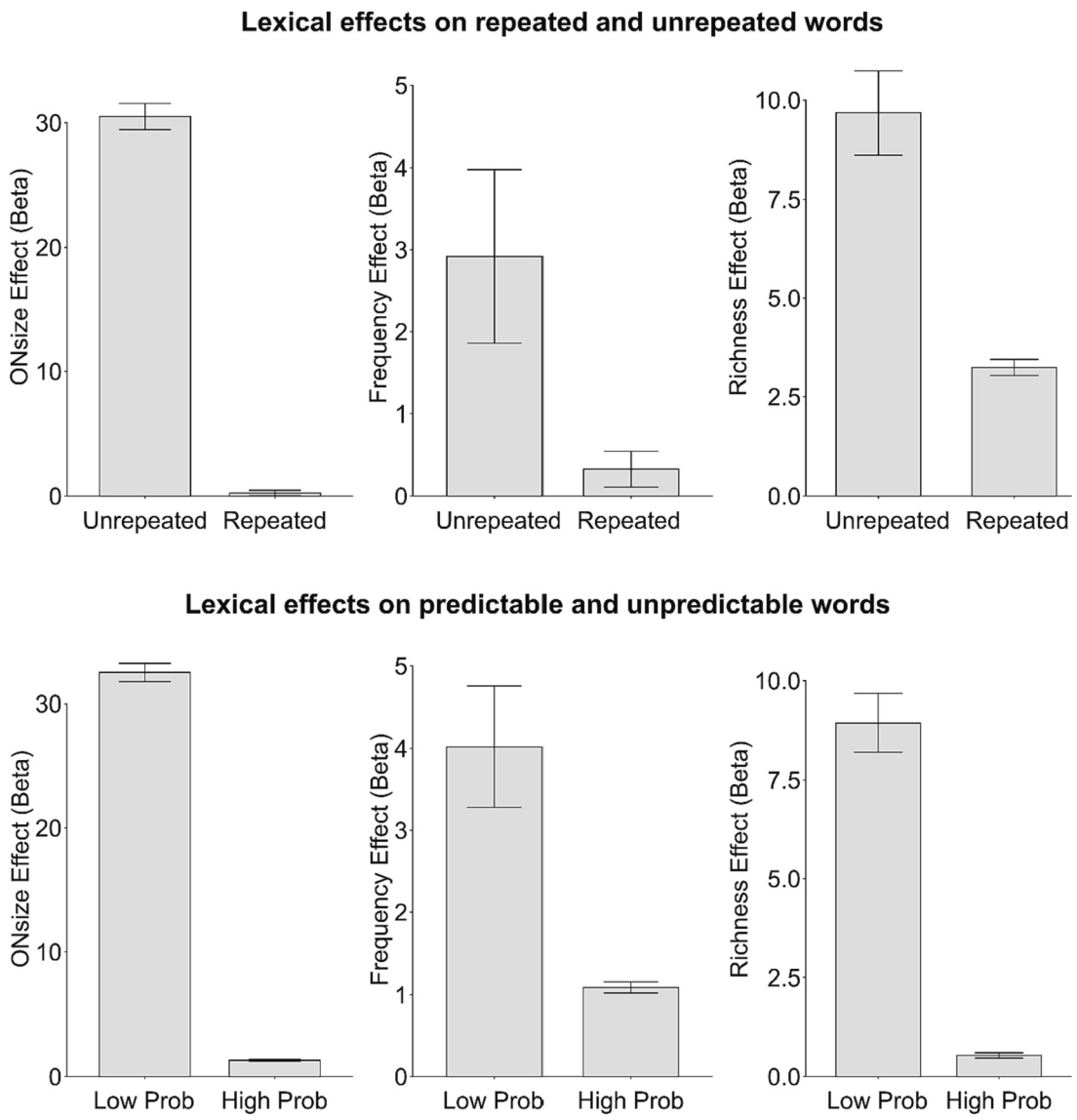

Fig. 9. Interaction between lexical variables and repetition priming (top row), and lexical probability (bottom row). In all bar charts, the y-axis shows the average estimate of the slope (i.e., the beta value) obtained by regressing ONsize, Frequency and Richness on the lexico-semantic prediction error. Error bars indicate ±1 standard error of the mean. The effects of all three lexical variables on the magnitude of lexico-semantic prediction error were reduced in the repeated (vs. non-repeated) conditions, and in the high (vs. low) probability conditions. The full time courses of all effects on the simulated N400 are shown in Supplementary Materials Fig. 4.

Predictive coding refers to a particular architecture and optimization algorithm that approximates this inferential process by actively generating top-down predictions and passing up unpredicted information — prediction error — to update internal states until they explain the bottom-up input and prediction error is minimized. Here, we show, for the first time, that the N400 ERP component can be simulated as the total lexical and semantic prediction error produced as predictive coding infers the meaning of incoming words from their orthographic forms. Using a predictive coding model with an architecture and algorithm that was initially developed to simulate a variety of phenomena in vision (Rao & Ballard, 1999; Spratling, 2013; Spratling, 2014), we were able to reproduce a wide range of lexical, priming and contextual effects in the N400 literature. We further show that predictive coding provides a biologically plausible link to neural activity, and a natural explanation for the temporal dynamics of the N400. Our findings therefore raise the possibly that predictive coding is used to implement language comprehension in the brain, and that the production of lexico-semantic prediction error (the N400) plays a key role in this process.

4.1. Predictive coding provides a plausible, functional link to evoked activity and can explain the rise-and-fall morphology of the N400

The central functional role that prediction error plays in inference, as well as its natural biologically plausible link with neural activity distinguishes predictive coding from previous models that have also simulated the N400 as a difference value. Specifically, two previous models have also operationalized the N400 as a "prediction error" — the difference between the current semantic state and an ideal target vector (Rabovsky & McRae, 2014), or between a word input vector and the model's prior lexical prediction (Fitz & Chang, 2019). However, in both these models, the error was calculated outside the model's architecture, and so there was no direct link between these difference values and ongoing neural activity. Two other models operationalized the N400 as an implicit "change of state" induced by an input — the difference in activity between the state of the model from before until after the input is encountered (Brouwer et al., 2017; Rabovsky et al., 2018). However, it is unclear why a larger implicit change in state would produce a larger

evoked response. Indeed, to simulate the N400, the difference between the two states was calculated externally by the modeler, rather than being produced by the model itself. In addition, this difference value played no functional role in language comprehension (although it was posited to play a role in learning Rabovsky et al., 2018, see Future Directions).

In predictive coding, the causal role that prediction error plays in comprehension is transparent: At each level of representation, prediction error serves as an signal for updating states at the level above. In addition, its biological linking function is clear and intuitive: Prediction error is computed and stored within dedicated error units whose individual activity sum to produce the N400. On the assumption that the activation of error units simulates the increased firing of error-computing units in the brain, an increase in lexico-semantic prediction error should produce more post-synaptic activity and a larger evoked N400 response at the scalp surface.

Together, the functional role of prediction error in inference and its direct link to neural activity allowed us to simulate the rise-and-fall time course of the N400. Only one previous model has successfully simulated the morphology of the N400 (Cheyette & Plaut, 2017; Laszlo & Armstrong, 2014; Laszlo & Plaut, 2012). In that model, instead of being operationalized as a difference value, the N400 was operationalized simply as the total activity produced by a semantic layer. By incorporating several neurobiological constraints into the model's architecture, the authors were able to show that this semantic activity exhibited similar dynamics to the actual N400.[6] In predictive coding, the time course of the simulated N400 emerges simply as a consequence of prediction error serving as an update signal in the optimization algorithm (see also Friston, 2005 for more general discussion). Specifically, when new unpredicted orthographic input is presented, the model's states fail to produce accurate predictions/reconstructions, resulting in a rise in lexico-semantic prediction error, explaining the rise in N400 amplitude. Then, as this error is used to update these higher-level states, they converge to stable values, producing more accurate top-down predictions that suppress lexico-semantic prediction error, resulting in the fall in N400 amplitude.

4.2. A single measure — lexico-semantic prediction error — can capture lexical, priming and broader effects of context on the N400

The multiplicity of factors that can affect the amplitude of the N400 has raised questions about whether this ERP component can be understood as a univariate error signal (Federmeier, 2022). Our model tackles this challenge directly. We demonstrate that lexical, priming *and* higher-level contextual effects on the N400, as well as their interactions, can be captured by a single dependent measure: *lexico-semantic prediction error*.[7]

4.2.1. Simulations of word-level phenomena and priming

Similar to previous word-level models, predictive coding was able to explain the sensitivity of the N400 to various lexical characteristics when words were presented in isolation, including the effects of frequency, semantic richness and orthographic neighborhood size. Also similar to previous word-level and some sentence-level models, our model was able to explain why the amplitude of the N400 is attenuated when words are repeated or semantically primed (see Table 1).

Notably, our model was able to simulate effects on the N400 not only on real words, but also on pseudowords, which thus far have only been successfully simulated by Laszlo and Plaut (2012). Similar to Laszlo and Plaut (2012), and mirroring the empirical data (Heilbron et al., 2022; Laszlo & Federmeier, 2011), our simulations showed that the effect of orthographic neighborhood size on pseudoword letter strings (*W-*E*-E-N) is similar to that on words. This is because, as originally discussed by Laszlo and Federmeier (2011), the co-activation of orthographic neighbors will result in the activation of their associated semantic features, regardless of whether the eliciting stimulus is a word or a non-word. In our model, co-activating a larger group of neighbors resulted in a greater lexico-semantic prediction error.

In addition, our model was able to simulate, for the first time, the larger N400 produced by pseudowords than by words (cf. Bentin, 1987), even when orthographic neighborhood size is held constant (for empirical findings, see Heilbron et al., 2022; Meade et al., 2018; Braun et al., 2006; but see Laszlo & Federmeier, 2011). In our model, the reason for this is that pseudoword string inputs (e.g., *W-E-E-N) activated multiple lexical units in parallel (e.g. *weed*, *teen*, *wren* etc.), and, unlike for real word inputs, the model did not converge on a single representation, preventing it from fully suppressing the lexico-semantic prediction error.

4.2.2. Simulations of broader contextual effects through top-down prediction

Similar to previous sentence-level models, our model was able to simulate several effects of broader context on the N400. These included the graded effects of lexical predictability (Kutas & Hillyard, 1984; DeLong et al., 2005; simulated by Rabovsky et al., 2018; Fitz & Chang, 2019; see also Brouwer et al., 2017), and the null effect of contextual constraint over and above lexical predictability (Kutas & Hillyard, 1984; Federmeier et al., 2007; simulated by Rabovsky et al., 2018 and by Fitz & Chang, 2019). We also simulated the anticipatory semantic overlap effect (Federmeier & Kutas, 1999; Kutas & Hillyard, 1984), which has previously only been simulated by Rabovsky et al., 2018). Finally, we also simulated contextual effects that have not yet been reported in previous models, including the effect of constraint on the anticipatory semantic overlap effect (empirical findings: Federmeier & Kutas, 1999; Ito et al., 2016), and, most notably, the effect of anticipatory orthographic overlap on both words (empirical findings: Laszlo & Federmeier, 2009; Ito et al., 2016; DeLong et al., 2019) and pseudowords (empirical findings: Laszlo & Federmeier, 2009).

Of note, the approach that we took to model these broader effects of context was different from that taken by previous models of the N400. In these previous models, a higher-level representation of the context was computed from the word sequence encountered thus far. In Fitz and Chang (2019)'s model, this contextual representation enabled next-word prediction, and the attenuation of the N400 to expected inputs was simulated as a reduced lexical prediction error. However, as noted earlier, this error was computed outside the model itself. In the change-of-state models by Rabovsky et al., 2018 and Brouwer et al., 2017, this contextual representation implicitly encoded an *event*, which carried information about upcoming expected semantic features. Rabovsky et al. (2018) showed that, as a consequence of these implicit semantic predictions, the shift in state at the event-level layer (the simulated N400 effect) was smaller to lexically expected than unexpected inputs. The authors further showed that these implicit semantic predictions could explain why the N400 was also smaller to lexically unexpected but

[6] We note, however, that not all architectural assumptions of this model were biologically motivated (see Nour Eddine et al., 2022, pages 138–140 for discussion). We also note that the use of total semantic activity (instead of a difference value) to simulate the N400 led to certain challenges in modeling priming effects. Specifically, if the N400 simply reflected total semantic activity, then primed words should generate a *larger* N400 response than unprimed words. To solve this problem, the authors implemented a decay-driven inhibition mechanism. However, while this allowed them to successfully simulate the priming effect as a smaller amplitude N400 (Bentin et al., 1985; Rugg, 1985), instead of facilitating the processing of primed targets, this inhibition essentially interfered with access to their semantic features (see Nour Eddine et al., 2022, page 141 for discussion).

[7] As we discuss below under Future Directions, prediction error generated at a higher event level of representation may additionally contribute to the larger N400 generated by highly implausible/anomalous words, relative to unpredictable but plausible words (see Wang et al., 2023).

semantically overlapping inputs than to unexpected unrelated inputs, i.e. the anticipatory semantic overlap effect. Finally, in Brouwer et al.'s model (Brouwer & Crocker, 2017), the N400 was simulated as a shift within a lower "semantic retrieval" layer that received input from the higher event-level layer. However, note that even when the input was expected, it induced a relatively large shift in this layer (see Nour Eddine et al., 2022, page 150, for discussion).

Several qualitative psycholinguistic theories of predictive language comprehension, however, have proposed that prediction goes beyond implicitly anticipating upcoming semantic features. These frameworks posit that implicitly predicted semantic information is actively propagated down the linguistic hierarchy in a top-down fashion (see Kuperberg & Jaeger, 2016, Section 3, pp. 39–45). Thus, according to these theories, the attenuation of the N400 is not only driven by the overlap between an event's predicted semantic predictions and the semantic features of the bottom-up input, which facilitates the mapping of the incoming word's semantic features on to a prior event-level representation, but also by overlap at the *lexical* level, i.e. facilitation of the mapping between the incoming word's lexical representation and these pre-activated semantic features (see DeLong et al., 2005; Federmeier, 2007; Lau et al., 2008). This *top-down lexico-semantic facilitation* account receives support from recent MEG findings showing that the effects of contextual predictability on the N400 in plausible sentences selectively localizes to regions that support lexico-semantic processing within the left temporal cortex (Wang, Nour Eddine, Brothers, Jensen, & Kuperberg, 2024), as opposed to more widespread regions across the left lateralized language network (including higher-level regions that encode event-level information over a longer time scale).

In our simulations, the predictive coding algorithm implemented precisely the type of top-down prediction mechanisms that allowed for top-down lexico-semantic facilitation. Before presenting the bottom-up input, a probability distribution of predicted inputs at the highest conceptual layer of the model's hierarchy led to the pre-activation of expected semantic features within the semantic state units — *semantic pre-activation* (although, as discussed under Future Directions, because our model did not directly infer a higher-level event-level from sequential inputs, we provided this input externally). Critically, the feedback connections between the semantic state units and the bottom-up lexical error units additionally allowed these pre-activated semantic states to generate top-down lexical predictions (reconstructions) that suppressed the lexical prediction error produced by expected bottom-up inputs as they became available (see Supplementary Fig. 3), leading to *lexico-semantic facilitation*.[8] Moreover, just as originally posited by Federmeier and Kutas (1999), the same top-down *lexical* reconstruction mechanism was able to explain why lexical-level prediction error was attenuated to inputs that were lexically unexpected but shared semantic features with expected inputs (the anticipatory semantic overlap effect: *They wanted the hotel to look more like a tropical resort. So along the driveway, they planted rows of pines* < *tulips*; expected: *palms*).

Of most theoretical significance, the top-down propagation of predictions down the linguistic hierarchy allowed us, for the first time, to simulate the *anticipatory orthographic overlap effect*. This describes the attenuation of the N400 to unexpected inputs that share orthographic features with highly expected continuations — both real words (e.g. "*The genie granted his third and final…*" expected: WISH; DISH < CLAW; Laszlo & Federmeier, 2009; Ito et al., 2016; DeLong et al., 2019) and pseudowords (*WUSH < CLAW: Laszlo & Federmeier, 2009).

Crucially, to explain this effect, it is necessary not only to pre-activate *semantic* features and generate *lexical* reconstructions that suppress lexical prediction error, but also to pre-activate *lexical* features and generate *orthographic* reconstructions that suppress orthographic prediction error. Facilitation at the lexical-orthographic interface may not occur under all reading conditions (e.g. Nieuwland, 2019; Nieuwland et al., 2018). However, there is clear evidence that predictions at the lexico-orthographic interface can be generated in strongly constraining contexts (Wang, Brothers, Jensen, & Kuperberg, 2023), particularly when reading relatively slowly (Ito et al., 2016), and/or when the broader experimental environment encourages top-down prediction (e.g. a high proportion of highly predictable sentences, see DeLong, Chan, & Kutas, 2021). These effects cannot easily be explained by previous sentence-level models of the N400, which lack the necessary top-down feedback connections between lexical representation and orthographic features. They can, however, be explained by predictive coding, which provides a mechanism that allows anticipated information to be propagated further down the hierarchy. Specifically, in a strongly constraining context, the top-down lexical-level reconstructions computed two iterations back will induce a top-down lexical bias that pre-activates lexical state units corresponding to the expected input (*wish*) before the bottom-up input becomes available. Then, as the lexically unexpected orthographically overlapping input (*dish*) becomes available, these pre-activated lexical states generate orthographic reconstructions that suppress the orthographic prediction error, which in turn, results in a smaller prediction error at the lexical and semantic levels (and therefore a smaller simulated N400).

Of course, predictive coding is not the only architecture that allows for the top-down propagation of information down a representational hierarchy. For example, feedback connections across hierarchically organized linguistic representations are also incorporated in classic Interactive Activation and Competition (IAC) models of visual and spoken word recognition (Chen & Mirman, 2012; McClelland & Elman, 1986; McClelland & Rumelhart, 1981). However, because IAC architectures incorporate lateral inhibitory connections between lexical units, they would not be able to simulate some of the contextual effects that we successfully simulated here using predictive coding.

For example, consider how the anticipatory orthographic overlap effect might be modeled in an interactive-activation setting: In a strongly constraining context, the top-down pre-activation of the lexical representation of the expected word (*wish*) would immediately begin to inhibit the lexical representations of other words, regardless whether they are orthographically related or unrelated to the expected lexical item (e.g. *dish* or *claw*). Therefore, if an unexpected orthographically overlapping word is subsequently encountered (e.g. "dish"), its lexical representation (*dish*) would initially be just as difficult to access as the lexical representation of an unexpected unrelated input (*claw*). This would predict no difference in the N400 between these two conditions (i.e. *dish* = *claw*), contrary to the empirical findings (*dish* < *claw*). Indeed, at later stages of processing, IAC architectures would predict a competitive interference effect, with even *more* difficulty in settling on the lexical representation of an unexpected orthographically overlapping input than an unexpected unrelated input (*dish* > *claw*). This is because when the unexpected orthographically overlapping target ("dish") is encountered, it would not only activate its own lexical representation (*dish*) but it would also provide further activation to its neighbor — the incorrectly predicted lexical item (*wish*). This would

[8] Note that this implies a functional distinction between *semantic pre-activation* (the pre-activation of expected semantic states before new bottom-up input becomes available to the semantic level) and *top-down lexico-semantic facilitation* (the facilitated mapping of expected lexical inputs on to these pre-activated semantic states). In our predictive coding model, semantic pre-activation was induced by the top-down bias term, based on reconstructions from the higher conceptual level that were computed two iterations back. Top-down lexico-semantic facilitation occurred both as a consequence of this top-down semantic bias (semantic pre-activation) and the suppression of lexical prediction error by top-down lexical reconstructions that were computed one iteration back. This distinction between two types of top-down information computed by the predictive coding algorithm — the top-down bias and the top-down reconstructions — is important because it implies that processing of a bottom-up input can be facilitated at a lower level of representation (through the suppression of prediction error by top-down reconstructions), even if the state units at this lower level have not yet been pre-activated (through the top-down bias term) before the bottom-up input becomes available.

result in more competitive inhibition and more difficulty settling on the unexpected orthographically overlapping target, *dish,* than an unexpected unrelated target, *claw* (see Laszlo & Federmeier, 2009, page 334–335 for discussion).[9]

In contrast, in predictive coding, there are no lateral lexical-level inhibitory connections amongst state units, and so there is no direct competition amongst activated lexical items. Instead, as the unexpected orthographically overlapping input ("dish") appears, the pre-activated lexical state units (*wish*) generate reconstructions (W-I-S-H) that suppress the orthographic prediction error produced by the three overlapping orthographic units, (I, S, and H). As a result, only the unreconstructed part of the input (**D** from D-I-S-H) passes up to the lexical and semantic levels. This small orthographic prediction error, in turn, resulted in a smaller prediction error at these levels too, i.e. a smaller N400.

Notably, our simulations showed that even though unexpected orthographically overlapping inputs produced a smaller lexico-semantic prediction error, they still took *longer* to converge on their correct lexical and semantic state representation, similar to what would be expected in an IAC framework. However, instead of being driven by mutual inhibition amongst activated lexical units, this occurred because the inappropriately suppressed orthographic prediction error to "dish" induced only a weak update in lexical and semantic state units on each iteration of the algorithm[10](in our simulations, this prolonged the time course of the lexico-semantic error produced by unexpected orthographically overlapping inputs relative to unexpected unrelated inputs, see Fig. 8). On the assumption that the time it takes for a model to settle on an input's correct lexical representation determines behavioral response times, then predictive coding makes the interesting prediction that when an input's form but not its meaning overlaps with that of a pre-activated lexical representation, then this should result in an *attenuation* of the N400 response, but *longer* behavioral response times. We return to the idea that below under Future Directions.

4.2.3. Simulations of interaction between contextual and lexical effects

The free top-down and bottom-up flow of information across hierarchical levels also provides a natural explanation for interactions between contextual predictability and frequency (empirical findings for Predictability x Frequency, see Dambacher et al., 2006), and semantic richness (Predictability x Richness, see Holcomb et al., 1999) — both simulated for the first time here. Mirroring the empirical findings, we showed that the effects of frequency and semantic richness were weaker when the critical words were contextually predictable. In both cases, pre-activation of the state units generated reconstructions that rapidly suppressed the production of lexico-semantic prediction error in response to expected inputs, regardless of their lexical characteristics.

Of note, in our simulations, the effects of contextual predictability interacted not only with frequency and semantic richness, but also with orthographic neighborhood size; that is, the effect of orthographic neighborhood size on the simulated N400 was smaller in more predictable contexts. The empirical data speaking to whether contextual predictability interacts with orthographic neighborhood size are less clear than for the interactions described above. To our knowledge, no previous study has orthogonally manipulated lexical predictability and orthographic neighborhood size in a controlled experimental design. However, Payne et al. (2015, 2018) compared the effect of orthographic neighborhood size on the N400 produced by words appearing later in sentences (more predictable) and words appearing earlier in sentences (less predictable). In contrast to what we show in the present simulations, these authors reported that the orthographic neighborhood size effect on the N400 (high > low) was just as large on more expected versus less expected words.

We suggest that the reason for this discrepancy is that the comprehenders in Payne et al.'s studies were not generating predictions all the way down to the lexical-orthographic interface — a necessary step for producing this interaction. Specifically, in our simulations, unexpected words with many orthographic neighbors produce a large lexico-semantic prediction error and a large N400 because these neighbors activated their lexical and semantic state units (just as for words presented in the absence of a prior context). However, in high constraint contexts, the pre-activation of the lexical state units led to the generation of orthographic reconstructions that immediately suppressed the orthographic prediction error to the expected high neighborhood inputs. This, in turn, suppressed the broad activation of lexical neighbors, and their corresponding semantic features, resulting in a small lexico-semantic prediction error and a small N400 on these expected high neighborhood words, thereby explaining why contextual predictability interacted with orthographic neighborhood size. However, as discussed earlier in relation to the anticipatory orthographic overlap effect, there is evidence that top-down lexical-orthographic predictions are not routinely generated under all reading conditions (Ito et al., 2016; Nieuwland, 2019). In Paynes' studies (Payne et al., 2015; Payne & Federmeier, 2018), the SOA was relatively short (500 ms), and the predictive validity of the environment was low (with a high proportion of sentences with syntactic prose or completely scrambled), explaining why predictability and orthographic neighborhood size did not interact on the N400 in their study. On this account, one should see an interaction between these two variables under reading conditions that are known to encourage the generation of lower-level lexical-orthographic predictions (e.g. reading at slower pace; a high predictive validity

[9] It less clear what an IAC architecture would predict for the anticipatory semantic overlap effect, where the pre-activation phase is followed by the presentation of a lexically unexpected but semantically overlapping input. In this case, the target item might also receive lateral inhibition from the pre-activated lexical competitor. However, this might be offset to some degree by top-down activation from semantic features that are share between the target and competitor. As such, one might expect to see some attenuation on the N400 to lexically unexpected semantically overlapping continuations, relative to lexically unexpected semantically unrelated continuations, albeit less prominently than what we observed in our predictive coding simulations (see Brothers et al., 2023 for recent discussion).

[10] Note that the eventual selection of the correct lexical item (*dish*) over the pre-activated competitor (*wish*) also constitutes a type of competition. However, instead of competing through lateral inhibition, the two items compete for the bottom-up activation induced by the D — the unexplained bottom-up input/orthographic prediction error, which, in this case, is very small. As activity eventually rises over the correct target item (*dish*), it falls over *wish*. Thus predictive coding approximates a type of Bayesian reasoning known as "explaining away" (Pearl, 1988; see Spratling, De Meyer, & Kompass, 2009; Spratling, 2016a in relation to vision, and Brothers et al., 2023 for recent discussion in relation to language processing).

environment). An important goal for future empirical studies will be to test this hypothesis.[11]

4.3. Implications

Taken together, our findings are consistent with the theory that the brain uses predictive coding to comprehend language. Of course, the present findings do not provide definitive support for this theory: Previous models have also simulated several of the N400 effects reported here. However, as we have discussed, there are several aspects of predictive coding that make it particularly promising for explaining why the N400 has emerged as such a key neural signature of language processing. These include: (1) the central functional role that lexico-semantic prediction error (the N400) plays in mapping form on to meaning; (2) the intuitive and biologically plausible connection to neural activity; (3) the ability to explain why and how higher-level information can be propagated down to lower levels of the linguistic and cortical hierarchy, and interact with incoming information without competitive inhibition amongst lexical representations, and (4) the ability to explain why effects of both context and priming on the N400 localize to regions of the left temporal cortex that support lexico-semantic processing (semantic priming: Nobre & McCarthy, 1995; Lau, Weber, Gramfort, Hämäläinen, & Kuperberg, 2016; Lau, Gramfort, Hämäläinen, & Kuperberg, 2013; effects of lexical predictability in plausible sentences: Wang et al., 2023).

By showing that the N400 evoked response can be understood as the production of lexico-semantic prediction error in a predictive coding framework, our findings also directly link the large existing N400 literature to previous MEG and fMRI studies in speech perception (Blank & Davis, 2016; Sohoglu & Davis, 2020), and visual word recognition (Price & Devlin, 2011) in which larger neural response to unexpected versus expected inputs have been interpreted as prediction error generated at lower levels of language hierarchy.

More generally, our findings link the N400 to the large body of research in predictive coding across non-linguistic perceptual and cognitive domains (Clark, 2013; Spratling, 2016b). Indeed, our model was based directly on models that were originally developed to explain low-level visual phenomena (Rao & Ballard, 1999; Spratling, 2012, 2013, 2014): the basic structure of the predictive coding architecture and its connections, as well as the steps of the predictive coding algorithm, were largely unchanged. Thus, by mapping the univariate N400 evoked response on to a distinct computational element in this architecture (error units), and showing that its temporal dynamics can be explained by the algorithm's optimization goal (the minimization of prediction error), our findings lend support to the hypothesis that the brain employs the same canonical circuit motif (cf. Douglas, Martin, & Whitteridge, 1989) to process language as in other perceptual and cognitive operations (Aitchison & Lengyel, 2017; Bastos et al., 2012). Some evidence for this hypothesis comes from MEG findings suggesting that the differential activation of state versus error units by expected and unexpected inputs over the course of the predictive coding algorithm can account for the dynamics of multivariate brain activity within the same 300-500 ms time window (Wang et al., 2024).

4.4. Limitations and future directions

The current version of our model has several limitations, opening up several potential avenues for future expansion.

4.4.1. Expanding the hierarchy upwards

One important limitation is that we provided probabilistic predictions externally to a simplified conceptual layer at top of the hierarchy. In reality, however, this highest-level layer would encode a *situation model* (van Dijk & Kintsch, 1983; Zwaan & Radvansky, 1998) that generates probabilistic predictions of upcoming semantic features. We conceptualize this situation model as an ongoing high-level interpretation that encodes not only a surface representation of events that are inferred from the prior linguistic input, but also knowledge retrieved from long-term memory that is specifically relevant to the current communicative situation. To incorporate this type of situation model, the highest conceptual layer of the model would need to be expanded in two main ways.

First, it would need to incorporate a mechanism for inferring events that are based on the full sequence of bottom-up linguistic inputs, similar to previous sentence-level models of the N400 (Brouwer et al., 2017,; Rabovsky et al., 2018; Fitz & Chang, 2019). This would allow us to model certain additional sentence-level phenomena, such as the effects of semantic attraction on unexpected words (Kuperberg, Sitnikova, Caplan, & Holcomb, 2003; simulated by Brouwer et al., 2017, and by Rabovsky et al., 2018), and the effect of word position within sentences (Van Petten & Kutas, 1990, Van Petten & Kutas, 1990; simulated by Rabovsky et al., 2018).

Second, it would need to incorporate a mechanism for inferring the current communicative situation or topic under discussion, along with its associated schema-relevant real-world and episodic knowledge stored within long-term memory (see Franklin, Norman, Ranganath, Zacks, & Gershman, 2020 for one modeling approach). This would allow us to simulate several additional phenomena on the N400, including various discourse-level and pragmatic effects (e.g. Van Berkum, Zwitserlood, Hagoort, & Brown, 2003); Kuperberg, Paczynski, & Ditman, 2011; Van Berkum, Van den Brink, Tesink, Kos, & Hagoort, 2008).

Finally, expanding the highest conceptual layer may allow us to simulate the effects of semantic implausibility/anomaly on the N400 (Kuperberg et al., 2020; Nieuwland et al., 2020). In contrast to the effect of lexical probability on the N400, which, in plausible sentences, localizes to regions of the left temporal cortex that support lexico-semantic processing, the anomaly effect on the N400 additionally localizes to the left inferior frontal cortex at a higher level of the cortical hierarchy (Wang, Schoot, et al., 2023). It has been suggested that this additional evoked response reflects the generation of a higher-level prediction error within the N400 time window when the newly inferred highly implausible or anomalous event cannot be explained by the model's generative parameters (Wang, Brothers, et al., 2023; and see Rao & Ballard, 1999 for analogous discussion in the visual system). The incorporation of error units at the highest layer of the hierarchy would allow us to test this hypothesis.

[11] If one does see an interaction between Contextual predictability and ONSize under experimental conditions that encourage top-down lexical-orthographic prediction, then this would provide evidence (a) that the effect of ONsize indeed stems from activity at the lexico-orthographic interface, and (b) that the attenuation of this effect on expected words is indeed driven by the top-down suppression of orthographic prediction error by orthographic reconstructions, as shown by our simulations. If, however, if it turns out that there is no interaction between these variables under these conditions, i.e. that the effect of ONSize is impervious to contextual predictability, then this would provide evidence for an alternative account of the effect of ONSize on the N400: that, instead of being driven by activity at the lexico-orthographic interface, it is driven by differences in how high ONsize and low ONsize words are encoded at a lower sublexical orthographic level of representation. For example, in a bigram-based coding scheme where each orthographic unit represents one bigram, words with smaller orthographic neighborhood sizes (e.g., *kiwi*) would require fewer bigrams to be uniquely represented, requiring fewer state units to be activated than words with larger neighborhood sizes (e.g. *core*). Sublexical representations are even more unlikely to be pre-activated during routine natural language comprehension. Therefore, the orthographic neighborhood size effect would be relatively invariant to top-down factors like contextual predictability. If this turns out to be the case, expansions of the predictive coding hierarchy down to include more realistic sublexical representations would be necessary to test this hypothesis, as we discuss further below.

4.4.2. Expanding the hierarchy downwards

A second limitation of the current model is that we used a simplified orthographic code that directly mapped groups of letters onto lexical entries. In reality, however, there is a much deeper hierarchy of representations that lie between the model's inputs and the lexical and semantic layers, including early visual representations as well as sublexical orthographic representations (bigrams and trigrams; Vinckier et al., 2007; Dehaene, Cohen, Sigman, & Vinckier, 2005). Moreover, written word recognition does not only rely on direct mappings between orthography and semantics, but also on closely interacting lexical and sublexical representations that map indirectly on to semantics via phonology (Grainger & Holcomb, 2009; Harm & Seidenberg, 2004; Seidenberg & McClelland, 1989). We suggest that expanding the current predictive coding hierarchy downwards to include these lower-level representations would offer several advantages.[12]

First, it might help stabilize the peak latency of the simulated N400. In the empirical literature, the peak latency of the N400 priming effect and the lexical predictability effect is very stable (Federmeier & Laszlo, 2009). In contrast, in our simulations, the peak latency of these effects varied with the magnitude of lexico-semantic prediction error (the smaller the simulated N400, the earlier its peak latency). We suggest that this discrepancy may, in part, be due to the artificially truncated pathway between the input and the target lexical and semantic representations. Because the complete set of letters directly activated lexical state units, there was essentially no limit to how rapidly lexico-semantic states could converge and suppress prediction error. In contrast, during natural language comprehension, the deeper hierarchy of lower-level representations likely imposes a structural constraint on how quickly information reaches the higher lexical and semantic layers. Indeed, as shown in the Supplementary Materials, prediction error generated at the higher semantic level appeared to exhibit greater stability in terms of its latency compared to that generated at lower levels.

A second advantage of extending the model hierarchy downwards is that it would allow us to simulate a number of additional ERP effects that we were not able to simulate using the slot-based coding scheme used in current model. We adopted this code because of its simplicity and widespread use (Grainger & Jacobs, 1996; Harm & Seidenberg, 1999; McClelland & Rumelhart, 1981; Zorzi, Houghton, & Butterworth, 1998). However, there are known issues with this type of coding scheme (see Davis & Bowers, 2006), and the incorporation of more realistic sublexical bigram and trigram representations would enable us to simulate effects such as the attenuation of the N400 to targets preceded by primes with transposed letters versus control primes (*leomn - lemon* vs. *leuzn - lemon*, see Grainger, 2008; Grainger, Kiyonaga, & Holcomb, 2006; Carreiras, Vergara, & Perea, 2009; Meade, Mahnich, Holcomb, & Grainger, 2021). Additionally, expanding the model to include phonological representations that interact with sublexical and lexical entries would allow us to simulate the smaller N400 to targets preceded by pseudo-homophone primes (*brane – brain)* versus pseudoword primes (*brans – brain*; Grainger et al., 2006), as well as the effects of phonological neighborhood on written words (Carrasco-Ortiz, Midgley, Grainger, & Holcomb, 2017).

A third advantage of incorporating additional sublexical orthographic and phonological representations into the model is that it would allow us to explore whether predictive coding is also able to explain effects on several earlier negative-going ERP components observed before the N400 time-window between 150 and 300 ms (Grainger & Holcomb, 2009; Kiyonaga et al., 2007). These early ERP components have previously been interpreted within the hierarchical interactive activation framework as indexing activity at sublexical levels of representation (see Grainger & Holcomb, 2009). Within a predictive coding framework, they may reflect the generation of prediction errors generated at lower sublexical levels of the linguistic hierarchy (see Price & Devlin, 2011).

4.4.3. Simulations of behavior

In the present study, our focus was on simulating the N400 ERP response, which we operationalized as lexico-semantic prediction error — the total activity produced by lexical and semantic *error units* on each iteration of the algorithm. As we have emphasized, these error units work in close conjunction with separate sets of *state units* that encode the bottom-up input, regardless of its predictability. Thus, at the same time that error units are producing the transient evoked N400 response, these state units are accumulating activity that encodes the lexical identity and semantic features of the bottom-up input. This information encoded within the state units may play an important role in guiding decision-making during behavioral tasks. In current work, we are showing that by placing a decision threshold on lexical state activity and examining the time (number of iterations) it takes for the algorithm to cross this threshold, it is possible to simulate various effects in the behavioral literature (Nour Eddine, Brothers, Wang and Kuperberg, 2023).

Indeed, the separation of state and error units makes predictive coding particularly promising for understanding why effects on the N400 and behavior often pattern together, but at other times they dissociate. For example, in repetition priming, semantic priming, and contextual predictability, the smaller the N400, the greater the behavioral facilitation. In predictive coding, this is because, at the same time as top-down reconstructions are suppressing lexico-semantic prediction error (within error units), the top-down bias provides a head-start for state units to converge on the correct expected conceptual and semantic representations. In other situations, however, the N400 is attenuated even when there is evidence of behavioral interference (e.g. Holcomb et al., 2002). As discussed earlier in relation to the anticipatory orthographic overlap effect, predictive coding can, in principle, account for this type of dissociation. For example, if an incorrect lexical state unit is pre-activated (via the top-down bias term), and this state does not share semantic features with the expected input, then when a target input is encountered it will produce top-down orthographic reconstructions that suppress orthographic prediction error, meaning that less lexico-semantic prediction error will be propagated up the hierarchy. This will result in both a smaller N400 and a longer time for state units to converge on the correct lexical and semantic representations. We are currently carrying out experiments and simulations to directly test this hypothesis.

4.4.4. Learning and adaptation

Another focus for future research will be to examine the relationship between prediction error, comprehension and learning within the predictive coding framework. The close link between prediction error and learning is well documented in both non-linguistic (Rescorla & Wagner, 1972; Rumelhart, Hinton, & Williams, 1986) and linguistic domains (e.g., Chang, Dell, & Bock, 2006; Elman, 1990). This link has also been emphasized in previous models of the N400 that proposed that prediction error (Fitz & Chang, 2019; Rabovsky & McRae, 2014) or changes-in-state (Rabovsky et al., 2018) are computed for the purpose of downstream learning through backpropagation. As noted earlier, in these previous models, prediction error played no functional role in comprehension, whereas in the current predictive coding model, locally-computed prediction error played a crucial role in inference/comprehension. This, however, doesn't imply that the prediction error computed during predictive coding isn't *also* used for downstream

[12] Another direction for future research would be to construct a predictive coding model of spoken language comprehension that takes acoustic inputs, and that includes layers encoding phonetic features and phonemes, in addition to lexical and semantic representations. This model would allow researchers to simulate the N400 generated during spoken language comprehension. It would also enable us to simulate the production of lower-level phonemic prediction error, which, in previous fMRI and MEG studies, has been hypothesized to drive the larger neural response produced by unpredictable versus predictable spoken inputs (Blank & Davis, 2016; Sohoglu & Davis, 2020).

learning. Indeed, it has been shown that under certain theoretical assumptions, this error converges to the learning signal used for back-propagation (Millidge, Tschantz, & Buckley, 2020; Song, Lukasiewicz, Xu, & Bogacz, 2020; Whittington & Bogacz, 2019).

In the current implementation of our model, the connection weights between the layers were hand-coded rather than trained. This modeling choice allowed us to interpret the activity in state and error units, and explicitly link them to psycholinguistic representations. However, it would be possible to include a weight update step in our current algorithm, allowing the model to use prediction error minimization not only to drive shorter-term inference, but also to drive longer-term learning/adaptation, i.e. modifying the generative parameters that define weights across levels of representation (Rao & Ballard, 1997; Rao & Ballard, 1999; Spratling, 2012). Under this scenario, the model would alternate between short-term inference (fixing the connection weights and updating the state units) and longer-term learning (fixing the state units and updating the weights), capturing the idea that language processing and language learning/adaptation are closely intertwined throughout the lifespan (see Dell & Chang, 2014; Elman, 1990; Kleinschmidt & Jaeger, 2015).

4.4.5. Late positivities

Finally, it will be important for future models to simulate a set of later positive-going ERP components, observed beyond the N400 time window between 600 and 1000 ms. Two previous computational models have attempted to simulate late positivities (Brouwer et al., 2017; Fitz & Chang, 2019), but each has limitations. For example, neither model explains why these late positivities are most likely to be produced when comprehenders have established a high-level situation model (see Kuperberg et al., 2020; Brothers, Wlotko, Warnke, & Kuperberg, 2020 for empirical evidence). In addition, neither model distinguishes between two distinct types of late positivities (Kuperberg et al., 2020; Van Petten & Luka, 2012).

The first is a *frontally* distributed positivity that is produced when unexpected inputs can be *successfully* integrated into the prior context, but this involves a large update to the situation model, e.g. when the input violates a strong prior prediction (Federmeier et al., 2007; Kuperberg et al., 2020; Kutas, 1993), or if it is particularly informative, inducing the retrieval of new schema-relevant events from long-term memory (e.g. Brothers, Greene, & Kuperberg, 2020; Davenport & Coulson, 2011; Thornhill & Van Petten, 2012). In both these situations, the updated situation model will produce new top-down reconstructions containing residual semantic and lexical information that is not yet encoded in lower-level semantic and lexical state units, i.e. top-down error. It has been argued that the late frontal positivity indexes the generation of this late top-down error (Wang, Brothers, et al., 2023), as it is propagated down the generative hierarchy, serving to *retroactively* update the lower-level semantic and lexical state units (via the top-down bias term), ensuring that the information is consistent across all levels. We are currently carrying out simulations to test this hypothesis.

The second type of late positivity is a posteriorly distributed component, otherwise known as the P600. Late posterior positivities/P600s are produced by syntactic anomalies (e.g., Hagoort, Brown, & Groothusen, 1993), orthographic anomalies (e.g., Vissers et al., 2006), and highly implausible or semantically anomalous continuations (e.g., Münte, Heinze, Matzke, Wieringa, & Johannes, 1998; Kuperberg et al., 2003; Kuperberg, 2007; van de Meerendonk, Kolk, Chwilla, & Vissers, 2009). In these situations, the predictive coding algorithm will fail to converge (fail to minimize prediction error) within the N400 time window, leading to conflict and reprocessing (van de Meerendonk et al., 2009). It has been recently argued that, during this second stage of reprocessing, the P600 tracks the brain's rising confidence that this conflict stemmed from an error within the external input (as opposed to an internal processing error), i.e. the brain's rising confidence that the input cannot be used to update the situation model, given the generative model's prior state and parameters (see Kuperberg, Alexander, & Brothers, 2024). This type of confidence tracking may play an essential role in adapting to future external linguistic errors (e.g., Coulson, King, & Kutas, 1998; Hanulikova, van Alphen, van Goch, & Weber, 2012).

4.5. Conclusion

Predictive coding offers a simple, interpretable and biologically plausible framework to make sense of prediction in language comprehension. Our simulations show that lexico-semantic prediction error within this framework shares a remarkable range of features with the N400 event-related component, from its time course, to its sensitivity to top-down and bottom-up variables, and even interactions between these variables. By mapping the N400 on to a distinct element within a hierarchical generative modeling framework (lexico-semantic prediction error), we situate this key neural component within the broader context of predictive coding research. Most importantly, our findings raise the possibility that the brain uses predictive coding to infer meaning from the form of words during language comprehension. This paves the way towards understanding how specific disruptions of predictive coding might give rise to the pathological neural responses observed during language processing in neurodevelopmental disorders such as schizophrenia (Brown & Kuperberg, 2015; Fletcher & Frith, 2009).

CRediT authorship contribution statement

Samer Nour Eddine: Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Formal analysis, Conceptualization. **Trevor Brothers:** Writing – review & editing, Writing – original draft, Visualization, Supervision, Methodology, Formal analysis, Conceptualization. **Lin Wang:** Writing – original draft, Visualization, Supervision, Methodology, Conceptualization. **Michael Spratling:** Supervision, Software, Methodology, Conceptualization. **Gina R. Kuperberg:** Writing – review & editing, Writing – original draft, Visualization, Supervision, Conceptualization.

Declaration of competing interest

The authors declare no conflict of interest.

Data availability

The code to reproduce all simulations and analyses is available on anonymized GitHub (https://anonymous.4open.science/r/PredictiveCodingModel_N400-34CA/README.md) and OSF (https://osf.io/f7upd/?view_only=8dfb9f87a10b44aebb24e23de1d412b7).

Acknowledgments

This work was funded by the National Institute of Child Health and Human Development (R01MHD082527) to G.R.K. We thank Arim Choi Perrachione for her help with the figures. We also thank Jeff Stibel for his support of Drs. Kuperberg and Wang.

Appendix A. Supplementary data

The code to reproduce all simulations and analyses is available on GitHub (https://github.com/samer-noureddine/PredictiveCodingModel_N400).

The supplementary materials can be found at the following link (https://osf.io/n4cp6). This includes a more technical description of the predictive coding algorithm, weight matrices, hyperparameters, frequency implementation and a derivation of the state update step of the algorithm.

References

Aitchison, L., & Lengyel, M. (2017). With or without you: Predictive coding and Bayesian inference in the brain. *Current Opinion in Neurobiology, 46*, 219–227. https://doi.org/10.1016/j.conb.2017.08.010

Amsel, B. D. (2011). Tracking real-time neural activation of conceptual knowledge using single-trial event-related potentials. *Neuropsychologia, 49*(5), 970–983. https://doi.org/10.1016/j.neuropsychologia.2011.01.003

Baggio, G., & Hagoort, P. (2011). The balance between memory and unification in semantics: A dynamic account of the N400. *Language & Cognitive Processes, 26*(9), 1338–1367. https://doi.org/10.1080/01690965.2010.542671

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. https://doi.org/10.1016/j.jml.2012.11.001

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. https://doi.org/10.1016/j.neuron.2012.10.038

Bates, D. M., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

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

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

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

Bornkessel-Schlesewsky, I., & Schlesewsky, M. (2019). Toward a neurobiologically plausible model of language-related, negative event-related potentials. *Frontiers in Psychology, 10*, 298. https://doi.org/10.3389/fpsyg.2019.00298

Braun, M., Jacobs, A. M., Hahne, A., Ricker, B., Hofmann, M., & Hutzler, F. (2006). Model-generated lexical activity predicts graded ERP amplitudes in lexical decision. *Brain Research, 1073-1074*, 431–439. https://doi.org/10.1016/j.brainres.2005.12.078

Brothers, T., Greene, S., & Kuperberg, G. R. (2020). Distinct neural signatures of semantic retrieval and event updating during discourse comprehension. In *Paper presented at the 27th Annual Meeting of the Cognitive Neuroscience Society, Boston, MA*.

Brothers, T., & Kuperberg, G. R. (2021). Word predictability effects are linear, not logarithmic: Implications for probabilistic models of sentence comprehension. *Journal of Memory and Language, 116*. https://doi.org/10.1016/j.jml.2020.104174

Brothers, T., Morgan, E., Yacovone, A., & Kuperberg, G. R. (2023). Multiple predictions during language comprehension: Friends, foes, or indifferent companions? *Cognition, 241*, Article 105602. https://doi.org/10.1016/j.cognition.2023.105602

Brothers, T., Swaab, T. Y., & Traxler, M. J. (2015). Effects of prediction and contextual support on lexical processing: Prediction takes precedence. *Cognition, 136*, 135–149. https://doi.org/10.1016/j.cognition.2014.10.017

Brothers, T., Wlotko, E. W., Warnke, L., & Kuperberg, G. R. (2020). Going the extra mile: Effects of discourse context on two late positivities during language comprehension. *Neurobiology of Language, 1*(1), 135–160. https://doi.org/10.1162/nol_a_00006

Brouwer, H., & Crocker, M. W. (2017). On the proper treatment of the N400 and P600 in language comprehension. *Frontiers in Psychology, 8*, 1327. https://doi.org/10.3389/fpsyg.2017.01327

Brouwer, H., Crocker, M. W., Venhuizen, N. J., & Hoeks, J. C. J. (2017). A neurocomputational model of the N400 and the P600 in language processing. *Cognitive Science, 41*(Suppl. 6), 1318–1352. https://doi.org/10.1111/cogs.12461

Brown, C., & Hagoort, P. (1993). The processing nature of the N400: Evidence from masked priming. *Journal of Cognitive Neuroscience, 5*(1), 34–44. https://doi.org/10.1162/jocn.1993.5.1.34

Brown, M., & Kuperberg, G. R. (2015). A hierarchical generative framework of language processing: Linking language perception, interpretation, and production abnormalities in schizophrenia. *Frontiers in Human Neuroscience, 9*, 643. https://doi.org/10.3389/fnhum.2015.00643

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., … Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 1877–1901. https://doi.org/10.5555/3495724.3495883

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

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods, 46*(3), 904–911. https://doi.org/10.3758/s13428-013-0403-5

Carrasco-Ortiz, H., Midgley, K. J., Grainger, J., & Holcomb, P. J. (2017). Interactions in the neighborhood: Effects of orthographic and phonological neighbors on N400 amplitude. *Journal of Neurolinguistics, 41*, 1–10. https://doi.org/10.1016/j.jneuroling.2016.06.007

Carreiras, M., Vergara, M., & Perea, M. (2009). ERP correlates of transposed-letter priming effects: The role of vowels versus consonants. *Psychophysiology, 46*(1), 34–42. https://doi.org/10.1111/j.1469-8986.2008.00725.x

Chang, F., Dell, G. S., & Bock, J. K. (2006). Becoming syntactic. *Psychological Review, 113*(2), 234–272. https://doi.org/10.1037/0033-295x.113.2.234

Chater, N., Crocker, M. W., & Pickering, M. J. (1998). The rational analysis of inquiry: The case of parsing. In M. Oaksford, & N. Chater (Eds.), *Rational models of cognition* (pp. 441–468). New York: Oxford University Press.

Chen, Q., & Mirman, D. (2012). Competition and cooperation among similar representations: toward a unified account of facilitative and inhibitory effects of lexical neighbors. *Psychological review, 119*(2), 417–430. https://doi.org/10.1037/a0027175

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. https://doi.org/10.1016/j.cognition.2016.10.016

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *The Behavioral and Brain Sciences, 36*(3), 181–204. https://doi.org/10.1017/S0140525X12000477

Coulson, S., King, J. W., & Kutas, M. (1998). Expect the unexpected: Event-related brain responses to morphosyntactic violations. *Language & Cognitive Processes, 13*(1), 21–58. https://doi.org/10.1080/016909698386582

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. https://doi.org/10.1016/j.brainres.2006.02.010

Davenport, T., & Coulson, S. (2011). Predictability and novelty in literal language comprehension: An ERP study. *Brain Research, 1418*, 70–82. https://doi.org/10.1016/J.Brainres.2011.07.039

Davis, C. J., & Bowers, J. S. (2006). Contrasting five different theories of letter position coding: Evidence from orthographic similarity effects. *Journal of Experimental Psychology. Human Perception and Performance, 32*(3), 535–557. https://doi.org/10.1037/0096-1523.32.3.535

Deacon, D., Dynowska, A., Ritter, W., & Grose-Fifer, J. (2004). Repetition and semantic priming of nonwords: Implications for theories of N400 and word recognition. *Psychophysiology, 41*(1), 60–74. https://doi.org/10.1111/1469-8986.00120

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

Delaney-Busch, N., Morgan, E., Lau, E., & Kuperberg, G. R. (2019). Neural evidence for Bayesian trial-by-trial adaptation on the N400 during semantic priming. *Cognition, 187*(June 2019), 10–20. https://doi.org/10.1016/j.cognition.2019.01.001

Dell, G. S., & Chang, F. (2014). The P-chain: Relating sentence production and its disorders to comprehension and acquisition. *Philosophical Transactions of the Royal Society, B: Biological Sciences, 369*(1634), 20120394. https://doi.org/10.1098/rstb.2012.0394

DeLong, K. A., Chan, W. H., & Kutas, M. (2019). Similar time courses for word form and meaning preactivation during sentence comprehension. *Psychophysiology, 56*(4), Article e13312. https://doi.org/10.1111/psyp.13312

DeLong, K. A., Chan, W. H., & Kutas, M. (2021). Testing limits: ERP evidence for word form preactivation during speeded sentence reading. *Psychophysiology, 58*(2), Article e13720. https://doi.org/10.1111/psyp.13720

DeLong, K. A., & Kutas, M. (2020). Comprehending surprising sentences: Sensitivity of post-N400 positivities to contextual congruity and semantic relatedness. *Language, Cognition and Neuroscience, 35*(8), 1044–1063. https://doi.org/10.1080/23273798.2019.1708960

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

van Dijk, T. A., & Kintsch, W. (1983). *Strategies of discourse comprehension*. New York: Academic Press.

Douglas, R. J., Martin, K. A. C., & Whitteridge, D. (1989). A canonical microcircuit for Neocortex. *Neural Computation, 1*(4), 480–488. https://doi.org/10.1162/neco.1989.1.4.480

Elman, J. L. (1990). Finding structure in time. *Cognitive Science, 14*(2), 179–211. https://doi.org/10.1207/s15516709cog1402_1

Elman, J. L., & McClelland, J. L. (1984). Speech perception as a cognitive process: The interactive activation model. In N. Lass (Ed.), *Vol. 10. Speech and language* (pp. 337–374). New York: Academic Press.

Federmeier, K. D. (2007). Thinking ahead: The role and roots of prediction in language comprehension. *Psychophysiology, 44*(4), 491–505. https://doi.org/10.1111/j.1469-8986.2007.00531.x

Federmeier, K. D. (2022). Connecting and considering: Electrophysiology provides insights into comprehension. *Psychophysiology, 59*(1), Article e13940. https://doi.org/10.1111/psyp.13940

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. https://doi.org/10.1006/Jmla.1999.2660

Federmeier, K. D., & Laszlo, S. (2009). Time for meaning: Electrophysiology provides insights into the dynamics of representation and processing in semantic memory. *Psychology of Learning and Motivation, 51*, 1–44. https://doi.org/10.1016/S0079-7421(09)51001-8

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. https://doi.org/10.1016/j.brainres.2006.06.101

Fitz, H., & Chang, F. (2019). Language ERPs reflect learning through prediction error propagation. *Cognitive Psychology, 111*, 15–52. https://doi.org/10.1016/j.cogpsych.2019.03.002

Fletcher, P. C., & Frith, C. D. (2009). Perceiving is believing: A Bayesian approach to explaining the positive symptoms of schizophrenia. *Nature Reviews. Neuroscience, 10*(1), 48–58.

Franklin, N. T., Norman, K. A., Ranganath, C., Zacks, J. M., & Gershman, S. J. (2020). Structured event memory: A neuro-symbolic model of event cognition. *Psychological Review, 127*(3), 327–361. https://doi.org/10.1037/rev0000177

Friston, K. J. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society of London B: Biological Sciences, 360*(1456), 815–836. https://doi.org/10.1098/Rstb.2005.1622

Grainger, J. (2008). Cracking the orthographic code: An introduction. *Language & Cognitive Processes, 23*(1), 1–35. https://doi.org/10.1080/01690960701578013
Grainger, J., & Holcomb, P. J. (2009). Watching the word go by: On the time-course of component processes in visual word recognition. *Lang & Ling Compass, 3*, 128–156.
Grainger, J., & Jacobs, A. M. (1996). Orthographic processing in visual word recognition: A multiple read-out model. *Psychological Review, 103*(3), 518–565. https://doi.org/10.1037/0033-295x.103.3.518
Grainger, J., Kiyonaga, K., & Holcomb, P. J. (2006). The time course of orthographic and phonological code activation. *Psychological Science, 17*(12), 1021–1026.
Hagoort, P., Baggio, G., & Willems, R. M. (2009). Semantic unification. In M. S. Gazzaniga (Ed.), *The cognitive neurosciences* (4th ed., pp. 819–836). Cambridge, MA: MIT Press.
Hagoort, P., Brown, C., & Groothusen, J. (1993). The syntactic positive shift (SPS) as an ERP measure of syntactic processing. *Language & Cognitive Processes, 8*(4), 439–483. https://doi.org/10.1080/01690969308407585
Hanulikova, A., van Alphen, P. M., van Goch, M. M., & Weber, A. (2012). When one person's mistake is another's standard usage: The effect of foreign accent on syntactic processing. *Journal of Cognitive Neuroscience, 24*(4), 878–887. https://doi.org/10.1162/jocn_a_00103
Harm, M. W., & Seidenberg, M. (1999). Phonology, reading acquisition, and dyslexia: Insights from connectionist models. *Psychological Review, 106*(3), 491–528.
Harm, M. W., & Seidenberg, M. S. (2004). Computing the meanings of words in reading: Cooperative division of labor between visual and phonological processes. *Psychological Review, 111*(3), 662–720. https://doi.org/10.1037/0033-295X.111.3.662
Hauk, O., Davis, M. H., Ford, M., Pulvermuller, 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. https://doi.org/10.1016/j.neuroimage.2005.11.048
Heilbron, M., Armeni, K., Schoffelen, J.-M., Hagoort, P., & de Lange, F. P. (2022). A hierarchy of linguistic predictions during natural language comprehension. *Proceedings of the National Academy of Sciences of the United States of America, 119*(32), Article e2201968119. https://doi.org/10.1073/pnas.2201968119
Holcomb, P. J. (1988). Automatic and attentional processing: An event-related brain potential analysis of semantic priming. *Brain and Language, 35*(1), 66–85.
Holcomb, P. J., & Grainger, J. (2006). On the time course of visual word recognition: An event-related potential investigation using masked repetition priming. *Journal of Cognitive Neuroscience, 18*(10), 1631–1643.
Holcomb, P. J., Grainger, J., & O'Rourke, T. (2002). An electrophysiological study of the effects of orthographic neighborhood size on printed word perception. *Journal of Cognitive Neuroscience, 14*(6), 938–950. https://doi.org/10.1162/089892902760191153
Holcomb, P. J., Kounios, J., Anderson, J. E., & West, W. C. (1999). Dual-coding, context-availability, and concreteness effects in sentence comprehension: An electrophysiological investigation. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 25*(3), 721–742.
Holcomb, P. J., & Neville, H. J. (1990). Auditory and visual semantic priming in lexical decision: A comparison using event-related brain potentials. *Language & Cognitive Processes, 5*(4), 281–312. https://doi.org/10.1080/01690969008407065
Ito, A., Corley, M., Pickering, M., Martin, A. E., & Nieuwland, M. S. (2016). Predicting form and meaning: Evidence from brain potentials. *Journal of Memory and Language, 86*, 157–171. https://doi.org/10.1016/j.jml.2015.10.007
Kiyonaga, K., Grainger, J., Midgley, K., & Holcomb, P. J. (2007). Masked cross-modal repetition priming: An event-related potential investigation. *Language & Cognitive Processes, 22*(3), 337–376.
Kleinschmidt, D. F., & Jaeger, F. T. (2015). Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel. *Psychological Review, 122*(2), 148–203. https://doi.org/10.1037/a0038695
Kounios, J., Green, D. L., Payne, L., Fleck, J. I., Grondin, R., & McRae, K. (2009). Semantic richness and the activation of concepts in semantic memory: Evidence from event-related potentials. *Brain Research, 1282*, 95–102. https://doi.org/10.1016/j.brainres.2009.05.092
Kounios, J., & Holcomb, P. J. (1994). Concreteness effects in semantic processing: ERP evidence supporting dual-coding theory. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 20*(4), 804–823. https://doi.org/10.1037/0278-7393.20.4.804
Kuperberg, G. R. (2007). Neural mechanisms of language comprehension: Challenges to syntax. *Brain Research, 1146*, 23–49. https://doi.org/10.1016/j.brainres.2006.12.063
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. https://doi.org/10.1080/23273798.2015.1130233
Kuperberg, G. R., Alexander, E., & Brothers, T. (2024). *The posterior P600 does not reflect error correction: An information seeking account of linguistic error processing*.
Kuperberg, G. R., Brothers, T., & Wlotko, E. (2020). A tale of two positivities and the N400: Distinct neural signatures are evoked by confirmed and violated predictions at different levels of representation. *Journal of Cognitive Neuroscience, 32*(1), 12–35. https://doi.org/10.1162/jocn_a_01465
Kuperberg, G. R., & Jaeger, T. F. (2016). What do we mean by prediction in language comprehension? *Language, Cognition and Neuroscience, 31*(1), 32–59. https://doi.org/10.1080/23273798.2015.1102299
Kuperberg, G. R., Kreher, D. A., Sitnikova, T., Caplan, D. N., & Holcomb, P. J. (2007). The role of animacy and thematic relationships in processing active English sentences: Evidence from event-related potentials. *Brain and Language, 100*(3), 223–237. https://doi.org/10.1016/j.bandl.2005.12.006
Kuperberg, G. R., Paczynski, M., & Ditman, T. (2011). Establishing causal coherence across sentences: An ERP study. *Journal of Cognitive Neuroscience, 23*(5), 1230–1246. https://doi.org/10.1162/jocn.2010.21452
Kuperberg, G. R., Sitnikova, T., Caplan, D., & Holcomb, P. J. (2003). Electrophysiological distinctions in processing conceptual relationships within simple sentences. *Cognitive Brain Research, 17*(1), 117–129. https://doi.org/10.1016/S0926-6410(03)00086-7
Kutas, M. (1993). In the company of other words: Electrophysiological evidence for single-word and sentence context effects. *Language & Cognitive Processes, 8*, 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. https://doi.org/10.1146/annurev.psych.093008.131123
Kutas, M., & Hillyard, S. A. (1980). Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science, 207*(4427), 203–205. https://doi.org/10.1126/science.7350657
Kutas, M., & Hillyard, S. A. (1984). Brain potentials during reading reflect word expectancy and semantic association. *Nature, 307*(5947), 161–163. https://doi.org/10.1038/307161a0
Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software, 82*(13).
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. https://doi.org/10.1016/j.bandl.2014.03.002
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. https://doi.org/10.1016/j.jml.2009.06.004
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. https://doi.org/10.1111/j.1469-8986.2010.01058.x
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. https://doi.org/10.1016/j.bandl.2011.09.001
Lau, E. F., Gramfort, A., Hämäläinen, M. S., & Kuperberg, G. R. (2013). Automatic semantic facilitation in anterior temporal cortex revealed through multimodal neuroimaging. *The Journal of Neuroscience, 33*(43), 17174–17181. https://doi.org/10.1523/Jneurosci.1018-13.2013
Lau, E. F., Phillips, C., & Poeppel, D. (2008). A cortical network for semantics: (De) constructing the N400. *Nature Reviews. Neuroscience, 9*(12), 920–933. https://doi.org/10.1038/nrn2532
Lau, E. F., Weber, K., Gramfort, A., Hämäläinen, M. S., & Kuperberg, G. R. (2016). Spatiotemporal signatures of lexico-semantic prediction. *Cerebral Cortex, 26*(4), 1377–1387. https://doi.org/10.1093/cercor/bhu219
Lee, C.-L., & Federmeier, K. D. (2008). To watch, to see, and to differ: An event-related potential study of concreteness effects as a function of word class and lexical ambiguity. *Brain and Language, 104*(2), 145–158. https://doi.org/10.1016/j.bandl.2007.06.002
Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition, 106*(3), 1126–1177. https://doi.org/10.1016/j.cognition.2007.05.006
McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology, 18*(1), 1–86. https://doi.org/10.1016/0010-0285(86)90015-0
McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: I. An account of basic findings. *Psychological Review, 88*(5), 375–407. https://doi.org/10.1037//0033-295x.88.5.375
Meade, G., Mahnich, C., Holcomb, P. J., & Grainger, J. (2021). Orthographic neighborhood density modulates the size of transposed-letter priming effects. *Cognitive, Affective, & Behavioral Neuroscience, 21*(5), 948–959. https://doi.org/10.3758/s13415-021-00905-w
Meade, G., Midgley, K. J., Dijkstra, T., & Holcomb, P. J. (2018). Cross-language neighborhood effects in learners indicative of an integrated lexicon. *Journal of Cognitive Neuroscience, 30*(1), 70–85. https://doi.org/10.1162/jocn_a_01184
van de Meerendonk, N., Kolk, H. H. J., Chwilla, D. J., & Vissers, C. T. W. M. (2009). Monitoring in language perception. *Lang & Ling Compass, 3*(5), 1211–1224. https://doi.org/10.1111/j.1749-818X.2009.00163.x
Michaelov, J., Coulson, S., & Bergen, B. K. (Sept. 2023). So Cloze Yet So Far: N400 Amplitude Is Better Predicted by Distributional Information Than Human Predictability Judgements," in. *IEEE Transactions on Cognitive and Developmental Systems, 15*(3), 1033–1042. https://doi.org/10.1109/TCDS.2022.3176783
Millidge, B., Tschantz, A., & Buckley, C. J. (2020). *Predictive coding approximates backprop along arbitrary computation graphs*. Cornell University (arXiv).
Misra, M., & Holcomb, P. J. (2003). Event-related potential indices of masked repetition priming. *Psychophysiology, 40*(1), 115–130. https://doi.org/10.1111/1469-8986.00012
Mumford, D. (1992). On the computational architecture of the neocortex. II. The role of cortico-cortical loops. *Biological Cybernetics, 66*(3), 241–251. https://doi.org/10.1007/BF00198477
Münte, T. F., Heinze, H. J., Matzke, M., Wieringa, B. M., & Johannes, S. (1998). Brain potentials and syntactic violations revisited: No evidence for specificity of the syntactic positive shift. *Neuropsychologia, 36*(3), 217–226.
Narayanan, S., & Jurafsky, D. (2002). Combining structure and probabilities in a Bayesian model of human sentence processing. In *Paper presented at the CUNY Conference on Human Sentence Processing, New York*.

Nieuwland, M. S. (2019). Do "early" brain responses reveal word form prediction during language comprehension? A critical review. *Neuroscience and Biobehavioral Reviews, 96*, 367–400. https://doi.org/10.1016/j.neubiorev.2018.11.019

Nieuwland, M. S., Barr, D. J., Bartolozzi, F., Busch-Moreno, S., Darley, E., Donaldson, D. I., … Wolfsthurn, S. (2020). Dissociable effects of prediction and integration during language comprehension: Evidence from a large-scale study using brain potentials. *Philosophical Transactions of the Royal Society of London B: Biological Sciences, 375*(1791), Article 20180522. https://doi.org/10.1098/rstb.2018.0522

Nieuwland, M. S., Politzer-Ahles, S., Heyselaar, E., Segaert, K., Darley, E., Kazanina, N., … Huettig, F. (2018). Large-scale replication study reveals a limit on probabilistic prediction in language comprehension. *Elife, 7*, Article e33468. https://doi.org/10.7554/eLife.33468

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

Norris, D. (2006). The Bayesian reader: Explaining word recognition as an optimal Bayesian decision process. *Psychological Review, 113*(2), 327–357. https://doi.org/10.1037/0033-295X.113.2.327

Nour Eddine, S., Brothers, T., & Kuperberg, G. R. (2022). The N400 in silico: A review of computational models. In K. Federmeier (Ed.), *Vol. 76. Psychology of learning and motivation* (pp. 123–206). Academic Press.

Nour Eddine, S., Brothers, T., Wang, L., & Kuperberg, G. R. (2023). Contextual facilitation in language comprehension: Insights from a unified predictive coding framework. In *Paper presented at the 15th Annual Meeting of the Society for the Neurobiology of Language, Marseille, France*.

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. https://doi.org/10.1016/j.brainres.2018.02.021

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. https://doi.org/10.1111/psyp.12515

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. San Mateo, California: Morgan Kaufmann.

Price, C. J., & Devlin, J. T. (2011). The interactive account of ventral occipitotemporal contributions to reading. *Trends in Cognitive Sciences, 15*(6), 246–253. https://doi.org/10.1016/J.Tics.2011.04.001

R Core Team. (2022). *R: A language and environment for statistical computing (version 4.2.2)*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org.

Rabovsky, M. (2020). Change in a probabilistic representation of meaning can account for N400 effects on articles: A neural network model. *Neuropsychologia, 143*, Article 107466. https://doi.org/10.1016/j.neuropsychologia.2020.107466

Rabovsky, M., Hansen, S. S., & McClelland, J. L. (2018). Modelling the N400 brain potential as change in a probabilistic representation of meaning. *Nature Human Behaviour, 2*(9), 693–705. https://doi.org/10.1038/s41562-018-0406-4

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. https://doi.org/10.1016/j.cognition.2014.03.010

Rabovsky, M., Sommer, W., & Abdel Rahman, R. (2012a). Implicit word learning benefits from semantic richness: Electrophysiological and behavioral evidence. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 38*(4), 1076–1083. https://doi.org/10.1037/a0025646

Rabovsky, M., Sommer, W., & Abdel Rahman, R. (2012b). The time course of semantic richness effects in visual word recognition. *Frontiers in Human Neuroscience, 6*, 11. https://doi.org/10.3389/fnhum.2012.00011

Rao, R. P. N., & Ballard, D. H. (1997). Dynamic model of visual recognition predicts neural response properties in the visual cortex. *Neural Computation, 9*(4), 721–763. https://doi.org/10.1162/neco.1997.9.4.721

Rao, R. P. N., & Ballard, D. H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience, 2*(1), 79–87. https://doi.org/10.1038/4580

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In W. E. Prokasy, & A. H. Black (Eds.), *Classical conditioning II: Current research and theory* (pp. 64–99). New York: Appleton- Century-Crofts.

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

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

Rugg, M. D., Doyle, M. C., & Melan, C. (1993). An event-related potential study of the effects of within- and across-modality word repetition. *Language & Cognitive Processes, 8*(4), 357–377. https://doi.org/10.1080/01690969308407582

Rugg, M. D., & Nieto-Vegas, M. (1999). Modality-specific effects of immediate word repetition: Electrophysiological evidence. *Neuroreport, 10*(12), 2661–2664. https://doi.org/10.1097/00001756-199908200-00041

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature, 323*(6088), 533–536. https://doi.org/10.1038/323533a0

Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. *Biometric Bulletin, 2*(6), 110–114.

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

Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition, 128*(3), 302–319. https://doi.org/10.1016/j.cognition.2013.02.013

Sohoglu, E., & Davis, M. H. (2020). Rapid computations of spectrotemporal prediction error support perception of degraded speech. *Elife, 9*. https://doi.org/10.7554/eLife.58077

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

Spratling, M. W. (2008). Predictive coding as a model of biased competition in visual attention. *Vision Research, 48*(12), 1391–1408. https://doi.org/10.1016/j.visres.2008.03.009

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. https://doi.org/10.1162/NECO_a_00222

Spratling, M. W. (2013). Image segmentation using a sparse coding model of cortical area V1. *IEEE Transactions on Image Processing, 22*(4), 1631–1643. https://doi.org/10.1109/tip.2012.2235850

Spratling, M. W. (2014). A single functional model of drivers and modulators in cortex. *Journal of Computational Neuroscience, 36*(1), 97–118. https://doi.org/10.1007/s10827-013-0471-7

Spratling, M. W. (2016a). A neural implementation of Bayesian inference based on predictive coding. *Connection Science, 28*(4), 346–383. https://doi.org/10.1080/09540091.2016.1243655

Spratling, M. W. (2016b). Predictive coding as a model of cognition. *Cognitive Processing, 17*(3), 279–305. https://doi.org/10.1007/s10339-016-0765-6

Spratling, M. W. (2017). A review of predictive coding algorithms. *Brain and Cognition, 112*, 92–97. https://doi.org/10.1016/j.bandc.2015.11.003

Spratling, M. W., De Meyer, K., & Kompass, R. (2009). Unsupervised learning of overlapping image components using divisive input modulation. *Computational Intelligence and Neuroscience, 2009*, Article 381457. https://doi.org/10.1155/2009/381457

Szewczyk, J. M., & Federmeier, K. D. (2022). Context-based facilitation of semantic access follows both logarithmic and linear functions of stimulus probability. *Journal of Memory and Language, 123*. https://doi.org/10.1016/j.jml.2021.104311

Taylor, W. (1953). 'Cloze' procedure: A new tool for measuring readability. *Journalism Quarterly, 30*, 415–433.

Thornhill, D. E., & Van Petten, C. (2012). Lexical versus conceptual anticipation during sentence processing: Frontal positivity and N400 ERP components. *International Journal of Psychophysiology, 83*(3), 382–392. https://doi.org/10.1016/j.ijpsycho.2011.12.007

Van Berkum, J. J. A., Hagoort, P., & Brown, C. M. (1999). Semantic integration in sentences and discourse: Evidence from the N400. *Journal of Cognitive Neuroscience, 11*(6), 657–671. https://doi.org/10.1162/089892999563724

Van Berkum, J. J. A., Van den Brink, D., Tesink, C. M. J. Y., Kos, M., & Hagoort, P. (2008). The neural integration of speaker and message. *Journal of Cognitive Neuroscience, 20*(4), 580–591.

Van Berkum, J. J. A., Zwitserlood, P., Hagoort, P., & Brown, C. M. (2003). When and how do listeners relate a sentence to the wider discourse? Evidence from the N400 effect. *Cognitive Brain Research, 17*(3), 701–718.

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

Van Petten, C., & Luka, B. J. (2012). Prediction during language comprehension: Benefits, costs, and ERP components. *International Journal of Psychophysiology, 83*(2), 176–190. https://doi.org/10.1016/j.ijpsycho.2011.09.015

Vinckier, F., Dehaene, S., Jobert, A., Dubus, J. P., Sigman, M., & Cohen, L. (2007). Hierarchical coding of letter strings in the ventral stream: Dissecting the inner organization of the visual word-form system. *Neuron, 55*(1), 143–156. https://doi.org/10.1016/j.neuron.2007.05.031

Vissers, C. T., Chwilla, D. J., & Kolk, H. H. (2006). Monitoring in language perception: The effect of misspellings of words in highly constrained sentences. *Brain Research, 1106*(1), 150–163.

Wang, L., Brothers, T., Jensen, O., & Kuperberg, G. R. (2023). Dissociating the pre-activation of word meaning and form during sentence comprehension: Evidence from EEG representational similarity analysis. *Psychonomic Bulletin & Review*. https://doi.org/10.3758/s13423-023-02385-0

Wang, L., Kuperberg, G., & Jensen, O. (2018). Specific lexico-semantic predictions are associated with unique spatial and temporal patterns of neural activity. *Elife, 7*, Article e39061. https://doi.org/10.7554/eLife.39061

Wang, L., & Kuperberg, G. R. (2023). Better together: Integrating multivariate with univariate methods, and MEG with EEG to study language comprehension. *Language, Cognition and Neuroscience.*. https://doi.org/10.1080/23273798.2023.2223783

Wang, L., Nour Eddine, S., Brothers, T. A., Jensen, O., & Kuperberg, G. R. (2024). *Predictive Coding explains the dynamics of neural activity within the left ventromedial temporal lobe during reading comprehension*.

Wang, L., Schoot, L., Brothers, T., Alexander, E., Warnke, L., Kim, M., … Kuperberg, G. R. (2023). Predictive coding across the left fronto-temporal hierarchy during language comprehension. *Cerebral Cortex, 33*(8), 4478–4497. https://doi.org/10.1093/cercor/bhac356

Whittington, J. C. R., & Bogacz, R. (2019). Theories of error back-propagation in the brain. *Trends in Cognitive Sciences, 23*(3), 235–250. https://doi.org/10.1016/j.tics.2018.12.005

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. https://doi.org/10.1016/j.neuroimage.2012.04.054

Xiang, M., & Kuperberg, G. (2015). Reversing expectations during discourse comprehension. *Language, Cognition and Neuroscience, 30*(6), 648–672. https://doi.org/10.1080/23273798.2014.995679

Zorzi, M., Houghton, G., & Butterworth, B. (1998). Two routes or one in reading aloud? A connectionist dual-process model. *Journal of Experimental Psychology. Human Perception and Performance, 24*(4), 1131–1161.

Zwaan, R. A., & Radvansky, G. A. (1998). Situation models in language comprehension and memory. *Psychological Bulletin, 123*(2), 162–185. https://doi.org/10.1037/0033-2909.123.2.162