





# An implemented predictive coding model of lexico-semantic processing explains the dynamics of univariate and multivariate activity within the left ventromedial temporal lobe during reading comprehension

Lin Wang<sup>a,b,#</sup> , Samer Nour Eddine<sup>b,#</sup>, Trevor Brothers<sup>b,c</sup>, Ole Jensen<sup>d,e</sup>, Gina R. Kuperberg<sup>a,b,\*</sup> 

<sup>a</sup> Department of Psychiatry and the Athinoula A. Martinos Center for Biomedical Imaging, Massachusetts General Hospital, Harvard Medical School, Charlestown, MA, 02129, USA

<sup>b</sup> Department of Psychology, Tufts University, Medford, MA, 02155, USA

<sup>c</sup> North Carolina Agricultural and Technical State University, North Carolina, NC, 27411, USA

<sup>d</sup> Department of Experimental Psychology, University of Oxford, Oxford OX2 6GG, United Kingdom

<sup>e</sup> Oxford Centre for Human Brain Activity, Wellcome Centre for Integrative Neuroimaging, Department of Psychiatry, University of Oxford, Oxford OX3 7JX, United Kingdom

## ARTICLE INFO

### Keywords:

Predictive coding  
Language comprehension  
Modeling  
MEG  
N400  
RSA

## ABSTRACT

During language comprehension, the larger neural response to unexpected versus expected inputs is often taken as evidence for predictive coding—a specific computational architecture and optimization algorithm proposed to approximate probabilistic inference in the brain. However, other predictive processing frameworks can also account for this effect, leaving the unique claims of predictive *coding* untested. In this study, we used MEG to examine both univariate and multivariate neural activity in response to expected and unexpected inputs during word-by-word reading comprehension. We further simulated this activity using an implemented predictive coding model that infers the meaning of words from their orthographic form. Consistent with previous findings, the univariate analysis showed that, between 300 and 500 ms, unexpected words produced a larger evoked response than expected words within a left ventromedial temporal region that supports the mapping of orthographic word-forms onto lexical and conceptual representations. Our model explained this larger evoked response as the enhanced lexico-semantic prediction error produced when prior top-down predictions failed to suppress activity within lexical and semantic “error units”. Critically, our simulations showed that despite producing minimal prediction error, expected inputs nonetheless reinstated top-down predictions within the model’s lexical and semantic “state” units. Two types of multivariate analyses provided evidence for this functional distinction between state and error units within the ventromedial temporal region. First, *within* each trial, the same individual voxels that produced a larger response to unexpected inputs between 300 and 500 ms produced unique temporal patterns to expected inputs that resembled the patterns produced within a pre-activation time window. Second, *across* trials, and again within the same 300–500 ms time window and left ventromedial temporal region, pairs of expected words produced spatial patterns that were more similar to one another than the spatial patterns produced by pairs of expected and unexpected words, regardless of specific item. Together, these findings provide compelling evidence that the left ventromedial temporal lobe employs predictive coding to infer the meaning of incoming words from their orthographic form during reading comprehension.

## 1. Introduction

One of the most robust findings in the neurobiology of language

processing is that unexpected inputs produce a larger neural response than expected inputs. This effect has been described at multiple levels of linguistic representation and at multiple levels of the cortical hierarchy

\* Corresponding author.

E-mail address: [GKUPERBERG@mgh.harvard.edu](mailto:GKUPERBERG@mgh.harvard.edu) (G.R. Kuperberg).

# Equal contributors.

<https://doi.org/10.1016/j.neuroimage.2024.120977>

Received 20 September 2024; Received in revised form 28 November 2024; Accepted 16 December 2024

Available online 16 December 2024

1053-8119/© 2024 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY-NC license (<http://creativecommons.org/licenses/by-nc/4.0/>).

(Blank and Davis, 2016; Caucheteux et al., 2023; Price and Devlin, 2011; Sohoglu and Davis, 2020; Wang et al., 2023; Kutas and Hillyard, 1984; DeLong et al., 2005). It is often attributed to the generation of “prediction error”, and taken as evidence for a computational framework known as *predictive coding* (Blank and Davis, 2016; Bornkessel-Schlesewsky and Schlewsky, 2019; Caucheteux et al., 2023; Kuperberg et al. 2020; Price and Devlin, 2011; Rabovsky and McRae, 2014; Sohoglu and Davis, 2020; Wang et al., 2023; Xiang and Kuperberg, 2015). However, predictive coding is *not* the only framework that can explain predictive effects of context (Aitchison and Lengyel, 2017; Falandays et al. 2021; Luthra et al. 2021). Nor is it the only framework to posit the generation of “prediction error” (e.g. Fitz and Chang, 2019; Rabovsky et al. 2018).

For example, consider the well-known N400 — an evoked neural response observed at the scalp surface between 300 and 500 ms following word onset (Kutas and Federmeier, 2011). We have known since the 1980s that the amplitude of the N400 is smaller to incoming words that are predictable versus unpredictable during sentence comprehension (Kutas and Hillyard, 1984; DeLong et al., 2005). This finding cannot be explained without assuming that processing the prior context induces a change in state that implicitly anticipates/pre-activates future inputs (see Kuperberg and Jaeger, 2016, Introduction). Indeed, several studies have reported evidence of anticipatory neural activity in predictive contexts, even before new bottom-up input becomes available (e.g. DeLong et al., 2005; Wicha et al. 2004; Piai et al. 2015; Wang et al. 2018a, 2018b, 2024a; Leon-Cabrera et al. 2019; Grisoni et al. 2021). The reduced N400 response to expected inputs that confirm prior predictions has been interpreted as reflecting the reduced “neural effort” associated with “lexical access” (Lau et al. 2008), “integration” (Hagoort et al. 2004)<sup>1</sup> or semantic access/retrieval (Kutas and Federmeier, 2011; Van Berkum, 2009; Kuperberg 2016).

More recently, the intuition that the amplitude of the N400 reflects the influence of prior predictions on processing new bottom-up input has been formalized by several computational models that operationalize this evoked response as a “prediction error” (Rabovsky and McRae, 2014; Brouwer et al. 2017; Rabovsky et al. 2018; Fitz and Chang, 2019). In this context, the term “prediction error” does *not* imply that an input violates a strong prior prediction, or that it is a linguistic error (an anomaly). Rather, prediction error is defined simply as the *difference* between the pattern of activity predicted by the model, and a “target” pattern of activity. Although these different models operationalize “prediction” and “prediction error” in quite different ways (see Nour Eddine et al. 2022, for a comprehensive review), they all capture the intuition that the greater the match between the model’s predictions and the pattern of activity associated with the target, the smaller the magnitude of prediction error - the simulated N400. Crucially, however, *none* of these models actually implemented predictive coding.

So what is predictive coding and what distinguishes it from these other computational models of predictive processing?

<sup>1</sup> An early dichotomy drawn between “lexical access” and “integration” in relation to the N400 originated from models that drew sharp distinction between an initial process of “accessing” lexical representations before “integrating” them with the prior context. Within more interactive frameworks, these processes are closely intertwined (Kuperberg et al., 2020; Laszlo & Federmeier, 2011), leading 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). One advantage of computational models over these more qualitative frameworks is their ability to specify precise assumptions. A comprehensive review of computational models of the N400 is provided by Nour Eddine, Brothers, & Kuperberg (2022). Detailed discussions about the relationship between predictive coding and more classic accounts of the N400 are provided by Wang, Schoot, et al. (2023) and by Kuperberg et al. (2020).

### 1.1. Hierarchical predictive coding

Hierarchical predictive coding refers to a particular two-unit computational architecture and algorithm that was initially developed in the visual system to simulate extra-classical receptive field effects (Rao and Ballard, 1999; Spratling, 2012, 2013, 2014; see also Mumford, 1992; and see Walsh et al. 2020, for a recent review). It commits to a specific biologically plausible arrangement of feedforward and feedback connections that link successive layers of the cortical hierarchy (see Bastos et al., 2012; Shipp, 2016 for discussions about the biological plausibility of predictive coding). It also commits to a specific type of optimization algorithm that approximates Bayesian inference.

In hierarchical predictive coding, prior predictive contexts do not simply induce state changes that *implicitly* predict/pre-activate future inputs; rather, this anticipated information is propagated down the cortical hierarchy in a top-down fashion via feedback connections in attempts to *reconstruct* information encoded within “state units” at the level below.<sup>2</sup> As new bottom-up information becomes available, any information that matches these top-down predictions is simply reinstated within lower-level state units, while any information that cannot be explained by the top-down predictions activates lower-level “error units”, producing prediction error. Therefore, in contrast to the models of the N400 described above, where prediction error is computed externally by the modeler, the prediction error produced in predictive coding is computed within the model itself, and simply corresponds to the total activity produced by lower-level error units encoding information within incoming words that cannot be explained by higher-level representations of the prior context.

Crucially, and again in contrast with previous computational models, which conceptualized prediction error/the N400 as a byproduct of processing (Brouwer et al. 2017) or as a downstream signal computed purely for learning (Fitz and Chang, 2019; Rabovsky et al., 2018), the prediction error computed in predictive coding plays a direct functional role in inference — the process of inferring underlying causes from bottom-up inputs (it can also play a role in learning over longer time scales, see Rao and Ballard, 1997; Whittington and Bogacz, 2019). Specifically, any unpredicted information within the input (prediction error) is passed back up the hierarchy via feedforward connections,<sup>3</sup> where it is used to update the information encoded in higher-level “state units”, allowing them to generate more accurate predictions on the next iteration of the algorithm. Therefore, over multiple iterations of the predictive coding algorithm, the top-down predictions become more accurate, the magnitude of prediction error (the total activity produced by error units) is minimized, and state units at multiple levels of the hierarchy converge on the representations that best explain the bottom-up input.

<sup>2</sup> As we will discuss, some *qualitative* models of predictive language processing have similarly posited a top-down propagation of predicted information to pre-activate lexical representations at a lower level of the language hierarchy (e.g. DeLong et al., 2005; Federmeier, 2007; Lau et al., 2008; see Kuperberg & Jaeger, 2016 section 3 for discussion). This type of top-down predictive pre-activation could, in principle, be implemented by any architecture with feedback connections that allow for the propagation of information from higher to lower levels of the cortical hierarchy. What further distinguishes predictive coding is its commitment to functionally distinct state and error units that co-exist within these lower-level regions.

<sup>3</sup> Passing only the difference between observed and predicted data (prediction error) up the cortical hierarchy provides an efficient coding scheme that reduces information redundancy across the cortex. The term “predictive coding” was originally used to describe this type of efficient coding (Srinivasan, Laughlin, & Dubs, 1982), without implying the use of a specific algorithm for inference. However, *hierarchical* predictive coding, as first described by Rao and Ballard (1999), not only performs efficient encoding but also approximates Bayesian inference.

In recent work, we developed and implemented a predictive coding model of lexico-semantic processing (Nour Eddine et al. 2024). This model is based directly on biologically plausible models that were originally developed to explain low-level visual phenomena (Rao and Ballard, 1999; Spratling, 2013, 2014); that is, we imported the core structure of the predictive coding architecture, including its unique connectivity and its unique two-unit structure directly from these foundational models, changing only the model's internal representations. Our model also implements the steps of a particular predictive coding optimization algorithm that approximates Bayesian inference (Spratling, 2014, 2016). Top-down predictions are allowed to propagate down the hierarchy, pre-activating information at the semantic and lexical levels of representation, and the N400 is operationalized simply as the total amount of activity produced by error units at these levels (i.e. lexico-semantic prediction error) as the model converges to infer word meaning from orthographic inputs.

We showed that this model is able to simulate a range of contextual and lexical effects on the amplitude of the N400, including lexical predictability, priming, word frequency, concreteness, and orthographic neighborhood, as well as their interactions (Nour Eddine et al., 2024). We further showed that the dynamics of the predictive coding algorithm naturally explains the rise-and-fall time course of the N400: When unexpected inputs are first encountered, error units at the lexical and semantic levels are activated, producing an increase in lexico-semantic prediction error, mirroring the rise in the N400 amplitude. Then, as this prediction error is used to update higher-level states, the top-down predictions become more accurate and so the magnitude of the lexical and semantic prediction error (the total activity produced by error units at these levels) is suppressed, resulting in a fall of the simulated N400. In this way, predictive coding provides a direct functional link between the magnitude of prediction error and the time course of the evoked N400 response.

### 1.2. Neuroanatomical evidence for predictive coding within the left ventromedial temporal lobe

Neuroanatomical evidence in support of predictive coding comes from two recent MEG studies of reading comprehension (Wang et al. 2018b, 2023). In the first study (Wang et al., 2018a), we used multivariate methods to show that, in predictive contexts, specific pre-activated neural patterns uniquely encoded expected upcoming individual words, even before new bottom-up input was presented (see also Wang et al., 2024a). These patterns localized to regions of the left ventromedial temporal lobe, which are thought to encode orthographic forms (e.g., Dehaene and Cohen, 2011), lexical representations (Hirshorn et al., 2016; Woolnough et al., 2021), and conceptual representations of individual words (Lambon-Ralph et al. 2017). These findings therefore provided direct neural evidence for the top-down predictive pre-activation of lower-level lexical representations.

In a second MEG study, we found that, in plausible sentences, instead of localizing to higher levels of the language hierarchy (e.g. the left inferior frontal cortex) that are thought to encode contextual representations over a longer time-scale, the larger evoked response to unexpected (versus expected) inputs between 300 and 500 ms selectively localized to left-lateralized temporal regions that support the processing of individual words (Wang et al., 2023). This again included parts of the left ventromedial temporal lobe, which are known to support access to lexical representations that map orthographic form on to meaning during word reading (Hirshorn et al., 2016; Woolnough et al., 2021). These findings are therefore consistent with the claim that the N400

stems from prediction error produced at the lexico-semantic level.<sup>4</sup>

Together, these findings provide evidence that the left ventromedial temporal lobe implements predictive coding to infer word meaning from orthographic form during reading comprehension. However, they do not provide definitive evidence for this theory. In principle, any architecture with long-range feedback connections could allow implicitly predicted information to be propagated down from higher to lower levels of the cortical hierarchy. To show that the left ventromedial temporal lobe implements predictive coding, it is necessary to show that the same voxels that produce a larger neural response between 300 and 500 ms to *unexpected* inputs (by activating lexico-semantic error units) *also* reinstate pre-activated lexical and semantic representations (within state units) upon encountering *expected* inputs. It is also necessary to show that the process of reinstating these prior lexico-semantic predictions between 300 and 500 ms can be explained by the dynamics of the predictive coding algorithm.

### 1.3. The present study

In the present study, we tested these hypotheses by collecting MEG data as participants read strongly constraining sentences that ended either with expected or unexpected but plausible words (e.g., “In the crib, there is a sleeping baby/child”). We began by conducting a univariate analysis, which showed that the larger evoked response to unexpected (versus expected) words between 300 and 500 ms localized to the left ventromedial temporal lobe, replicating our previous findings (Wang et al., 2023). We then took the ventral and medial portions of this region as functional Regions of Interest (ROIs) and carried out two types of multivariate analyses, which were able to capture *patterns* of neural activity elicited by expected inputs between 300 and 500 ms, despite their small evoked response. To guide our interpretations, we carried out the same analyses on simulated activity extracted from the lexical and semantic layers of our predictive coding model.

**Hypothesis 1.** *Between 300–500 ms, expected inputs reinstate pre-activated lexico-semantic predictions within the same left ventromedial temporal region that produces a larger evoked response to unexpected inputs.*

As noted above, in previous work (Wang et al., 2018b), we provided evidence that in predictive contexts, the brain pre-activates specific lexical representations that encode expected upcoming words within the left ventromedial temporal lobe before new bottom-up input becomes available. In this previous study, this lexical pre-activation effect manifest as distinct item-specific *temporal patterns* of activity within an early pre-activation time window. We now reasoned that if upon confirming prior predictions, expected inputs reinstate these pre-activated temporal patterns, then, between 300 and 500 ms, expected inputs should produce temporal patterns that are more similar to the pre-activated patterns than those produced by unexpected inputs (i.e., a within-trial similarity effect: *expected* > *unexpected*; see Hubbard and Federmeier, 2021 for consistent EEG evidence on the scalp surface). Critically, if these predictions are reinstated within state units that co-exist with error units within the same ventromedial temporal region, this within-trial similarity effect should be detected in the same voxels of the left ventromedial temporal lobe that produce a larger overall univariate response to unexpected inputs.

We first carried out simulations using our predictive coding model to show that the reinstatement of prior top-down predictions within lexical and semantic state units indeed yielded a within-trial similarity effect (*expected* > *unexpected*). We then tested for this effect on the neural data

<sup>4</sup> Implausible/semantically anomalous words additionally produce an enhanced evoked response between 300-500 ms within the left inferior frontal cortex, which may reflect the generation of higher-level prediction error when an implausible interpretation cannot be explained by still higher-level predictions based on longer-term real-world knowledge (Wang et al., 2024).

after extracting the unique temporal patterns at each voxel within our left ventromedial ROIs for each sentence (a) within the early pre-activation time window identified in our previous study (Wang et al., 2018b), and (b) within the 300–500 ms time window following the onset of each expected and unexpected input.

**Hypothesis 2.** *The process of reinstating prior predictions to expected inputs within the left ventromedial temporal lobe activates state units that are functionally distinct from the error units activated by unexpected inputs.*

Next, we directly tested the hypothesis that, over the course of the predictive coding algorithm, regardless of the specific item activated, the process of converging on expected lexico-semantic representations within the left ventromedial temporal regions between 300 and 500 ms activates computational units (state units) that are functionally distinct from the units activated by unexpected inputs (both state and error units). Obviously, MEG does not have the spatial resolution to directly detect activity from individual error and state units. However, we reasoned that if the signal detected within each individual voxel within this left ventromedial temporal region reflects a random mixture of state and error activity, then despite producing a *smaller* overall univariate response between 300 and 500 ms, pairs of *expected* words should produce *spatial* patterns (across voxels) that are *more similar* to one another than pairs of expected and unexpected words (i.e., a cross-trial similarity effect: *within-expected* > *between-expected-unexpected*) (see de Gardelle et al. 2013a, 2013b, for evidence that this type of functional distinction can give rise to differences in fine-grained spatial patterns that can be detected using multivariate analyses).

We first verified that this was the case by carrying out simulations using our predictive coding model. After mixing and projecting the activity produced by state and error units at the lexical and semantic layers into a 20-voxel sampling space, we demonstrated that, regardless of the specific item activated, the differential activation of state versus error units by expected versus unexpected inputs indeed yielded a cross-trial similarity effect (*within-expected* > *between-expected-unexpected*).<sup>5</sup> We then tested for the same effect on the neural data by examining the spatial patterns across all voxels within our left ventromedial ROIs at each time point following the onset of each expected and unexpected input between 300 and 500ms.

## 2. Methods

In this Methods section, we first summarize the architecture and algorithm of the predictive coding model (see Nour Eddine et al., 2024). We then describe the MEG data collection and preprocessing. In the Results section, we describe the computational simulations themselves along with the univariate analyses and two types of multivariate analyses carried out on the simulated data and on the source-localized MEG data.

### 2.1. Predictive coding model

We built a hierarchical predictive coding model of lexico-semantic processing that infers concepts based on orthographic inputs. This model employs the same architectural principles and predictive coding algorithm used in previous predictive coding models that have simulated other perceptual and cognitive phenomena (Spratling, 2014, 2016). The model is described in detail by Nour Eddine et al. (2024),

<sup>5</sup> Importantly, we used a metric of spatial similarity – Pearson’s  $r$  – that is mathematically independent of the magnitude of response. Therefore, any greater similarity amongst pairs of expected items could not trivially be explained by the univariate effect (expected items produced less overall activity than unexpected items) (cf. Guggenmos, Sterzer, & Cichy, 2018; Haxby et al., 2001; Jimura & Poldrack, 2012; Walther et al., 2016; for discussion see Wang & Kuperberg, 2024, pages 10-12).

where we showed that the magnitude of lexico-semantic prediction error produced as the model converged mirrors the functional sensitivity of the N400 to various lexical variables, priming, contextual effects, as well as their higher-order interactions (see <https://github.com/samer-nour-eddine/REPN400>). The model is briefly described below.

#### 2.1.1. Architecture

The architecture, shown in Fig. 1A, is comprised of four hierarchically-organized layers — three levels of linguistic representation (orthographic, lexical, semantic) and a conceptual layer at the top, which encodes the probabilities of 1579 upcoming concepts, each corresponding to one lexical item. This highest layer was used to provide top-down predictive pre-activation in our simulations (see below). The lowest *orthographic* layer, which was used to provide bottom-up inputs in our simulations, encodes one of 26 letter identities (A-Z) at each of four possible spatial positions. The middle *lexical* level encodes 1579 four-letter words in the model’s lexicon (e.g., *baby*, *lime*). The third *semantic* level encodes 12,929 unique semantic features (e.g.  $\langle \text{small} \rangle$ ,  $\langle \text{cry} \rangle$ ,  $\langle \text{cute} \rangle$ ).

Similar to the classic Interactive Activation and Competition (IAC) (Chen and Mirman, 2012; McClelland and Rumelhart, 1981) and TRACE (McClelland and Elman, 1986), rather than training the model, we incorporated linguistic representations at each level of the hierarchy, and hand-coded the connection weights that described the mappings between successive levels as weight matrices,  $V$  and  $W$ . For example, an orthographic-lexical matrix specified the mappings between the positions of individual letters (e.g. ‘B’, ‘A’, ‘B’, ‘Y’) and individual lexical items (e.g. ‘*baby*’), while a lexical-semantic matrix specifies mappings between each lexical unit (e.g. ‘*baby*’) and a specific set of semantic features (e.g.  $\langle \text{small} \rangle$ ,  $\langle \text{cry} \rangle$ ,  $\langle \text{cute} \rangle$ , etc.). Note that it is also possible for the model to learn these parameters without any modification to the architecture.

Consistent with predictive coding principles, each of the three linguistic levels of representations incorporates two types of computational units — state units, which encode the internal representations being inferred, and error units, which encode the residual difference between the information encoded within state units and the top-down predictions (otherwise referred to as reconstructions) from the level above. In addition, as for all predictive coding architectures, at each linguistic level of the hierarchy, the state and error units share one-to-one connections. Across linguistic levels, the state units communicate with error units via many-to-many connections.

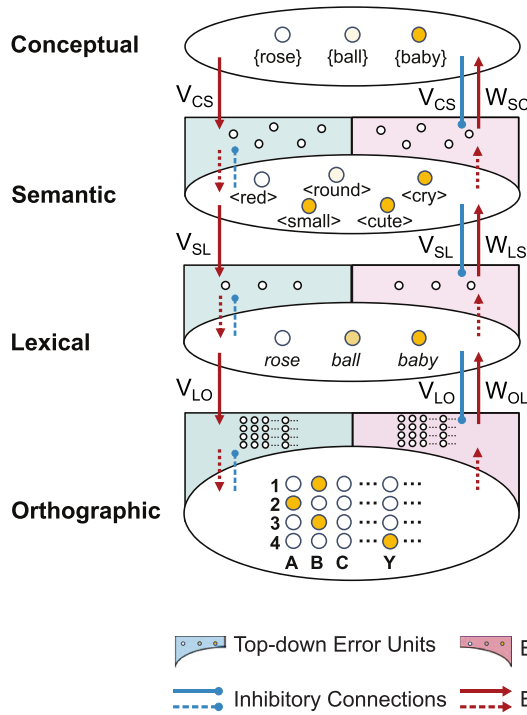
#### 2.1.2. Algorithm

Predictive coding implements an optimization algorithm that approximates Bayesian inference. Our model implements the Predictive Coding/Biased Competition-Divisive Input Modulation algorithm (PC/BC-DIM) (Spratling, 2008, 2016), which computes prediction errors using division, as opposed to the subtraction method used in the original predictive coding model by Rao and Ballard (1999). This allows for fast convergence of the algorithm and ensures that the activity of all units remains non-negative, similar to how biological neurons function.

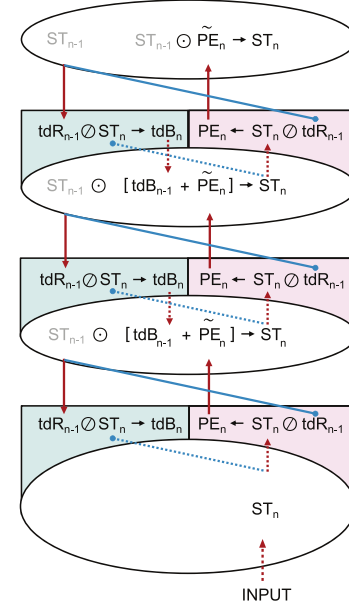
At each level of representation, the activity within state units can be thought of as a changing target pattern that state units at the level above try to predict or reconstruct. As shown in Fig. 1B, on each iteration of the algorithm, the state units at the level above generate a top-down prediction/reconstruction of the target pattern at the lower level. Error units at the lower level then calculate the residual difference between this top-down reconstruction and the target state pattern. We incorporated two types of error units (Spratling, 2016): (a) “bottom-up error units”, which computed the residual information in the state units that was not encoded in the top-down reconstructions from the level above (known as “prediction error”), and (b) “top-down error units”, which compute the residual information in the top-down reconstructions that was not present in state units at that level (termed “top-down bias”). Bottom-up prediction error and top-down bias are both calculated by



## A. Model Architecture



## B. Model Algorithm



**Fig. 1. Predictive coding model. A. Model Architecture.** The model includes state units and error units at three levels of linguistic representation (Orthographic, Lexical and Semantic) and an additional highest Conceptual layer that only contains state units. **Key elements:** (1) State units: Represented as small circles within large ovals at each level of linguistic representation and at the conceptual layer. (2) Error units: Represented as small circles within half arcs at each linguistic level. Bottom-up error units (shown in pink) compute the residual information in the state units that was not encoded in the top-down reconstructions from the level above (known as “prediction error”); Top-down error units (shown in green) compute the residual information in the top-down reconstructions that was not present in state units at the same level (termed “top-down bias”). **Connections:** (1) Blue solid arrows: Inhibitory feedback connections from higher-level state units to lower-level bottom-up error units, which allow any top-down reconstructions that explain/match activity within lower-level state units to suppress lower-level prediction error; (2) Red solid arrows: Excitatory feedforward connections, from lower-level bottom-up error units to higher-level state units, which pass the prediction error forward to update the higher level state units, so that these state units produce better top-down reconstructions on the next iteration; (3) Blue dotted arrows: Inhibitory within-level connections from state units to top-down error units; (4) Red dotted arrows: Excitatory within-level connections from state units to bottom-up error units, or from top-down error units to state units.  $V_{LO}/W_{OL}$  represent connections between the lexical and orthographic level;  $V_{SL}/W_{LS}$  represent connections between the semantic and lexical level; and  $V_{CS}/W_{SC}$  represent connections between conceptual and semantic level. **Activity Patterns:** We schematically depict the model’s activity patterns after settling on the lexical representation of the item, /baby/. Different shades of yellow indicate the strength of activity in each state unit. **Orthographic level:** four state units are activated: corresponding to the letters B, A, B and Y in their respective position; **Lexical level:** the unit for /baby/ is strongly activated, and the unit for *ball* is partly activated due to shared letters with *baby*; **Semantic level:** units corresponding to the semantic features of *baby* (<cute>, etc.) show varying levels of activation; **Conceptual level:** the unit for the representation of /baby/ is strongly activated. Across all levels, activity within error units is minimal because the model has settled. **B. Model Algorithm.** Schematic illustration of the predictive coding algorithm operating on the  $n^{\text{th}}$  iteration, following the presentation of bottom-up orthographic input. Each variable’s subscript indicates the iteration on which it was computed. The same three steps occur in sequence at each level of representation: (1) **State units update:** State units are updated based on the top-down bias from the previous iteration and the prediction error from the level below on the current iteration:  $ST_n = ST_{n-1} \odot [tdB_{n-1} + \tilde{PE}_n]$ . Values are copied to the top-down and bottom-up error units at the same level. (2) **Prediction error and top-down bias computation:** Prediction error is calculated via elementwise division:  $PE_n = ST_n \oslash tdR_{n-1}$ . This prediction error is passed up to state units at the level above by transforming its dimensionality:  $\tilde{PE}_n = W \cdot PE_n$ ; top-down bias is also calculated via elementwise division:  $tdB_n = tdR_{n-1} \oslash ST_n$ ; and top-down bias is copied to state units at the same level for the next iteration. (3) **Top-down reconstruction computation:** State units generate top-down reconstructions of activity at the level below via linear transformation by the V (generative) matrix,  $tdR_n = V \cdot ST_n$ ; and these reconstructions are passed down to the error units at the level below. Linear transformations of variables were implemented through the hand-coded weight matrices W (feedforward) and V (feedback) described above.

element-wise division (i.e., Prediction Error = State  $\oslash$  Reconstruction; Top-down Bias = Reconstruction  $\oslash$  State). The prediction error is passed up to update state units at the level above, allowing them to generate more accurate top-down predictions on the next iteration of the algorithm. In contrast, the top-down bias modifies the target state pattern at the same level, bringing it closer to the prediction from the level above. Thus, at each iteration of the algorithm, the state at each level of the hierarchy is modified in two ways: one that helps it better predict its lower-level target pattern (driven by prediction error), and another that helps it serve as a better target for a higher-level state (driven by top-down bias). Over multiple iterations, the magnitude of prediction error and top-down bias decrease, and the model reaches a global state that can accurately explain the bottom-up input at multiple levels of

representation.

## 2.2. MEG study

### 2.2.1. Design and stimuli

We developed a set of 240 highly constraining Chinese sentence contexts. Each context was paired with either an expected or an unexpected but plausible critical word (e.g. “In the crib, there is a sleeping baby/child”, see Wang et al., 2018b, for a detailed description). The expected and unexpected words were matched on frequency, extracted from the SUBTLEX-CH database (Cai and Brysbaert, 2010) (mean expected  $\pm$  SD:  $55 \pm 130$  vs. unexpected:  $27 \pm 97$ ,  $t_{(96)} = 1.73$ ,  $p = .09$ ), and on visual complexity, which was operationalized by aggregating the

number of strokes of all characters of each word (*expected*:  $17 \pm 5$  vs. *unexpected*:  $17 \pm 4$ ,  $t_{(239)} = 0.56$ ,  $p = .58$ ).

In a cloze norming study, 30 participants, who did not take part in the MEG study, were presented with the contexts and asked to produce the most likely next word. The expected words had a cloze probability of 88 % (SD: 12 %). Unexpected words were not produced by any of the participants in the cloze norming tests, and therefore had a cloze probability of zero.

The stimuli were divided into two lists, each containing 240 sentences. All sentence contexts appeared in each list, with half of the sentences ending with expected words and the other half ending with unexpected words. Within each list, the sentences were pseudo-randomized so that participants did not encounter more than three expected or unexpected critical words in succession.

### 2.2.2. Participants

The study was approved by the Institutional Review Board (IRB) of the Institute of Psychology, Chinese Academy of Sciences, and all participants signed a written consent form and were paid for their time. Initially, 34 native Chinese speakers participated but the data of eight participants were subsequently excluded because of technical problems, leaving a final MEG dataset of 26 participants (mean age 23 years, range 20 – 29; 13 males). All participants were right-handed, had normal or corrected-to-normal vision, and had no history of language or neurological impairments.

### 2.2.3. Experimental procedure

MEG data were collected while participants sat comfortably in a dimly-lit shielded room. Each participant read 240 sentences, which were presented on a projection screen, word by word (gray font on a black background). Each trial began with a blank screen (1600 ms), and each word was presented with a long Stimulus Onset Asynchrony of 1000 ms (200 ms presentation with an inter-stimulus interval of 800 ms). The final word of each sentence was presented together with a period, followed by an inter-trial interval of 2000ms. Participants were asked to read the sentences for comprehension. To encourage comprehension, following 1/6th of the trials (at random), participants read a statement that referred back to the semantic content of the sentence that they had just read, and pressed one of two buttons with their left hand depending on whether they judged it to be true or false. Following the remainder of the trials, the Chinese word “继续” (meaning “NEXT”) appeared and participants simply pressed another button with their left hand within 5000 ms in order to progress to the next trial.

The 240 sentences were divided into eight blocks, each lasting about eight minutes. Between each block, participants were told that they could relax and blink, but to try to keep the position of their heads still. They then indicated verbally to the experimenter when they were ready for the next block. The experiment lasted for about 1.5 h, including preparation, instructions, and a short practice session consisting of eight sentences.

### 2.2.4. MEG data acquisition

The MEG dataset was collected using a CTF Omega System with 275 axial gradiometers at the Institute of Biophysics, Chinese Academy of Sciences. Six sensors (MLF31, MRC41, MRF32, MRF56, MRT16, MRF24) were excluded from the data recordings because they were non-functional. The ongoing MEG signals were low-pass filtered at 300 Hz and digitized at 1200Hz. Head position with respect to the sensor array was monitored continuously with three fiducial coils placed at the forehead, and the left and right cheekbones. In addition, structural Magnetic Resonance Images (MRIs) were obtained from 25 participants using a 3.0T Siemens system. In order to facilitate alignment between these MRIs and the MEG coordinate system for source-level analysis, three markers were attached in the same position as the fiducial coils.

### 2.2.5. MEG pre-processing

The MEG data were analyzed using the Fieldtrip software package, an open-source MATLAB toolbox (Oostenveld et al. 2011). To minimize environmental noise, we applied third-order synthetic gradiometer correction during preprocessing. The MEG data were then segmented into 4000 ms epochs, time-locked from –2000 ms until 2000 ms after the onset of each critical word. Within each 4000 ms epoch, trials contaminated with muscle or MEG jump artifacts were identified and removed using a semi-automatic procedure. We then carried out an Independent Component Analysis (ICA; Bell and Sejnowski, 1997; Jung et al., 2000) and removed components associated with the eye movement and cardiac activity from the MEG signal (mean: 5.8, range: 3–8, std: 1.5). We also inspected the data visually and removed any remaining artifacts. On average, 96 % of trials were retained (115 trials in each of the two conditions). Finally, we applied a 30 Hz low pass filter to the artifact-free MEG data and applied a baseline correlation by subtracting the mean amplitude between –200 and 0 ms relative to the onset of the critical word from each trial.

We next projected the time series data collected from the MEG sensors into the source space using a beamforming approach (Van Veen et al. 1997). Participant-specific spatial filters were computed using the Linearly Constrained Minimum Variance (LCMV) method (Van Veen et al., 1997), based on a lead field matrix and the covariance matrix of the sensor-level data.

To obtain the lead field matrix, we used the fiducials to spatially co-register the individual anatomical MRIs to the MEG sensor array, and then created a single-shell head model based on the segmented MRI images (Nolte, 2003). After that, we divided the brain volume into voxels using a three-dimensional grid with 10 mm spacing, and then mapped this grid on to the Montreal Neurological Institute (MNI) brain template (Montreal, Quebec, Canada). In one participant whose MRI images were not available, we used the MNI template brain. To calculate the covariance matrix, we used data from the axial gradiometers from –1000 ms to 1000 ms relative to the word onset. We specified each participant’s spatial filter to have a fixed orientation, resulting in one spatial filter at each grid point. We then applied these spatial filters to the sensor-level data to estimate the source activity at each grid point.

## 3. Results

### 3.1. Univariate analysis: unexpected versus expected

#### 3.1.1. Simulations

In previous work, we used an implemented predictive coding model of lexico-semantic processing to simulate the larger N400 response to unexpected versus expected inputs as lexico-semantic prediction error (Nour Eddine et al., 2024) — the total activity produced by *error units* at the lexical and semantic layers of the model. We now carried out similar simulations to verify that this larger univariate response can still be explained if, instead of only summing across error units, we summed across *all* units at the lexical and semantic layers (i.e., *both* error and state units).

We simulated 120 expected and unexpected trials. For each trial, we first provided the model with 20 iterations of top-down predictive pre-activation of an expected item by clamping this item at the highest conceptual layer with 88 % of the total activation — the average constraint of the contexts in our experimental stimuli (all other items were clamped with uniform activation). Then, after unclamping this conceptual layer, we clamped the lowest orthographic layer with either an *expected* four-letter input that matched the pre-activated item, or an *unexpected* four-letter input that mismatched the pre-activated item, and we let the model run for 20 more iterations. We then summed activity across all lexical and semantic error and state units at each iteration, and subtracted the summed activity produced by expected inputs from that produced by unexpected inputs to compute a time course of the simulated univariate effect (unexpected *minus* expected).

As shown in Fig. 2A, the simulated univariate effect revealed a rise-and-fall waveform-like morphology, analogous to the N400 effect, rising rapidly and peaking at around iteration 4, and then falling again by around iteration 15. Averaged across iterations 2 and 11, the effect was significant across all simulated trials,  $t_{(119)} = 86.10$ ,  $p < .001$ .

To verify that this effect was mainly driven by error unit activity, we extracted the activity produced by the error and state units separately (each summed across the lexical and semantic levels). As shown in Fig. 2B, the activity produced by the error units (i.e. lexico-semantic prediction error, shown in red) was much larger to unexpected than expected inputs. In contrast, the activity produced by the state units (shown in blue) was minimal to both the expected and unexpected input. (As we describe in the next section, despite producing such a small magnitude response, these state units nonetheless converged on the lexico-semantic representations that encoded the bottom-up input). These findings therefore confirm that the univariate effect (*unexpected* > *expected*) was driven by lexico-semantic prediction error produced by error units, with minimal contribution from state units.

### 3.1.2. MEG

We carried out the univariate analysis on 174 voxels within a left fronto-temporal search volume, which we specified in the source space based on the Brainnetome atlas (Fan et al., 2016), see Supplementary Figure S1. In each participant, at each of the 174 voxels within this left fronto-temporal search volume, and at each sampling time point (from 200 ms before to 1000 ms following critical word onset), we separately averaged the source-localized MEG evoked response to the unexpected and expected words. We then computed their difference (*unexpected* minus *expected*) and averaged these difference values between 300 and 500ms. To determine where the effect was statistically significant, we carried out paired *t*-tests at each of the 174 voxels, and used a cluster-based permutation approach to account for multiple comparisons across voxels (Maris and Oostenveld, 2007). Specifically, adjacent voxels with *p*-values that exceeded a pre-set uncorrected value of 0.05 or less were considered spatial clusters, and the sum of the *t*-values within each spatial cluster was taken as the cluster mass statistic. We then created a null distribution by repeating this procedure 1000 times, but randomly shuffling the condition labels at each voxel and taking the largest cluster mass statistic in each randomization. If the cluster mass statistic of any observed clusters fell within the highest or lowest 2.5 % of the null distribution, we considered the effect to be significant.

Confirming our previous findings in plausible sentences (Wang et al., 2023), this analysis showed that the larger univariate response to unexpected versus expected words selectively localized to the regions of the left temporal lobe that are known to support lexico-semantic processing. As shown in Fig. 3, a cluster-based permutation test across the left fronto-temporal search volume revealed a significant cluster within an anterior-mid portion of the left ventromedial temporal lobe (*unexpected* > *expected*),  $p = .002$ . This cluster spanned two neuroanatomical subregions: (a) a left ventral temporal region (including the anterior inferior temporal and anterior-mid fusiform cortex, 15 voxels), which has been implicated in lexical processing, i.e. mapping orthographic forms on to sets of semantic features (Hirshorn et al., 2016; Woolnough et al., 2021), and (b) a left medial temporal region (including the parahippocampal gyrus and hippocampus, 5 voxels), which has been implicated in domain-general conceptual processing — mapping distributed semantic features on to unique concepts (Lambon-Ralph et al., 2017; Patterson et al. 2007).

We also observed a significant evoked effect (*unexpected* > *expected*) in the mid-portion of the left superior temporal cortex. This also replicates our previous findings (Wang et al., 2023), and may have reflected activity within a pathway that maps indirectly from orthography to meaning through phonology (Grainger and Holcomb, 2009; Harm and Seidenberg, 2004), see Supplementary Materials, Section 2.

### 3.2. Within-trial similarity analysis: reinstatement of prior item-specific predictions within each trial

A key claim of predictive coding is that, despite producing a smaller neural response between 300 and 500 ms, expected inputs should nonetheless reinstate the same lexico-semantic predictions that were pre-activated before this bottom-up input became available, and that this reinstatement effect should localize to the same region and 300–500 ms time window in which the univariate effect is observed. The aim of this within-trial similarity analysis was to test this hypothesis.

#### 3.2.1. Simulations

We began by verifying that our model did indeed reinstate prior top-down lexico-semantic predictions within individual trials. In Fig. 4A, left, we show in blue, the time course of pre-activated activity (from iterations  $-20$  to  $0$ ), extracted from the single lexical state unit that encoded the word, /poem/, after pre-activating the concept, {poem}, at the model's highest layer and allowing activity to propagate down the model's hierarchy. In Fig. 4A, right, also in blue, we show the post-activation time course (iterations  $0$  to  $20$ ) of the same word's lexical state unit, /poem/, after presenting the model with the expected orthographic input, "P-O-E-M". Immediately after the onset of the expected orthographic input, the expected lexical state unit, /poem/, rapidly reached a plateau between 2 and 11 iterations — the same iteration window that we used to quantify the magnitude of the N400 effect in the previous univariate simulation (see Fig. 2). Also in Fig. 4A, right, in red, we show the post-activation time course of a single unexpected word's lexical state unit, /soot/, after presenting the model with an unexpected orthographic input, "S-O-O-T". This lexical state unit also accumulated activity but at a slower rate, plateauing well after the 2–11 iteration window used to operationalize the N400 effect.

To confirm that the model reinstated the correct expected lexical item in all 120 simulated trials within this 2–11 post-activation iteration window, for each simulated trial, we extracted the most active lexical state unit within this 2–11 iteration window: In simulated expected trials, this most-active lexical state unit (e.g. /poem/) corresponded to both the top-down conceptual prediction (e.g. {poem}) as well as to the expected orthographic input ("P-O-E-M"). In contrast, for all unexpected trials, the most active lexical state unit (e.g. /soot/) did not correspond to the pre-activated concept ({poem}), but did correspond to the bottom-up orthographic input ("S-O-O-T").

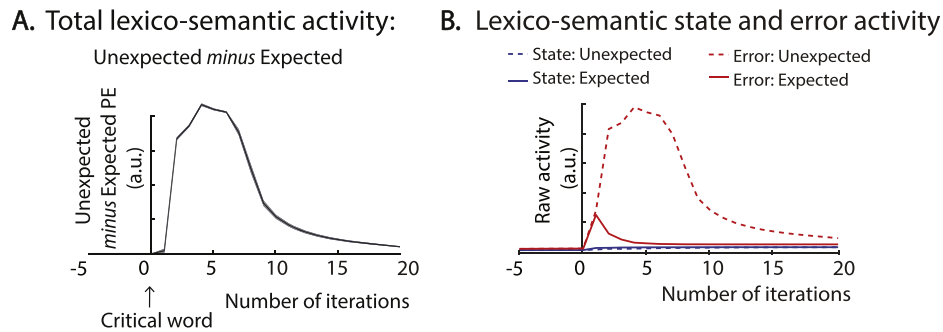
Finally, to show that only expected, and not unexpected inputs, reinstated prior predictions across both the lexical and semantic layers within the 2–11 iteration window of interest, we carried out a within-trial similarity analysis that was intended to mirror the analysis used to test for a prediction reinstatement effect on the neural data (see below). For each of the 120 simulated trials, we extracted the activity produced across all lexical and semantic state units (a) during the top-down pre-activation phase (from iterations  $-11$  to  $-2$ ), and (b) during the bottom-up activation phase (from iterations  $2$  to  $11$ ) after presenting the model with expected and unexpected inputs. Within each trial, we computed a Pearson's *r* value to quantify the similarity between the pre-activated patterns and the post-activation patterns produced by expected and unexpected inputs separately. As shown in Fig. 4B, the lexico-semantic state patterns produced by the expected inputs were more similar to the pre-activated state patterns than the lexico-semantic state patterns produced by the unexpected inputs (*expected*:  $0.64 \pm 0.02$  vs. *unexpected*:  $0.09 \pm 0.07$ ,  $t_{(119)} = 86.78$ ,  $p < .001$ ).

#### 3.2.2. MEG

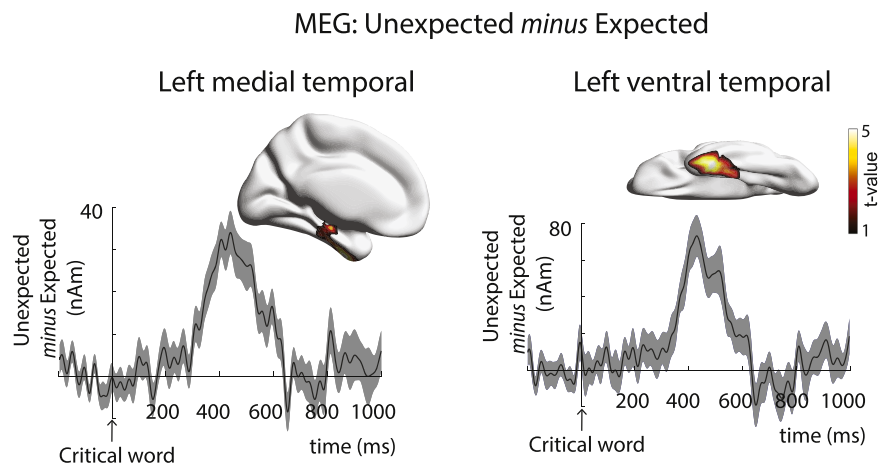
We then carried out a similar within-trial similarity analysis on the MEG data (Fig. 5A) to test the hypothesis that, between 300 and 500 ms, the same voxels within the left medial and ventral temporal regions that produced a larger univariate response to unexpected inputs would reinstate prior predictions to expected inputs.

In previous work (Wang et al., 2018b), we found evidence that, in

## Simulated univariate effect: Unexpected vs Expected



**Fig. 2. Simulated univariate N400 effect.** **A.** Time course of the simulated N400 effect. The difference in total lexico-semantic activity produced by unexpected and expected inputs (*Unexpected minus Expected*) at each iteration of the predictive coding algorithm post-target onset. At each iteration, the total lexico-semantic activity was computed by summing activity across all state and error units within the lexical and semantic layers of the model. **B.** Activity produced by lexico-semantic error and state units in response to *expected* and *unexpected* inputs. Note that the magnitude of activity produced by error units was far greater than that produced by state units on all iterations of the algorithm.



**Fig. 3. MEG univariate N400 effects (*Unexpected minus Expected*).** A significantly larger N400 response was evoked by unexpected compared to expected words between 300 and 500 ms post-stimulus onset within the anterior-mid ventromedial temporal lobe. The t-values are visualized on the MNI template brain using BrainNet Viewer (Xia et al. 2013). The time courses of the N400 effect (*Unexpected > Expected*) are extracted by averaging separately across voxels within the left medial and ventral portions of this cluster. Gray shading indicates standard errors.

predictive contexts, the generation of top-down item-specific lexico-semantic predictions manifested as unique fine-grained temporal patterns within the left ventromedial temporal lobe, across a 400 ms pre-activation time-window  $-900$  to  $-500$  ms, relative to the onset of the predicted upcoming word (i.e. 100–500 ms following the onset of the pre-critical word). In the current study, we reasoned that if either of our ROIs — the left anterior-mid medial temporal region (5 voxels) or the left anterior-mid ventral temporal region (10 voxels) — reinstate prior lexico-semantic predictions between 300 and 500 ms, then expected words should produce temporal patterns that are more similar to these pre-activated patterns than the temporal patterns produced by unexpected words.

To test this hypothesis, in each participant, for each trial, and at each voxel within each of these two ROIs, we extracted a vector that represented the fine-grained temporal pattern of neural activity produced by each expected and unexpected critical word between 300 and 500 ms post-target onset. We subdivided the 400 ms pre-activation time window identified in our previous study (Wang et al., 2018b) into three sliding 200 ms pre-activation time windows, and extracted a vector that represented fine-grained temporal pattern produced at each voxel within each of these pre-activation time windows:  $-900$  to  $-700$  ms,  $-800$  to  $-600$  ms,  $-700$  to  $-500$  ms prior to the critical word (or, equivalently, 100–300 ms, 200–400 ms, and 300–500 ms following the onset of the

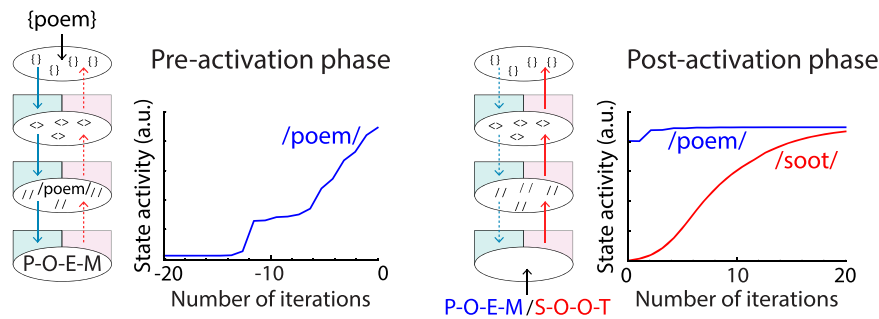
pre-critical word). Then, for each trial, at each voxel, we tested for a within-trial similarity effect by correlating (using Pearson's  $r$ ) the pre-activated patterns within each 200 ms sliding time window and the post-target patterns (300–500 ms) for expected and unexpected words separately (see Fig. 5A for a schematic depiction). We found that within the left anterior-mid medial temporal region of interest, the fine-grained temporal patterns produced by expected inputs between 300 and 500 ms were significantly more similar to the pre-activated patterns produced between  $-900$  ms and  $-700$  ms, than the patterns produced by unexpected inputs ( $t_{(25)} = 2.57$ ,  $p = .024$ ; Paired  $t$ -test with Bonferroni correction across the three pre-activation time windows), see Fig. 5B. Within the left anterior-mid ventral temporal region of interest, the effect was marginally significant ( $t_{(25)} = 1.97$ ,  $p = .08$ ).

A more exploratory within-trial similarity analysis across the remaining 159 voxels within the left fronto-temporal search volume showed that the prediction reinstatement effect extended to a more posterior portion of the left fusiform and medial temporal lobe,  $p = .012$  (cluster-corrected across all voxels, and Bonferroni corrected across the three pre-activated time windows).

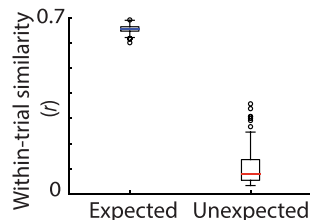
Finally, to exclude the possibility that any within-trial similarity effect was driven by differences in similarity between the lexical properties of the pre-critical words and the presented expected versus unexpected critical words, we calculated the absolute differences in



### A. Top-down prediction and prediction reinstatement of a single lexical state unit



### B. Simulated within-trial similarity effect



**Fig. 4. Top-down prediction and prediction reinstatement of a single lexical state unit and simulated within-trial similarity effect.** **A.** Time course of item-specific activity of example lexical state units during the top-down pre-activation (iterations  $-20$  to  $0$ ) and bottom-up post-activation (iterations  $0$  to  $20$ ) phases. We presented the model's conceptual layer with the top-down prediction, {poem}, and allowed activity to propagate down the model's hierarchy for 20 iterations. **Left:** During the pre-activation phase (iterations  $-20$  to  $0$ ), top-down predictions increased state activity in the lexical unit corresponding to the expected item /poem/. **Right:** During the post-activation phase, we presented an expected ('P-O-E-M') or an unexpected ('S-O-O-T') orthographic input for 20 iterations. Following the onset of the expected orthographic input ('P-O-E-M'; shown in blue), the corresponding lexical state unit rapidly accumulated activity, reaching a plateau within the same iteration window (2–11) in which we observed a larger univariate response to unexpected than expected inputs. In contrast, after the onset of the unexpected orthographic input ('S-O-O-T'; shown in red), activity accumulated within the corresponding lexical state unit at a slower rate, plateauing after the 2–11 iteration window that we used to operationalize the univariate effect. Note that in both plots, the y-axis represents arbitrary units, but the raw activation values in the pre-activation window are actually far smaller than those in the post-activation window. Despite these low absolute pre-activation values, they still offer a powerful head start during the post-activation phase because prediction error acts multiplicatively to update states on each iteration of the algorithm. **B.** Simulated within-trial similarity effect: *Expected* > *Unexpected*. Box plots showing that the similarity between post-stimulus state activity to expected inputs and pre-activated state activity was greater (larger average  $r$  values), than the similarity between post-stimulus state activity to unexpected inputs and pre-activated state activity (smaller average  $r$  values).

frequency and visual complexity between the pre-critical word and the presented expected and unexpected critical words in each sentence. We then computed the differences in these mean values and compared these differences against a null distribution generated by randomly shuffling the condition labels. These non-parametric tests failed to reveal any differences in similarity between the expected and unexpected conditions, either in frequency (*expected*:  $47 \pm 117$  vs. *unexpected*:  $25 \pm 92$ ,  $p = .13$ ) or visual complexity (*expected*:  $7 \pm 6$  vs. *unexpected*:  $7 \pm 6$ ,  $p = .74$ ).

#### 3.3. Cross-trial similarity analysis: Reinstatement of prior predictions within state units that are functionally distinct from the error units activated by unexpected inputs

The within-trial similarity analysis described above focused on the reinstatement of item-specific lexico-semantic predictions *within* each individual trial by examining the fine-grained *temporal patterns* produced by expected inputs in each of the ROIs defined by the univariate analysis between 300 and 500 ms. We next carried out a cross-trial similarity analysis to ask whether, regardless of specific item, the process of reinstating expected lexico-semantic representations within our two ROIs would activate state units that are functionally distinct from the error units activated by unexpected inputs. For this analysis, we focused on the fine-grained *spatial patterns* produced across all voxels within each of the ROIs across the 300–500 ms time window.

##### 3.3.1. Simulations

We began by carrying out simulations to ask how the differential

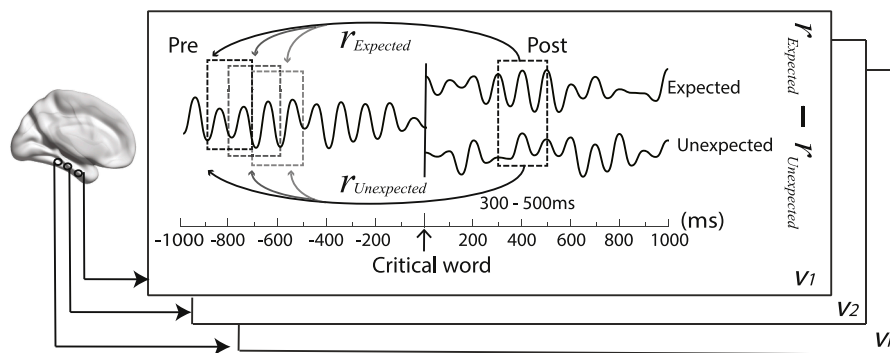
activation of state and error units by expected and unexpected inputs influences the similarity amongst fine-grained spatial patterns of activity.

In Fig. 6A, we show the time course of state and error activity to expected and unexpected inputs. These values were extracted from the lexical and semantic layers of the model, and normalized by their respective maximal values. After presenting the model with unexpected inputs, activity within error units rises rapidly, peaking at around iteration 7, whereas activity within state units rises more slowly. Activity within error units then falls again and is minimal by around iteration 15, whereas activity within state units continues to rise, beginning to plateau at around iteration 15. The responses to expected inputs are quite different: After presenting the model with expected inputs, activity is dominated by state units at both iterations 7 and 15, with minimal contribution from error units.

We reasoned that if the signal detected within each voxel reflects a random mixture of state and error activity, then as a consequence of the larger proportion of state (*versus* error) units activated by expected (*versus* unexpected) inputs at around iteration 7, pairs of expected inputs should produce spatial patterns that are more similar to one another than pairs of expected and unexpected inputs, despite producing minimal error activity. To verify this intuition, we carried out simulations to examine similarity amongst patterns of activity extracted from the predictive coding model after projecting them into a 20 voxel "sampling space" (Fig. 6B).

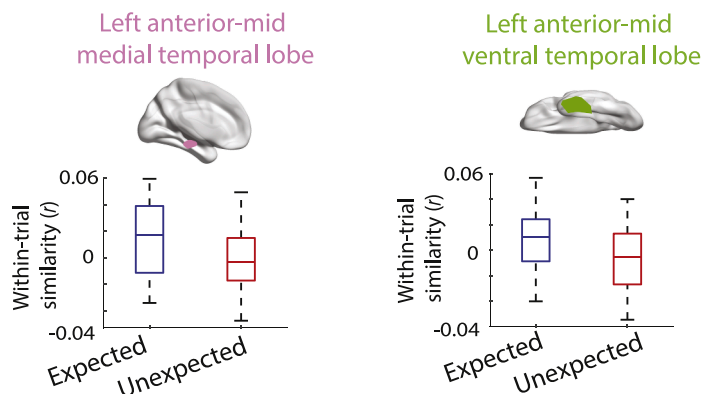
Specifically, for each of our simulated trials, on each iteration of the algorithm, we mixed and projected the summed lexical and semantic state and error activity (a  $2 \times 1$  state-error vector [*st er*]) into the

- A. (1) For each voxel, extract temporal patterns within pre-activation and post-activation time windows, and compute within-trial similarity,  $r$



- (2) Repeat at each voxel

- B. Within-trial similarity effect within regions of interest



**Fig. 5. MEG within-trial similarity analysis: Prediction reinstatement effect.** A. Schematic illustration of the MEG within-trial similarity analysis. (1) At each voxel,  $v$ , within the region of interest, we extracted a vector that represented the fine-grained temporal pattern produced between 300 and 500 ms after the onset of each expected and unexpected critical word. At the same voxel, we extracted vectors that represented the fine-grained temporal patterns produced within three overlapping 200 ms time windows prior to critical word onset ( $-900$  to  $-700$  ms,  $-800$  to  $-600$  ms,  $-700$  to  $-500$  ms). Within each trial, we computed Pearson's  $r$  values to quantify the similarity between the pre-activated temporal patterns and the post-target temporal patterns produced by unexpected and expected words separately, and subtracted these  $r$  values to yield within-trial similarity difference values (*Expected minus Unexpected*). (2) We repeated this analysis at each voxel within the region of interest ( $v_1, v_2, v_n$ ). B. MEG within-trial similarity effects. **Top:** Schematic depictions of the medial (in pink) and ventral (in green) portions of the left anterior-mid temporal lobe that showed a smaller univariate N400 response to expected relative to unexpected words in the univariate analysis, i.e. our two functional regions of interest. **Bottom:** Box plots display  $r$  values representing similarities between pre-activated temporal patterns (from  $-900$  to  $-700$  ms) and post-target temporal patterns (from 300 to 500 ms) for both *Expected* and *Unexpected* inputs. The *Expected* > *Unexpected* similarity effect was significant in the left anterior-mid medial temporal region, and approached significance in the left anterior-mid ventral temporal region.

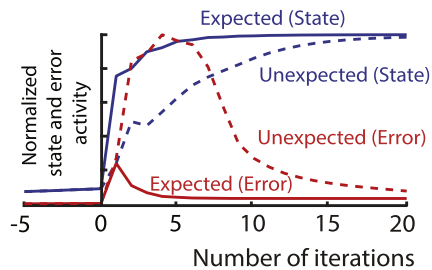
sampling space by multiplying it with a  $20 \times 2$  mixing matrix,  $M$ . Each individual element in this mixing matrix was a randomly-generated integer from 1 to 9, to reflect our assumption that any given voxel would capture a random mixture of state and error activity (see Fig. 6B for a schematic illustration using four voxels). We also added randomly generated Gaussian noise ( $\mu = 0, \sigma^2 = 3$ ) to the mixing space to simulate noise within the nervous system (see de Gardelle et al., 2013b; Faisal et al. 2008). Then, at each iteration of the predictive coding algorithm, for each expected and unexpected input, we extracted a vector that described the spatial patterns produced across all voxels within the sampling space. As shown schematically in Fig. 6C, we then computed the average similarity amongst the vectors for all pairs of expected words (*within-expected* similarity) and for all pairs of expected and unexpected words (*between-expected-unexpected* similarity) using Pearson's  $r$ , and subtracted these average similarity values from one another to construct time courses of the cross-trial similarity effect. Note that greater *within-expected* similarity than *between-expected-unexpected* similarity (*within-expected* > *between-expected-unexpected*) should be independent of any differences in signal magnitude; that is, it cannot be

trivially explained by the univariate effect (*unexpected* > *expected*).<sup>6</sup>

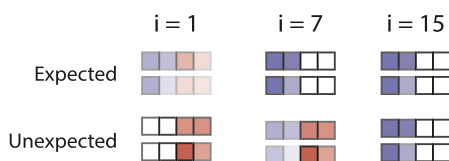
As shown in Fig. 6D, this cross-trial similarity analysis indeed revealed greater *within-expected* than *between-expected-unexpected* similarity, with a rise-and-fall time course: Starting from iteration 1, pairs of expected inputs produced patterns that were more spatially consistent than the patterns produced by pairs of expected and unexpected inputs. This difference peaked at iteration 7, after which expected and unexpected inputs became increasingly more similar, leading to a decrease in the effect. To test the significance of this effect, we averaged the

<sup>6</sup> This would not be true for the contrast, *within-unexpected* > *between-expected-unexpected*, which would be confounded by differences in signal magnitude between unexpected and expected inputs within this time window. The reason for this is that the greater signal-to-noise ratio in the unexpected condition would result in a more accurate estimation of the spatial similarity between any two unexpected words, resulting in a Pearson's  $r$  value that is larger (closer to the true Pearson's  $r$  value of 1) than between any two expected and unexpected words (see Wang & Kuperberg, 2024, page 10-12, for detailed discussion).

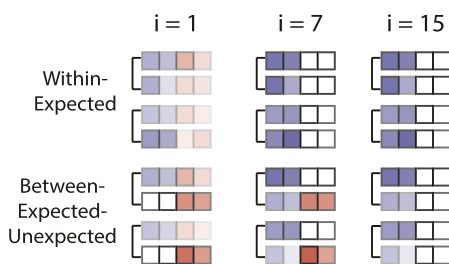
### A. Simulated error and state activity



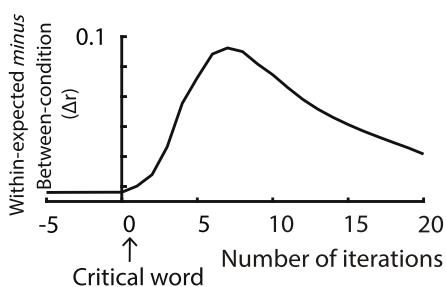
### B. Mix error & state outputs and project to 20-voxel space at each iteration $i$



### C. Calculate pairwise similarity amongst spatial patterns within this space



### D. Simulated cross-trial similarity effect: Within-Expected > Between-condition



(caption on next column)

**Fig. 6. Schematic depictions and results of cross trial similarity analysis of simulated data.** A. Outputs of the predictive coding model produced by lexico-semantic error and state units in response to *expected* and *unexpected* inputs. Normalized state and error activity was calculated by dividing their respective maximum value across the *expected* and *unexpected* conditions over all iterations. B. For each simulated trial, at each iteration,  $i$ , of the algorithm, we mixed and projected the state and error activity into a 20-voxel “sampling space”. We illustrate the resulting spatial pattern schematically with a four-element vector at three representative iterations: At  $i = 1$ , the activity produced by all expected inputs reflected a mixture of state and error activity, while the activity produced by unexpected inputs was dominated by error activity; at  $i = 7$ , the activity produced by all expected inputs was primarily driven by state activity, whereas the activity produced by unexpected inputs reflected a mixture of both state and error activity; at  $i = 15$ , the activity produced by both expected and unexpected inputs was dominated by state activity. C. At each iteration of the algorithm, we calculated the pairwise similarity between spatial patterns for all pairs of expected words (*within-expected* similarity) and all pairs of expected and unexpected words (*between-expected-unexpected* similarity). This process is illustrated schematically for two pairs at three representative iterations:  $i = 1$ ,  $i = 7$  and  $i = 15$ . D. We then subtracted these average similarity values from one another to yield a full time course of the simulated cross-trial similarity effect (*Within-expected* minus *Between-expected-unexpected*) within the 20-voxel sampling space.

similarity difference values between iterations 2–11 — the time window we used to quantify the simulated univariate effect — and carried out a two-sample  $t$ -test. This analysis revealed a significant effect across all pairs of simulated trials,  $t_{(28,678)} = 289.17$ ,  $p < .001$ .

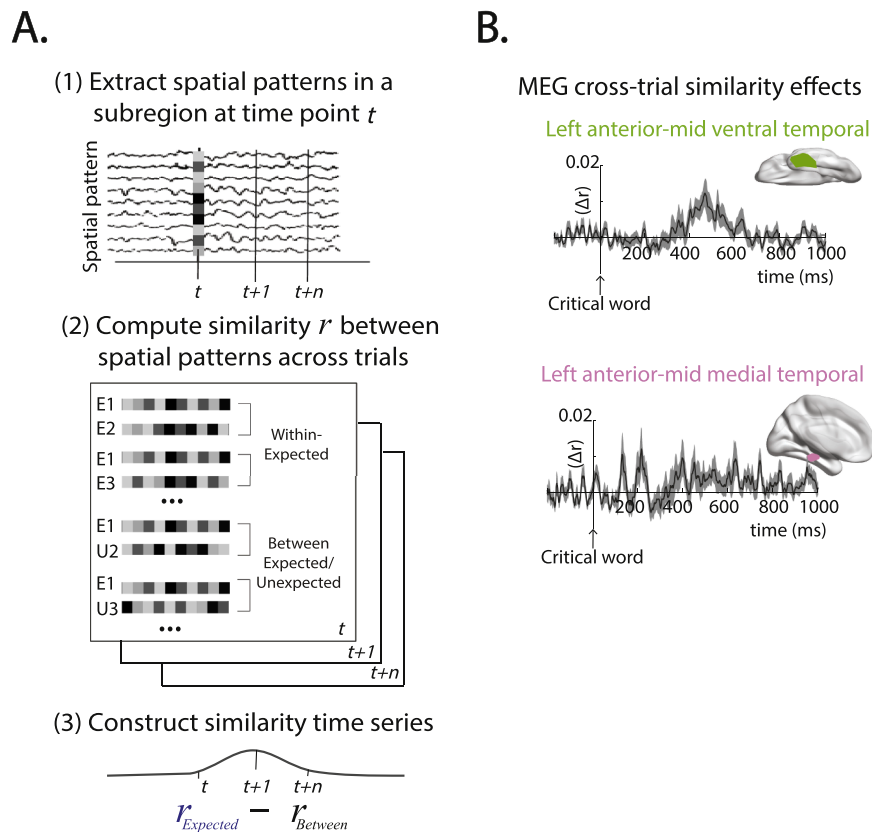
#### 3.3.2. MEG

Having established that the convergence of the predictive coding algorithm on expected inputs should give rise to an *increase* in similarity amongst pairs of expected words (versus pairs of expected and unexpected words), despite producing only a small univariate response, we tested for a *within-expected* > *between-expected-unexpected* effect in the MEG data.

We began by testing for a cross-trial similarity effect (*within-expected* > *between-expected-unexpected*) in the same left anterior-mid ventral temporal and the left anterior-mid medial temporal regions that produced a larger univariate response to *unexpected* versus *expected* inputs between 300 and 500ms. As depicted schematically in Fig. 7A, for each of these two ROIs, we quantified the similarity amongst fine-grained spatial patterns across all voxels for all pairs of expected words (*within-expected*) and for all pairs of expected and unexpected words (*between-expected-unexpected*) and computed the cross-trial similarity effect (*within-expected* minus *between-expected-unexpected*). As shown in Fig. 7B, within our 300–500 ms time window of interest, we found a significant cross-trial similarity effect (*within-expected* > *between-expected-unexpected*) within the left anterior-mid ventral temporal ROI (pairwise  $t$ -test:  $t_{(25)} = 2.86$ ,  $p = .008$ ). However, the effect in the left anterior-mid medial temporal ROI did not reach significance ( $t_{(25)} = 1.44$ ,  $p = .16$ ).<sup>7</sup>

To determine whether any other regions within the larger left fronto-temporal search volume also showed a cross-trial similarity effect, we carried out a similar cross-trial similarity analysis in 14 additional anatomical subregions that we defined within the larger search volume (see Supplementary Figure S1, Supplementary Table 1). Between 300–500 ms, we found significant cross-trial similarity effects (*within-expected* > *between-expected-unexpected*) in the mid-portion of the left lateral temporal cortex (including the left middle temporal cortex and the left mid- inferior temporal cortex). This may have reflected activity

<sup>7</sup> Note that unlike some representational analysis streams that simply ask whether patterns of activity can discriminate between two conditions (Kriegeskorte, Mur, & Bandettini, 2008; Nili et al., 2014), our analysis stream allowed us to look at the directionality of the effects.



**Fig. 7. Cross-trial similarity analysis and MEG cross-trial similarity effects:** **A.** Schematic illustration of the cross-trial similarity analysis conducted on the MEG source-localized data within one region of interest. (1) For each trial, at each time point,  $t$ , after the presentation of bottom-up input, we extracted a vector that represented the spatial pattern produced across all voxels. (2) At each time point, we computed Pearson's  $r$  values to quantify the similarity between the vectors produced by all pairs of Expected (E) words (e.g. between  $E_1$  and  $E_2$ , between  $E_1$  and  $E_3$ ) and all pairs of Expected (E) and Unexpected (U) words (e.g. between  $E_1$  and  $U_2$ , between  $E_1$  and  $U_3$ ). At each time point, we computed an average *Between-expected-unexpected* similarity  $r$  value (averaged across all *Between-expected-unexpected* pairs) and subtracted this from an average *Within-expected* similarity value (averaged across all *Within-expected* pairs), i.e. *Within-expected* minus *Between-expected-unexpected*. (3) We repeated this analysis at each time point, yielding a time series of the cross-trial similarity effect. **B.** MEG cross-trial similarity effects (*Within-expected* minus *Between-expected-unexpected*) within the two functional regions of interest that produced a larger univariate N400 response to unexpected than expected inputs (see Fig. 3). For each region, we show the time course of the cross-trial similarity effect, with standard errors indicated with gray shading. The left anterior-mid ventral temporal region showed greater *within-expected* than *between-expected-unexpected* cross-trial similarity between 300 and 500 ms, but the effect in the left anterior-mid medial temporal region did not reach significance.

in the pathway that maps indirectly from orthography to meaning through phonology, see Supplementary Materials (Section 2).

Finally, to exclude the possibility that any neural similarity effect (i.e. *within-expected* > *between-expected-unexpected*) was driven by the pairwise difference in the frequency or visual complexity of the critical words, we computed each of these values across all pairs of expected words and across all pairs of expected and unexpected words. Non-parametric tests did reveal some differences. However, these differences went in the opposite direction to the observed neural similarity effect; that is, pairs of expected words (*within-expected* pairs) were *less* similar to one another (i.e. showed greater differences) than pairs of expected and unexpected words (*between-expected-unexpected* pairs) in both frequency (*within-expected*:  $62 \pm 126$  vs. *between-expected-unexpected*:  $53 \pm 115$ ,  $p = .001$ ) and visual complexity (*within-expected*:  $6 \pm 4$  vs. *between-expected-unexpected*:  $5 \pm 4$ ,  $p = .001$ ).

#### 4. Discussion

Can the computational principles of predictive coding explain how the brain extracts meaning from the form of incoming words during language processing? To address this question, we used MEG to measure univariate and multivariate neural activity in response to expected and unexpected words during word-by-word reading comprehension, and we simulated this activity using an implemented predictive coding

model of lexico-semantic processing (Nour Eddine et al., 2024).

Replicating our previous findings (Wang et al., 2023), we found that between 300 and 500 ms, expected incoming words produced a smaller N400 response than unexpected but plausible words within the left ventromedial temporal lobe. Critically, we extend this previous work by demonstrating that, despite producing a smaller univariate response, expected words still generated consistent multivariate patterns within the same neuroanatomical region and time window. First, *within* each trial, individual voxels within this region produced unique temporal patterns between 300 to 500 ms that mirrored the temporal patterns that were predictively pre-activated by the prior context. Second, *across* trials, pairs of expected words produced spatial patterns that were more similar to one another than the spatial patterns produced by pairs of expected and unexpected words. Our simulations showed that the univariate effect as well as both multivariate effects could be explained by the dynamics of the predictive coding algorithm as it inferred the meaning of incoming words from their orthographic form.

##### 4.1. The larger N400 evoked response to unexpected versus expected inputs within the left anterior-mid ventromedial temporal lobe reflects the production of lexico-semantic prediction error during predictive coding

The larger evoked response produced by unexpected (versus expected) inputs (the N400 effect) selectively localized to regions of the



left temporal lobe that support lexico-semantic processing. These included the anterior-mid portion of the left medial and ventral temporal lobes, that play a key role in accessing lexical representations that map onto semantic features (Hirshorn et al., 2016; Woolnough et al., 2021) and accessing amodal concepts from these semantic features (Lambon-Ralph et al., 2017).

Our simulations using an implemented predictive coding model (Nour Eddine et al., 2024) confirmed that this specific architecture and algorithm was able to explain why unexpected words produced a larger overall neural response than expected words within this region: When expected bottom-up inputs confirmed prior top-down lexical and semantic predictions, lexical and semantic state units converged on the expected representation without strongly activating lexical and semantic error units i.e. lexico-semantic prediction error was minimal. Therefore, the total magnitude of the evoked neural response to expected inputs was relatively small. In contrast, when unexpected bottom-up inputs were encountered, they additionally activated lexical and semantic error units (because prior top-down lexical and semantic predictions failed to suppress activity within these error units), resulting in a larger lexico-semantic prediction error and a larger overall evoked neural response.

Our simulations also confirmed that the dynamics of the predictive coding algorithm explained the rise-and-fall time course of the evoked N400 to the unexpected inputs: The initial activation of lexical and semantic error units (prediction error), which drove the rise of the evoked response to unexpected inputs, served to update semantic and conceptual state units so that they converged on increasingly more accurate representations of the input. This, in turn, resulted in increasingly more accurate top-down predictions, which suppressed the prediction error, resulting in the subsequent fall of the evoked response to the unexpected inputs at the end of the N400 time window.

While these univariate findings are consistent with a predictive coding framework, they alone do not provide conclusive evidence for this theory. This is because other computational models can also explain the effect of predictability on the univariate N400 response during language comprehension (Brouwer et al., 2017; Fitz and Chang, 2019; Rabovsky et al., 2018), although in these models the effect was computed externally by the modeler rather than emerging from the model's dynamics. Moreover, because these previous models assume that a prior predictive context can change the state of the model before new bottom-up input becomes available, they can also potentially account for some previous reports of anticipatory neural activity observed before new bottom-up input becomes available (see Rabovsky, 2020; Yan, Kuperberg, & Jaeger, 2017, for discussion). For example, some researchers have reported effects of contextual constraint on a frontally-distributed anticipatory ERP component (e.g. Grisoni et al., 2021; Leon-Cabrera et al., 2019), which correlates with the magnitude of the N400 produced by subsequent inputs that confirm prior predictions (Grisoni et al., 2021). The presence of this type of anticipatory neural activity, however, doesn't necessarily entail that predicted upcoming information is actively propagated down the cortical hierarchy to pre-activate lower-level lexical representations within the temporal cortex, or that these lexical representations are reinstated by expected inputs. As we discuss next, our multivariate findings provide more direct evidence for these more specific claims of predictive coding.

#### 4.2. Within each trial, pre-activated item-specific representations are reinstated by expected inputs within the left ventromedial temporal lobe between 300 and 500ms

Consistent with the claim that in predictive contexts, lexical-level representations are pre-activated before new bottom-up input becomes available, in previous work we used MEG together with multivariate methods to demonstrate that predicted upcoming individual words (e.g. the word, "baby" in the context "In the crib, there is a sleeping..."; Wang et al., 2018b) are encoded as distinct temporal patterns within the left

ventromedial temporal lobe. These item-specific neural patterns were observed within a pre-activation time window that immediately followed the onset of the pre-target word. This early onset and relatively transient predictive pre-activation has also been observed in other studies (Wang et al., 2024a; Wang et al., 2020) and cannot be explained by activity produced by the pre-target word itself.

In the present study, our within-trial similarity analysis extended these previous findings by showing that when expected bottom-up input confirmed these prior top-down lexico-semantic predictions, the same temporal patterns were reinstated between 300 and 500ms. That is, within each trial, when an incoming target word was expected, it produced unique temporal patterns between 300 and 500 ms that were more similar to the *pre-activated* temporal patterns than when it was unexpected. This finding is consistent with a previous scalp-recorded EEG study that also found evidence the reinstatement of prior predictions (Hubbard and Federmeier, 2021). Critically, here we show that this prediction reinstatement effect localized to the same medial portion of the left anterior-mid temporal lobe that contributed to the larger univariate response to unexpected inputs. It also extended to a more posterior portion of the left ventromedial temporal lobe, which showed evidence of predictive pre-activation in our previous study (Wang et al., 2018b). Our simulations suggest that these patterns may have reflected the dynamic and interactive process of reinstating pre-activated mappings from form to meaning as the predictive coding algorithm settled on the expected conceptual and orthographic word-form representations between 300 and 500 ms (see Rogers et al., 2021 for a similar account).

A top-down propagation of predictions and reinstatement of pre-activated lower-level lexico-semantic representations by expected inputs is also posited by some qualitative models of predictive language comprehension (e.g. DeLong et al., 2005; Federmeier, 2007; Lau et al., 2008; see Kuperberg and Jaeger, 2016, section 3 for discussion) and could, in principle, be implemented by classic Interactive Activation and Competition (IAC) architectures (McClelland and Elman, 1986; McClelland and Rumelhart, 1981). Similar to predictive coding, these architectures also include feedback connections that allow predicted information to flow from higher to lower levels of the linguistic/cortical hierarchy, and feedforward connections that allow expected inputs to reinstate prior predictions when they become available to lower-level representations. However, unlike predictive coding, this type of single-unit architecture cannot easily explain why or how the same medial temporal voxels that produce unique item-specific temporal patterns by sharpening on *expected* inputs, should also produce a larger univariate N400 response to *unexpected* inputs (see Lee and Mumford, 2003, for discussion in the visual system).

In contrast, by explicitly positing not only the top-down generation of predictions, but also a functional distinction between error and state units that co-exist within the same neuroanatomical region, predictive coding is naturally able to account for this finding; that is, it can explain why, between 300 and 500 ms, the same region is able to both reinstate prior predictions to expected inputs (by activating state units) and produce a larger evoked response to unexpected inputs (by activating error units): Upon encountering unexpected inputs, prior top-down predictions fail to suppress activity within lexical and semantic *error units*, resulting in the larger overall univariate response within this region. However, when expected inputs are encountered, despite suppressing error activity (and therefore producing minimal prediction error and a small univariate evoked response), *state units* rapidly converge on the precise lexico-semantic representation that were previously pre-activated, explaining the within-trial similarity effect.

#### 4.3. Across trials, expected inputs activated state units that are functionally distinct from error units activated by unexpected inputs within the left ventromedial temporal lobe between 300 and 500ms

The distinction between state and error units is perhaps the most distinguishing feature of predictive coding (Friston, 2010; Mumford,

1992; Rao and Ballard, 1999; Spratling, 2016; Walsh et al., 2020). Although the analysis described above provides some evidence for this distinction, our cross-trial similarity analysis provides even more direct evidence for this claim.

The precise representation of state and error units within cortical microcircuitry remains unclear (Bastos et al., 2012; Mikulasch et al. 2022; Shipp, 2016), and MEG lacks the spatial resolution to directly detect these units. However, on the assumption that the signal detected within each voxel within the left ventromedial temporal region reflected a random mixture of activity from both state and error units, then if expected inputs only activate state units between 300 and 500 ms (by reinstating of prior top-down lexico-semantic predictions), while unexpected inputs activate both state and error units, then, on average, the spatial patterns produced by pairs of expected words should be more similar to each other than the spatial patterns produced by pairs of expected and unexpected words. This is precisely what we found: Between 300–500 ms post target onset, the spatial patterns produced by pairs of *expected* words within the anterior-mid portion of the left ventral temporal cortex (including anterior inferior temporal and anterior-mid fusiform regions) were significantly *more similar* to one another than those produced by pairs of expected and unexpected words. We therefore interpret this finding as evidence for a functional distinction between the units activated by expected and unexpected inputs (see de Gardelle et al., 2013a, 2013b, for similar findings in low-level perception).

This interpretation was confirmed by our simulations, which showed a similar cross-trial similarity effect within a 20-voxel sampling space (*within-expected* > *between-expected-unexpected*). These simulations not only reproduced this basic effect; they also showed that its rise-and-fall time course emerged as a direct consequence of the change in the proportions of state versus error units activated by expected versus unexpected inputs over the course of the predictive coding algorithm. Specifically, when the bottom-up input was first encountered, pairs of expected inputs produced patterns that were more spatially consistent than the patterns produced by pairs of expected and unexpected inputs, driving the rise of the *within-expected* > *between-expected-unexpected* similarity effect. However, as the predictive coding algorithm iteratively converged on the state representations that encoded the unexpected inputs, the spatial patterns produced by the expected and unexpected inputs became increasingly more similar (because both the unexpected and expected inputs only activated state units), leading to the subsequent fall of the effect (see Fig. 6 for a schematic illustration).

This cross-trial similarity effect cannot be trivially explained by the larger univariate response to unexpected inputs. The univariate analysis detected differences in the *strength* of neural activity within the left ventromedial region. In contrast, our cross-trial similarity analysis probed the pair-wise *similarity* between *patterns* of activity produced across voxels within this region, using a metric of similarity (Pearson's  $r$ ) that is mathematically independent of signal strength. Although increases in signal strength can inflate Pearson's  $r$  estimates,<sup>8</sup> our effect of interest (*within-expected* > *between-expected-unexpected*) went in the opposite direction to the univariate effect (*unexpected* > *expected*); that is, despite producing a *smaller* overall response within the left anterior-mid ventromedial temporal region between 300 and 500 ms, pairs of expected words produced *more similar* spatial patterns than pairs of expected and unexpected inputs in this same region within the same 300–500 ms time window.

This type of functional distinction cannot be explained by previous

<sup>8</sup> This is because a greater signal-to-noise ratio often leads to a more accurate estimation of Pearson's  $r$ , and so if a particular condition evokes a larger signal, then the average pairwise similarity amongst patterns within that condition will often also be greater (e.g., Guggenmos et al., 2018; Haxby et al., 2001; Jimura & Poldrack, 2012; Walther et al., 2016; see Wang & Kuperberg, 2024, pages 10-12, for detailed discussion).

models that successfully simulated the larger evoked N400 response to unexpected inputs as a prediction error produced outside the language comprehension system (Fitz and Chang, 2019; Rabovsky and McRae, 2014), or as change-of-state within the same connectionist units (Brouwer et al., 2017; Falandays et al., 2021; Luthra et al., 2021; Rabovsky et al., 2018).<sup>9</sup> Similarly, it cannot be explained by classic interaction activation models that also assume that expected and unexpected information are encoded within the same connectionist units (McClelland and Elman, 1986; McClelland and Rumelhart, 1981).

We therefore take these findings as strong evidence that distinct state and error units work closely together within the left ventromedial temporal region to implement the specific predictive coding optimization algorithm that approximates Bayesian inference.

#### 4.4. A division of labor within the left ventromedial temporal lobe

To sum up, consistent with the claims of predictive coding, both the within-trial and the cross-trial similarity analyses provided evidence that overlapping regions within the left ventromedial temporal lobe that produce a larger evoked responses to *unexpected* inputs, also converge on *expected* lexico-semantic representations between 300 and 500ms. There were, however, some interesting differences between the two types of multivariate effects.

The *within-trial* similarity effect manifested as unique item-specific *temporal patterns* across the 300–500 ms time window, which reinstated item-specific temporal patterns that were pre-activated before the expected bottom-up input became available. This item-specific reinstatement effect was most prominent in the *medial* portion of the anterior-mid temporal lobe, and it also extended to a more posterior ventromedial region that did not produce a univariate effect. Conversely, the cross-trial similarity effect (*within-expected* > *between-expected-unexpected*) manifested as differences in *spatial patterns* within the 300–500 ms time window that were *not* specific to individual items. Moreover, this effect localized primarily to the ventral portion of the left anterior-mid temporal lobe (the left anterior inferior temporal and anterior-mid fusiform cortex).

These findings raise the intriguing possibility of a division of labor within the left ventromedial temporal lobe during reading comprehension. Specifically, we suggest that the item-specific *temporal* patterns may have played a functional role in synthesizing specific sets of distributed features into individual discrete items (cf. Damasio, 1989). The anterior-mid medial temporal region may serve to bind unique combinations of multimodal *semantic* features, e.g. ⟨cute⟩, ⟨small⟩ and ⟨cry⟩, encoded across widespread cortical regions (Huth et al., 2016; Martin, 2007), into single amodal concepts, e.g. {baby} (Chen et al., 2016; Lambon-Ralph et al., 2017; Patterson et al., 2007). Analogously, the posterior ventral fusiform cortex (the visual word-form area) may have functioned to synthesize particular combinations of letters or bi/trigrams, e.g. “BA”, “AB” and “BY”, encoded within still more posterior occipital regions (Dehaene et al. 2005; Vinckier et al., 2007), into specific word-forms, e.g. “BABY” (Dehaene and Cohen, 2011; Price and Devlin, 2011).

In contrast, the *spatial* patterns observed within the left anterior-mid ventral temporal lobe may have played a more general role in *mapping* visual word-forms on to their semantic features, regardless of individual item. This would be consistent with previous work showing that anterior-mid portions of the left ventral temporal lobe encode lexical representations that implement such form-meaning mappings

<sup>9</sup> Carrying out analogous simulations using these previous models is not possible for two reasons: First, these models do not compute activity at each time step. Second, in these previous models, the increased activity produced by unexpected inputs (the univariate N400 response) was not computed by local error units within the model itself but was rather computed externally by the modeler.

(Caramazza, 1996; Hirshorn et al., 2016; Woolnough et al., 2021).

#### 4.5. The brain employs the same basic predictive coding mechanism to carry out inference during language processing as in non-linguistic domains

Predictive coding offers a biologically plausible framework for understanding the well-established effects of predictability on the N400 effect during language comprehension (Nour Eddine et al., 2024), including its neuroanatomical localization (Wang et al., 2023). Here, we show that its specific computational and architectural principles can also explain the timing, neuroanatomical localization and dynamics of *multivariate* activity produced within this crucial 300–500 ms time window. As such, the present findings directly link the neurobiology of reading comprehension with predictive coding research across multiple domains of perception and cognition (Clark, 2013; Rao & Ballard, 1999; Spratling, 2014; Spratling, 2016). More generally, they provide evidence that the same canonical circuit motif (cf. Douglas et al. 1989) that implements the predictive coding algorithm in these other domains (Bastos et al., 2012) may also support the extraction of meaning during language comprehension.

#### CRedit authorship contribution statement

**Lin Wang:** Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Investigation, Data curation, Conceptualization. **Samer Nour Eddine:** Writing – review & editing, Writing – original draft, Validation, Investigation, Conceptualization. **Trevor Brothers:** Writing – review & editing, Investigation. **Ole Jensen:** Writing – review & editing, Writing – original draft, Visualization, Methodology, Investigation, Conceptualization. **Gina R. Kuperberg:** Writing – original draft, Visualization, Supervision, Methodology, Investigation, Conceptualization.

#### Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this article.

#### Funding and acknowledgement

Research reported in this publication was supported by the Eunice Kennedy Shriver National Institute of Child Health & Human Development of the National Institutes of Health under Award Number R01HD082527. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. Drs. Kuperberg and Wang were also supported by Jeff Stibel.

#### Data and code availability statement

The MEG data that support the findings of this study are available on request from the corresponding author (GRK). Codes used for computational simulations are available at: <https://github.com/samer-nour-eddine/REPN400>

#### Supplementary materials

Supplementary material associated with this article can be found, in the online version, at [doi:10.1016/j.neuroimage.2024.120977](https://doi.org/10.1016/j.neuroimage.2024.120977).

#### Data availability

Data will be made available on request.

#### References

- Aitchison, L., Lengyel, M., 2017. With or without you: predictive coding and Bayesian inference in the brain. *Curr. Opin. Neurobiol.* 46, 219–227. <https://doi.org/10.1016/j.conb.2017.08.010>.
- 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>.
- Bell, A.J., Sejnowski, T.J., 1997. The “independent components” of natural scenes are edge filters. *Vision Res.* 37 (23), 3327–3338. [https://doi.org/10.1016/s0042-6989\(97\)00121-1](https://doi.org/10.1016/s0042-6989(97)00121-1).
- Blank, H., Davis, M.H., 2016. Prediction errors but not sharpened signals simulate multivoxel fMRI patterns during speech perception. *PLoS Biol.* 14 (11), e1002577.
- Bornkessel-Schlesewsky, I., Schlesewsky, M., 2019. Toward a neurobiologically plausible model of language-related, negative event-related potentials. *Front. Psychol.* 10, 298. <https://doi.org/10.3389/fpsyg.2019.00298>.
- Brouwer, H., Crocker, M.W., Venhuizen, N.J., Hoeks, J.C.J., 2017. A neurocomputational model of the N400 and the P600 in language processing. *Cogn. Sci.* 41 Suppl 6, 1318–1352. <https://doi.org/10.1111/cogs.12461>.
- Cai, Q., Brysbaert, M., 2010. SUBTLEX-CH: chinese word and character frequencies based on film subtitles. *PLoS. One* 5 (6), e10729. <https://doi.org/10.1371/journal.pone.0010729>.
- Caramazza, A., 1996. The brain’s dictionary. *Nature* 380 (6574), 485–486. <https://doi.org/10.1038/380485a0>.
- Caucheteux, C., Gramfort, A., King, J.R., 2023. Evidence of a predictive coding hierarchy in the human brain listening to speech. *Nature Human Behav.* <https://doi.org/10.1038/s41562-022-01516-2>.
- Chen, Q., Mirman, D., 2012. Competition and cooperation among similar representations: toward a unified account of facilitative and inhibitory effects of lexical neighbors”: correction to Chen and Mirman (2012). *Psychol. Rev.* 119 (4), 898–899. <https://doi.org/10.1037/a0030049>.
- Chen, Y., Shimotake, A., Matsumoto, R., Kunieda, T., Kikuchi, T., Miyamoto, S., Ralph, M.A.L., 2016. The ‘when’ and ‘where’ of semantic coding in the anterior temporal lobe: temporal representational similarity analysis of electrocorticogram data. *Cortex* 79, 1–13.
- Clark, A., 2013. Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behav. Brain Sci.* 36 (3), 181–204. <https://doi.org/10.1017/S0140525X12000477>.
- Damasio, A., 1989. The brain binds entities and events by multiregional activation from convergence zones. *Neural Comput.* 1 (1), 123–132. <https://doi.org/10.1162/neco.1989.1.1.123>.
- de Gardelle, V., Stokes, M., Johnen, V.M., Wyart, V., Summerfield, C., 2013a. Overlapping multivoxel patterns for two levels of visual expectation. *Front. Hum. Neurosci.* 7, 158. <https://doi.org/10.3389/fnhum.2013.00158>.
- de Gardelle, V., Waszczuk, M., Egner, T., Summerfield, C., 2013b. Concurrent repetition enhancement and suppression responses in extrastriate visual cortex. *Cerebral Cortex* 23 (9), 2235–2244. <https://doi.org/10.1093/cercor/bhs211>.
- Dehaene, S., Cohen, L., 2011. The unique role of the visual word form area in reading. *Trends Cogn. Sci. (Regul. Ed.)* 15 (6), 254–262.
- Dehaene, S., Cohen, L., Sigman, M., Vinckier, F., 2005. The neural code for written words: a proposal. *Trends Cogn. Sci. (Regul. Ed.)* 9 (7), 335–341.
- DeLong, K.A., Urbach, T.P., Kutas, M., 2005. Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nat. Neurosci.* 8 (8), 1117–1121. <https://doi.org/10.1038/nn1504>.
- Douglas, R.J., Martin, K.A.C., Whitteridge, D., 1989. A Canonical Microcircuit for Neocortex. *Neural Comput* 1 (4), 480–488. <https://doi.org/10.1162/neco.1989.1.4.480>.
- Faisal, A.A., Selen, L.P., Wolpert, D.M., 2008. Noise in the nervous system. *Nat. Rev. Neurosci.* 9 (4), 292–303. <https://doi.org/10.1038/nrn2258>.
- Falandays, J.B., Nguyen, B., Spivey, M.J., 2021. Is prediction nothing more than multi-scale pattern completion of the future? *Brain Res.* 1768, 147578. <https://doi.org/10.1016/j.brainres.2021.147578>.
- Fan, L., Li, H., Zhuo, J., Zhang, Y., Wang, J., Chen, L., Jiang, T., 2016. The human Brainnetome Atlas: a new brain atlas based on connective architecture. *Cereb. Cortex.* 26 (8), 3508–3526. <https://doi.org/10.1093/cercor/bhw157>.
- 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>.
- Fitz, H., Chang, F., 2019. Language ERPs reflect learning through prediction error propagation. *Cogn. Psychol.* 111, 15–52. <https://doi.org/10.1016/j.cogpsych.2019.03.002>.
- Friston, K.J., 2010. The free-energy principle: a unified brain theory? *Nat. Rev. Neurosci.* 11 (2), 127–138. <https://doi.org/10.1038/nrn2787>.
- Grainger, J., Holcomb, P.J., 2009. Watching the word go by: on the time-course of component processes in visual word recognition. *Lang Linguist Compass* 3, 128–156.
- Grisoni, L., Tomasello, R., Pulvermuller, F., 2021. Correlated brain indexes of semantic prediction and prediction error: brain localization and category specificity. *Cereb. Cortex.* 31 (3), 1553–1568. <https://doi.org/10.1093/cercor/bhaa308>.
- Guggenmos, M., Sterzer, P., Cichy, R.M., 2018. Multivariate pattern analysis for MEG: a comparison of dissimilarity measures. *Neuroimage* 173, 434–447. <https://doi.org/10.1016/j.neuroimage.2018.02.044>.
- Hagoort, P., Hald, L., Bastiaansen, M., Petersson, K.M., 2004. Integration of word meaning and world knowledge in language comprehension. *Science* (1979) 304 (5669), 438–441.



- Harm, M.W., Seidenberg, M.S., 2004. Computing the meanings of words in reading: cooperative division of labor between visual and phonological processes. *Psychol. Rev.* 111 (3), 662–720. <https://doi.org/10.1037/0033-295X.111.3.662>.
- Haxby, J.V., Gobbini, M.I., Furey, M.L., Ishai, A., Schouten, J.L., Pietrini, P., 2001. Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science* (1979) 293 (5539), 2425–2430.
- Hirshorn, E.A., Li, Y., Ward, M.J., Richardson, R.M., Fiez, J.A., Ghuman, A.S., 2016. Decoding and disrupting left midfusiform gyrus activity during word reading. *Proc. Nat. Acad. Sci.* 113 (29), 8162–8167. <https://doi.org/10.1073/pnas.1604126113>.
- Hubbard, R.J., Federmeier, K.D., 2021. Representational pattern similarity of electrical brain activity reveals rapid and specific prediction during language comprehension. *Cereb. Cortex*. 31 (9), 4300–4313. <https://doi.org/10.1093/cercor/bhab087>.
- Huth, A.G., de Heer, Griffiths, T.L., Theunissen, F.E., Gallant, J.L., 2016. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature* 532 (7600), 453–458. <https://doi.org/10.1038/nature17637>.
- Jimura, K., Poldrack, R.A., 2012. Analyses of regional-average activation and multivoxel pattern information tell complementary stories. *Neuropsychologia* 50 (4), 544–552.
- Jung, T.P., Makeig, S., Humphries, C., Lee, T.W., McKeown, M.J., Iragui, V., Sejnowski, T.J., 2000. Removing electroencephalographic artifacts by blind source separation. *Psychophysiology*. 37 (2), 163–178. <https://doi.org/10.1017/S0048577200980259>.
- Kriegeskorte, N., Mur, M., Bandettini, P., 2008. Representational similarity analysis – Connecting the branches of systems neuroscience. *Front. Syst. Neurosci.* 2, 4. <https://doi.org/10.3389/fnro.2008.004.2008>.
- 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., 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. *J. Cogn. Neurosci.* 32 (1), 12–35. [https://doi.org/10.1162/jocn\\_a.01465](https://doi.org/10.1162/jocn_a.01465).
- Kuperberg, G.R., Jaeger, T.F., 2016. What do we mean by prediction in language comprehension? *Lang. Cogn. Neurosci.* 31 (1), 32–59. <https://doi.org/10.1080/23273798.2015.1102299>.
- Kutas, M., Federmeier, K.D., 2011. Thirty years and counting: finding meaning in the N400 component of the event-related brain potential (ERP). *Annu. Rev. Psychol.* 62, 621–647. <https://doi.org/10.1146/annurev.psych.093008.131123>.
- 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>.
- Lambon-Ralph, M.A., Jefferies, E., Patterson, K., Rogers, T.T., 2017. The neural and computational bases of semantic cognition. *Nat. Rev. Neurosci.* 18 (1), 42–55. <https://doi.org/10.1038/nrn.2016.150>.
- 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>.
- Lau, E.F., Phillips, C., Poeppel, D., 2008. A cortical network for semantics: (De)constructing the N400. *Nature Rev. Neurosci.* 9 (12), 920–933. <https://doi.org/10.1038/nrn2532>.
- Lee, T.S., Mumford, D., 2003. Hierarchical Bayesian inference in the visual cortex. *J. Opt. Soc. Am. A* 20 (7), 1434. <https://doi.org/10.1364/josaa.20.001434>.
- Leon-Cabrera, P., Flores, A., Rodriguez-Fornells, A., Moris, J., 2019. Ahead of time: early sentence slow cortical modulations associated to semantic prediction. *Neuroimage* 189, 192–201. <https://doi.org/10.1016/j.neuroimage.2019.01.005>.
- Luthra, S., Li, M.Y.C., You, H., Brodbeck, C., Magnuson, J.S., 2021. Does signal reduction imply predictive coding in models of spoken word recognition? *Psychon. Bull. Rev.* 28 (4), 1381–1389. <https://doi.org/10.3758/s13423-021-01924-x>.
- Maris, E., Oostenveld, R., 2007. Nonparametric statistical testing of EEG- and MEG-data. *J. Neurosci. Methods* 164 (1), 177–190. <https://doi.org/10.1016/j.jneumeth.2007.03.024>.
- Martin, A., 2007. The representation of object concepts in the brain. *Ann. Rev. Psychol.* 58, 25–45.
- McClelland, J.L., Elman, J.L., 1986. The TRACE model of speech perception. *Cogn. Psychol.* 18 (1), 1–86. [https://doi.org/10.1016/0010-0285\(86\)90015-0](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. *Psychol. Rev.* 88 (5), 375–407. <https://doi.org/10.1037/0033-295x.88.5.375>.
- Mikulasch, F.A., Rudelt, L., Wibrall, M., Priesemann, V., 2022. Where is the error? Hierarchical predictive coding through dendritic error computation. *Trends Neurosci.* <https://doi.org/10.1016/j.tins.2022.09.007>.
- Mumford, D., 1992. On the computational architecture of the neocortex. II. The role of cortico-cortical loops. *Biol. Cybern.* 66 (3), 241–251. <https://doi.org/10.1007/BF00198477>.
- Nili, H., Wingfield, C., Walther, A., Su, L., Marslen-Wilson, W., Kriegeskorte, N., 2014. A toolbox for representational similarity analysis. *PLoS Comput. Biol.* 10 (4), e1003553.
- Nolte, G., 2003. The magnetic lead field theorem in the quasi-static approximation and its use for magnetoencephalography forward calculation in realistic volume conductors. *Phys. Med. Biol.* 48 (22), 3637–3652.
- Nour Eddine, S., Brothers, T., & Kuperberg, G.R. (2022). The N400 in silico: a review of computational models. In K. Federmeier (Ed.), *Psychology of Learning and Motivation* (Vol. 76, pp. 123–206): Academic Press.
- Nour Eddine, S., Brothers, T., Wang, L., Spratling, M., Kuperberg, G.R., 2024. A predictive coding model of the N400. *Cognition* 246, 105755. <https://doi.org/10.1016/j.cognition.2024.105755>.
- Oostenveld, R., Fries, P., Maris, E., Schoffelen, J.-M., 2011. FieldTrip: open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Comput. Intell. Neurosci.* 2011, 1. <https://doi.org/10.1155/2011/156869>.
- Patterson, K., Nestor, P.J., Rogers, T.T., 2007. Where do you know what you know? The representation of semantic knowledge in the human brain. *Nature Rev. Neurosci.* 8 (12), 976–987. <https://doi.org/10.1038/nrn2277>.
- Piai, V., Roelofs, A., Rommers, J., Maris, E., 2016. Beta oscillations reflect memory and motor aspects of spoken word production. *Hum Brain Mapp* 36 (7), 2767–2780. <https://doi.org/10.1002/hbm.22806>.
- Price, C.J., Devlin, J.T., 2011. The interactive account of ventral occipitotemporal contributions to reading. *Trends Cogn. Sci. (Regul. Ed.)* 15 (6), 246–253. <https://doi.org/10.1016/J.Tics.2011.04.001>.
- Rabovsky, M., 2020. Change in a probabilistic representation of meaning can account for N400 effects on articles: a neural network model. *Neuropsychologia* 143, 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. *Nat. Hum. Behav.* 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>.
- Rao, R.P.N., Ballard, D.H., 1997. Dynamic model of visual recognition predicts neural response properties in the visual cortex. *Neural Comput.* 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. *Nat. Neurosci.* 2 (1), 79–87. <https://doi.org/10.1038/4580>.
- Rogers, T.T., Cox, C.R., Lu, Q., Shimotake, A., Kikuchi, T., Kunieda, T., Lambon Ralph, M.A., 2021. Evidence for a deep, distributed and dynamic code for animacy in human ventral anterior temporal cortex. *Elife* 10, e66276. <https://doi.org/10.7554/eLife.66276>.
- Shipp, S., 2016. Neural elements for predictive coding. *Front Psychol.* 7, 1792. <https://doi.org/10.3389/fpsyg.2016.01792>.
- 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>.
- Spratling, M.W., 2008. Predictive coding as a model of biased competition in visual attention. *Vis. Res.* 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 Comput.* 24 (1), 60–103. [https://doi.org/10.1162/NECO\\_a.00222](https://doi.org/10.1162/NECO_a.00222).
- Spratling, M.W., 2013. Image segmentation using a sparse coding model of cortical area V1. *IEEE Trans. Image Process.* 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. *J. Comput. Neurosci.* 36 (1), 97–118. <https://doi.org/10.1007/s10827-013-0471-7>.
- Spratling, M.W., 2016. Predictive coding as a model of cognition. *Cogn. Process* 17 (3), 279–305. <https://doi.org/10.1007/s10339-016-0765-6>.
- Srinivasan, M.V., Laughlin, S.B., Dubs, A., 1982. Predictive coding: a fresh view of inhibition in the retina. *Proc. R. Soc. Lond. Ser. B: Biol. Sci.* 216 (1205), 427–459. <https://doi.org/10.1098/rspb.1982.0085>.
- Van Berkum, J.J.A., 2009. The neurophysiology of ‘simple’ utterance comprehension: an ERP review. In: Sauerland, U., Yatsushiro, K. (Eds.), *Semantics and Pragmatics: From Experiment to Theory*. Basingstoke: Palgrave Macmillan, pp. 276–316.
- Van Veen, B.D., van Drongelen, W., Yuchtman, M., Suzuki, A., 1997. Localization of brain electrical activity via linearly constrained minimum variance spatial filtering. *IEEE Trans. Biomed. Eng.* 44 (9), 867–880. <https://doi.org/10.1109/10.623056>.
- 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>.
- Walsh, K.S., McGovern, D.P., Clark, A., O’Connell, R.G., 2020. Evaluating the neurophysiological evidence for predictive processing as a model of perception. *Ann. N. Y. Acad. Sci.* 1464 (1), 242–268. <https://doi.org/10.1111/nyas.14321>.
- Walther, A., Nili, H., Ejaz, N., Alink, A., Kriegeskorte, N., Diedrichsen, J., 2016. Reliability of dissimilarity measures for multi-voxel pattern analysis. *Neuroimage* 137, 188–200. <https://doi.org/10.1016/j.neuroimage.2015.12.012>.
- Wang, L., Brothers, T., Jensen, O., Kuperberg, G.R., 2024. Dissociating the pre-activation of word meaning and form during sentence comprehension: evidence from EEG Representational Similarity Analysis. *Psychon. Bull. Rev.* 31 (2), 862–873.
- Wang, L., Brothers, T., Jensen, O., Kuperberg, G.R., 2024a. Dissociating the pre-activation of word meaning and form during sentence comprehension: Evidence from EEG representational similarity analysis. *Psychonomic Bulletin & Review* 31 (2), 862–873.
- Wang, L., Hagoort, P., Jensen, O., 2018a. Language prediction is reflected by coupling between frontal gamma and posterior alpha oscillations. *J. Cogn. Neurosci.* 30 (3), 432–447. [https://doi.org/10.1162/jocn\\_a.01190](https://doi.org/10.1162/jocn_a.01190).
- Wang, L., Kuperberg, G.R., 2024. Better together: integrating multivariate with univariate methods, and MEG with EEG to study language comprehension. *Language, cognition and neuroscience* 39 (8), 991–1019.
- Wang, L., Kuperberg, G., Jensen, O., 2018b. Specific lexico-semantic predictions are associated with unique spatial and temporal patterns of neural activity. *Elife* 7, e39061. <https://doi.org/10.7554/eLife.39061>.



- Wang, L., Kuperberg, G.R., 2024. Better together: integrating multivariate with univariate methods, and MEG with EEG to study language comprehension. *Lang Cogn. Neurosci.* 39 (8), 991–1019.
- 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>.
- Wang, L., Wlotko, E., Alexander, E.J., Schoot, L., Kim, M., Warnke, L., Kuperberg, G.R., 2020. Neural evidence for the prediction of animacy features during language comprehension: evidence from MEG and EEG Representational Similarity Analysis. *J. Neurosci.* 40 (16), 3278–3291. <https://doi.org/10.1101/709394>.
- Whittington, J.C.R., Bogacz, R., 2019. Theories of error back-propagation in the brain. *Trends Cogn. Sci. (Regul. Ed.)* 23 (3), 235–250. <https://doi.org/10.1016/j.tics.2018.12.005>.
- Wicha, N.Y., Moreno, E.M., Kutas, M., 2004. Anticipating words and their gender: an event-related brain potential study of semantic integration, gender expectancy, and gender agreement in Spanish sentence reading. *J. Cogn. Neurosci.* 16 (7), 1272–1288. <https://doi.org/10.1162/0898929041920487>.
- Woolnough, O., Donos, C., Rollo, P.S., Forseth, K.J., Lakretz, Y., Crone, N.E., Tandon, N., 2021. Spatiotemporal dynamics of orthographic and lexical processing in the ventral visual pathway. *Nat. Hum. Behav.* 5, 389–398. <https://doi.org/10.1038/s41562-020-00982-w>.
- Xia, M., Wang, J., He, Y., 2013. BrainNet Viewer: a network visualization tool for human brain connectomics. *PLoS. One* 8 (7), e68910. <https://doi.org/10.1371/journal.pone.0068910>.
- Xiang, M., Kuperberg, G., 2015. Reversing expectations during discourse comprehension. *Lang. Cogn. Neurosci.* 30 (6), 648–672. <https://doi.org/10.1080/23273798.2014.995679>.
- Yan, S., Kuperberg, G.R., Jaeger, T.F., 2017. Prediction (or not) during language processing. A commentary on Nieuwland et al. (2017) and DeLong et al. (2005). *bioRxiv*. <https://doi.org/10.1101/143750>.