Predictive coding across the left fronto-temporal hierarchy during language comprehension

*Lin Wang[1,2], Lotte Schoot[1,2], Trevor Brothers[,1,2], Edward Alexander[1,2], Lena Warnke[1], Minjae Kim[2], Sheraz Khan[1,3], Matti Hämäläinen[1], *Gina R. Kuperberg[1,2]

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

2 Department of Psychology, Tufts University, Medford, MA, 02155, USA

3 Massachusetts Institute of Technology (MIT)

Corresponding authors' email addresses:

gkuperberg@mgh.harvard.edu

lwang48@mgh.harvard.edu

Classification: Biological Sciences, Psychological and Cognitive Sciences

Abstract

We used MEG and ERPs to track the time-course and localization of evoked activity produced by *expected*, *unexpected plausible* and *implausible* words during incremental language comprehension. We suggest that the full pattern of results can be explained within a hierarchical predictive coding framework in which increased evoked activity reflects the activation of residual information that was not already represented at a given level of the fronto-temporal hierarchy ("error" activity). Between 300-500ms, the three conditions produced progressively larger responses within left temporal cortex (lexico-semantic prediction error), while *implausible* inputs produced a selectively enhanced response within inferior frontal cortex (prediction error at the level of the event model). Between 600-1000ms, *unexpected plausible* words activated left inferior frontal and middle temporal cortices (feedback activity that produced top-down error), while highly *implausible* inputs activated left inferior frontal cortex, posterior fusiform (unsuppressed orthographic prediction error/reprocessing), and medial temporal cortex (possibly supporting new learning). Therefore, predictive coding may provide a unifying theory that links language comprehension to other domains of cognition.

Key words: ERP, hierarchy, MEG, prediction, probabilistic

Introduction

One of the most amazing feats of human cognition is our ability to comprehend language by transforming streams of linguistic input into a high-level representation of real-world events (an *event model*) (Radvansky and Zacks 2011). Language comprehension can be understood as a process of *probabilistic inference*: the use of prior knowledge, encoded within an internal generative model, to infer the underlying high-level representation that best “explains” the bottom-up input (Kuperberg and Jaeger 2016). According to an influential theory of brain function, probabilistic inference is approximated using an optimization algorithm known as *predictive coding* (Mumford 1992; Rao and Ballard 1999; Friston 2005; Clark 2013; Spratling 2017).

Predictive coding claims that higher levels of the cortical hierarchy, representing information over longer spatiotemporal scales, continually generate top-down predictions of activity at lower cortical levels, which represents information over shorter spatiotemporal scales. As new bottom-up information becomes available to lower levels, top-down predictions attempt to *suppress* the activity produced by these inputs. Any residual (unexplained) information encoded in the input, i.e. *prediction error*, is then used to update higher-level representations, allowing them to produce more accurate predictions. Inference is complete when, over multiple iterations of the algorithm, prediction error is minimized across all levels of the cortical hierarchy.

In the language domain, studies of speech perception (Blank and Davis 2016; Sohoglu and Davis 2020) and visual word recognition (Price and Devlin 2011; Heilbron, Richter, et al. 2020) have shown that predictable inputs produce less neural activity than unpredictable inputs in regions that encode low-level phonological and orthographic information. This has been taken as evidence for the suppression of lower-level prediction error by accurate top-down predictions. An important open question is whether the principles of hierarchical predictive coding can also account for the neuroanatomical localization and timing of evoked neural activity produced during higher-level language comprehension, which requires us to infer information represented at *multiple* time-scales as new linguistic inputs unfold in real time.

300-500ms

MEG studies have shown that single words, presented without any prior context, produce robust event-related (evoked) responses: Increases in time-locked neural activity, relative to baseline. These evoked responses begin in early perceptual regions, and are followed by subsequent peaks as activity rapidly flows from posterior to anterior regions of ventral and lateral temporal cortices, and then to inferior frontal and orbitofrontal cortices. Between 300-500ms, the mid- and anterior temporal and inferior frontal cortices are activated in parallel, particularly over the left hemisphere (e.g. Dale et al. 2000; Dhond et al., 2001; Halgren et al., 2002; Marinkovic et al. 2003). This feedforward sweep of activity propagates to the scalp surface, manifesting in both EEG and MEG as the N400 waveform, which is thought to reflect the initial influence of the incoming stimulus on the current state of semantic memory (Kutas and Federmeier 2011).

A large body of studies using event-related potentials (ERPs) has established that an N400 is triggered by each meaningful content word in a sentence (Van Petten and Kutas 1990, 1991; Payne et al. 2015). During sentence comprehension, the amplitude of the N400 is highly sensitive to the lexical predictability of an incoming word in relation to its prior context (Kutas and Hillyard 1984; DeLong et al. 2005; Federmeier et al. 2007), with *expected* words evoking a smaller N400 than *unexpected* but *plausible* words (e.g. *"For the sleepover, you should bring a blanket and a pillow/game"*). There is also evidence for an effect of contextual plausibility on the N400, over and above the effect of lexical predictability (Kuperberg et al. 2020; Nieuwland et al. 2020), with *unexpected plausible* words evoking a smaller N400 than *implausible* words (e.g. *"For the* *sleepover, you should bring a blanket and a game/sneeze"*). However, it remains unclear which regions in the fronto-temporal hierarchy contribute to the effects of lexical predictability and contextual plausibility. fMRI studies have had difficulty addressing this question because the sluggish hemodynamic response is largely insensitive to transient, feedforward neural activity (Furey et al. 2006), like that reflected by the N400 (Lau et al. 2008; Geukes et al. 2013; Lau et al. 2013; Lau et al. 2016). While previous MEG and intracranial EEG studies have shown that the effects of sentence context on the N400 can localize to *both* temporal and inferior frontal cortices (e.g. Helenius et al. 1998; Halgren et al. 2002; Maess et al. 2006; Ihara et al. 2007; Heilbron, Armeni, et al. 2020), none of these previous studies have separately manipulated lexical predictability and contextual plausibility. Determining which levels of the fronto-temporal hierarchy are modulated by these two factors is important because it can help distinguish between different neurobiological frameworks of language comprehension, including predictive coding.

One possibility is that the effects of lexical predictability and contextual plausibility on the N400 *both* localize to temporal and inferior frontal regions, with graded increases in activity across *expected, unexpected plausible* and *implausible* words. This would be consistent with a recent computational model (Rabovsky et al. 2018) that interprets the N400 as an implicit update in a single hidden layer that maps directly from lexical inputs to an event model. Assuming that this single state is distributed across both frontal and temporal cortices, the more *expected* the new input, the smaller the update, and the smaller the evoked response it should produce across these two regions. We will refer to this as a *distributed state* account (see Figure 1A for summary).

A second possibility is that, in plausible sentences, the N400 effect of lexical predictability localizes to regions of the left temporal cortex that support lexical processing (i.e. *expected* < *unexpected* = *implausible*), while, in unpredictable sentences, the N400 effect of contextual plausibility localizes to the inferior frontal cortex (i.e. *expected* = *unexpected* < *implausible*). This account is motivated by theories that draw a sharp distinction between lexico-semantic "access" (Lau *et al.* 2008; Lau et al. 2009) and event-level "integration" (Brothers et al. 2015; Nieuwland *et al.* 2020). Unlike the *distributed state* account, it assumes a *hierarchical* architecture in which semantic information associated with individual words (lexico-semantic information) is encoded at a shorter time-scale at *lower* levels of the fronto-temporal hierarchy (temporal cortex; Price 2012), while a full representation of the prior context is encoded and maintained over a longer time span within higher-level regions (e.g. the prefrontal cortex; Lerner et al. 2011). According to this framework, in plausible sentences, it should be easier to "access" lower-level lexico-semantic representations of *expected* than *unexpected plausible* words because prior lexico-semantic predictions would have already *pre-activated* these representations within the left temporal cortex (see Wang et al. 2018). This would be consistent with MEG and intracranial studies showing that the semantic priming effect localizes to regions within the left temporal cortex (Nobre and McCarthy 1995; Lau *et al.* 2013; Lau and Nguyen 2015), and that this facilitation effect is enhanced when top-down prediction is encouraged (Lau *et al.* 2016). In contrast, according to this account, the effect of contextual implausibility is not driven by differences in prior lexico-semantic prediction. Instead, inferring an implausible event leads to additional difficulties in "integrating" it with real-world knowledge, with this "integration" being mediated by the left inferior frontal cortex (Hagoort et al. 2004). We will refer to this as the *prediction-integration* account[1] (see Figure 1B for summary).

A third possibility is that the effects of lexical predictability and contextual plausibility on

[1] Another version of this *prediction-integration* account assumes a serial processing architecture in which lexical access occurs in the temporal cortex in the N400 time window, but high-level integration occurs at a later stage of processing, between 600-1000ms within frontal regions (Brouwer and Hoeks 2013).

the N400 localize *only* to temporal regions (*expected* < *unexpected plausible* < *implausible).* This account also assumes a hierarchical architecture, with the N400 reflecting activity at the lower lexico-semantic level. In contrast to the *prediction-integration* account, however, it assumes that comprehenders not only predict semantic features associated with specific lexical items (lexico-semantic predictions), but also semantic features that are associated with whole semantic categories (e.g. <animate> *versus* <inanimate>). When new input becomes available, these semantic-level predictions will also serve to facilitate the retrieval/access of *unexpected plausible* (versus *implausible*) lexico-semantic representations. Therefore, on this account, *all* contextual effects on incoming words — both lexical predictability and contextual plausibility — are hypothesized to "bottom-out" at the lower lexico-semantic level of representation (MacDonald et al. 1994; Smith and Levy 2013). Updates of higher-level states, which are maintained over a longer time scale, are assumed *not* to influence the N400, although the precise reasons for this are not usually specified. We will refer to this account as the *lexico-semantic facilitation* account (see Figure 1C for summary).

**** Insert Figure 1 here ***

Finally, hierarchical predictive coding predicts a fourth possible pattern: Graded N400 modulation across the three conditions within temporal regions, but non-graded modulation within the inferior frontal cortex (see Figure 1D). Similar to the *prediction-integration* and the *lexico-semantic* facilitation accounts, predictive coding assumes a hierarchical architecture. Here, we assume that, during language comprehension, lexico-semantic information associated with individual words is encoded at a shorter time-scale within the temporal cortex, and that a higher-level *event model* is represented over a longer time scale within the prefrontal cortex.

Within the temporal cortex, predictive coding posits that the prior context generates predictions that pre-activate semantic features, both of expected individual words and whole semantic categories, similar to the *lexico-semantic facilitation* account. Therefore, predictive coding also predicts progressively larger N400 responses across the three conditions within the temporal cortex (*expected* < *unexpected plausible* < *implausible*). However, instead of directly attributing these evoked effects to changes induced by the input in lexico-semantic *states*, predictive coding attributes them to differences in the magnitude of lexico-semantic prediction error[2] – activity produced by a distinct population of lexico-semantic *error units* that cannot be suppressed/explained by top-down predictions from the higher-level event model (see Figure 2, left; for computational simulations see Nour Eddine 2021; Nour Eddine *et al.* in press).

Within the inferior frontal cortex, predictive coding attributes differences in amplitude of the N400 to differences in the magnitude of prediction error produced at the *higher* level of the event model, i.e. activity within higher-level error units that is not suppressed by predictions based on real-world knowledge (see Figure 2, right). Within this framework, there should be no difference in the magnitude of higher-level prediction error produced by *expected* and *unexpected plausible* inputs because, in both cases, the event model is plausible (i.e. consistent with real-world

[2] Note that in predictive coding, the term, "prediction error" does *not* necessarily correspond to a *linguistic error/anomaly* or to the *violation* of a strong top-down lexical-level prediction, but rather simply to unpredicted information encoded in the bottom-up input (see Xiang and Kuperberg 2015; Bornkessel-Schlesewsky and Schlesewsky 2019; Kuperberg *et al.* 2020; Nour Eddine 2021; Nour Eddine et al. in press). We also note that the prediction error computed as part of the predictive coding algorithm is distinct from other descriptions of prediction error in the computational N400 literature - (a) a difference between a current semantic state and an "ideal" target (Rabovsky and McRae 2014), or (b) a difference between a word input and a model's prior lexical prediction (Fitz and Chang 2019). It is also distinct from "change-of-state" computational accounts in which the N400 is proposed to directly reflect the difference in state from before until after a new input is encountered (Brouwer et al. 2017; Rabovsky *et al.* 2018). In these standard connectionist architectures, the metric used to operationalize the N400 is calculated *outside* the neural network, solely for the purposes of longer-term learning (see Open Questions in Discussion). In contrast, in predictive coding, unpredicted/unexplained information at each level of the cortical hierarchy ("prediction error") is explicitly computed within a distinct set of error units, as part of an inference algorithm (see Nour Eddine et al. in press for discussion).

knowledge), and so any higher-level prediction error/evoked response is suppressed. However, in the case of *implausible* inputs, higher-level prediction error is no longer suppressed (see Rao and Ballard 1999). Therefore, predictive coding predicts an enhanced N400 response in the inferior frontal cortex only to *implausible* inputs (*expected* = *unexpected plausible* < *implausible*).

**** Insert Figure 2 here ***

600-1000ms

Beyond the N400 time-window, previous ERP studies have shown that unpredicted inputs can also produce additional evoked responses between 600-1000ms, particularly in rich, schema-constraining contexts. This later-stage activity manifests at the scalp surface as a set of positive-going waveforms between 600-1000ms, with different scalp distributions depending on whether the interpretation is plausible or implausible (Van Petten and Luka 2012; DeLong et al. 2014; Brothers et al. 2020; Kuperberg *et al.* 2020). Unexpected words that yield *plausible* interpretations produce a late *frontally-distributed* positivity (Federmeier *et al.* 2007; Brothers *et al.* 2020; Kuperberg *et al.* 2020), while anomalous inputs that yield highly *implausible* interpretations produce a late *posteriorly-distributed* positivity, also known as the semantic P600 (Kuperberg et al. 2003; Kuperberg 2007; van de Meerendonk et al. 2009; Brothers *et al.* 2020; Kuperberg *et al.* 2020).

Various theoretical proposals have been offered to explain these late ERP responses. For example, it has been proposed that the *late frontal positivity* reflects lexical inhibition (Ness and Meltzer-Asscher 2018), or a large late shift of the event model (Brothers *et al.* 2020; Kuperberg *et al.* 2020), while the *late posterior positivity/P600* has been linked to reprocessing of the perceptual input that is triggered by linguistic conflict (van de Meerendonk *et al.* 2009; Brothers *et al.* 2020;

Kuperberg *et al.* 2020; Brothers et al. 2021). However, we know very little about the neuroanatomical sources of this later-stage evoked activity. Although fMRI is more sensitive to later-stage feedback activity than earlier stimulus-driven feedforward transient responses (Furey *et al.* 2006) (like the N400), the sluggish hemodynamic response also captures much later activity that extends beyond 1000ms, and that is likely to reflect processing of subsequent words and offline reflections about the meaning of a sentence as a whole. Even though several MEG and intracranial studies have explored the effects of semantic context during sentence comprehension, they have generally failed to report effects beyond the N400 time-window.

Hierarchical predictive coding provides a biologically-motivated computational framework for understanding these late evoked responses. Within a non-stationary hierarchical generative framework (cf. Qian et al. 2012; Gershman et al. 2014), new information that is inconsistent with the comprehender's prior event model will induce attempts to retrieve new schema from long-term memory (Franklin et al. 2020; Kuperberg 2021). In the case of an *unexpected plausible* input, this retrieval will be successful, resulting in the generation of *new* top-down predictions. Predictive coding posits that, as these new predictions flow down the cortical hierarchy between 600-1000ms, they will produce "top-down error", i.e. they will activate error units that carry residual top-down information that is not already encoded within prior states (see Rao and Ballard 1997, 1999), both at the level of the event model and at the lexico-semantic level. This framework therefore predicts that, relative to *expected* inputs, *unexpected plausible* inputs will produce late evoked effects both within inferior frontal and temporal cortices (see Figure 3, left).

In contrast, in the case of a highly *implausible* input that yields an anomalous interpretation (e.g. "*...they cautioned the *drawers*"), the brain is unable to retrieve new schema, and it will therefore continue to generate incorrect top-down predictions. This will result in a failure to switch off prediction error at still lower levels of the cortical hierarchy that encode lower-level linguistic representations. This account therefore predicts that highly *implausible* words will produce an enhanced later evoked response within lower-level regions such as the posterior fusiform cortex, which supports sublexical orthographic processing (Price and Devlin 2011; Heilbron, Richter*, et al.* 2020), i.e. orthographic reprocessing (see Figure 3, right).

**** Insert Figure 3 here ***

This study

Here, we used MEG and ERPs to track the time-course and localization of evoked neural activity, produced in the first 1000ms following word onset, during incremental language comprehension. Participants read predictable multi-sentence discourse contexts, e.g. *"The lifeguards received a report of sharks right near the beach. Their immediate concern was to prevent any incidents in the sea. Hence, they cautioned the…"*. Critical words in the final sentence were (a) *expected* (e.g. *"swimmers"*), (b) *unexpected* but *plausible* (e.g. *"trainees"*), or (c) highly *implausible* (e.g. *"drawers"*)[3]. To examine the full time-course and neuroanatomical localization of the evoked responses produced by each condition, we carried out a distributed source localization analysis of the MEG data, which is relatively undistorted by conductivities of the skull and scalp. To test our a *priori* hypotheses, we carried out planned comparisons between each pair of conditions across a left-lateralized search region where we expected effects to be maximal (particularly the effects of lexical predictability, see Federmeier 2022), and within three 200ms time-windows of interest (300-500ms, corresponding to the N400 time-window, and 600-800ms

[3] Participants also read a set of *low constraint plausible* scenarios, as well as a set of *low constraint implausible* fillers (see Methods). The comparison between the *high constraint unexpected* and *low constraint unexpected* plausible continuations will be the focus of a separate report.

and 800-1000ms, corresponding to the first and second halves of the time-window associated with the late positivity ERP effects). We also report activity at every 100ms across both hemispheres (left hemisphere findings are shown in the Results section, and right hemisphere results are shown in Supplementary Materials).

Finally, in addition to collecting MEG data, we also collected EEG data simultaneously. This was important because MEG and EEG do not capture precisely the same underlying signal (Ahlfors et al. 2010; see Supplementary Materials for further discussion). Therefore, by collecting EEG data using the same stimuli and in the same participants, we were able to replicate previous ERP findings (Kuperberg *et al.* 2020), and link them to the source-localized evoked effects detected by MEG.

Materials and Methods

Materials

Participants read three types of three-sentence discourse scenarios, each with a constraining context, and a critical noun in the third sentence: (1) *Expected*, in which the critical word was predictable; (2) *Unexpected plausible*, in which the critical word was plausible but unpredictable because it was out of keeping with the schema set up by the prior context, and (3) *Implausible*, in which the critical word violated the animacy constraints of the preceding verb (which constrained either for an animate or an inanimate noun).

The stimuli were based on three of the conditions used in a previous ERP study (Kuperberg *et al.* 2020). A full description is provided there as well as in Supplementary Materials. Briefly, the discourse contexts of each scenario were constraining (average cloze probability of the most probable word: 68%), as quantified in a cloze norming study that was carried out in participants who were recruited through Amazon Mechanical Turk. These contextual constraints came from the entirety of the discourse context — the first two sentences plus the first few words of the third sentence before the critical word. In all scenarios, these first few words of the third sentence included an adjunct phrase of 1-4 words, followed by a pronominal subject that referred back to the first two sentences, a verb and a determiner.

To create the *expected* scenarios, each context was paired with the noun with the highest cloze probability for that context. To create the *unexpected plausible* scenarios, each context was paired with a noun of zero (or very low) cloze probability, but that was still plausible in relation to this context. To create the *implausible* scenarios, each context was paired with a noun that violated the animacy constraints of the preceding verb, see Table 1. In all scenarios, the critical noun was followed by three additional words to complete the sentence.

When constructing these materials, we attempted to minimize the possibility of "bottom-up priming" between words in the prior context and the critical word. First, the critical noun and the preceding verb were never strongly associated, as confirmed in a separate cloze norming study (described in Kuperberg *et al.* 2020, see pages 17-18). In that cloze norming study, participants provided continuations for very short sentence contexts constructed around the verb (proper name + verb + determiner, e.g. "Jane cautioned the…"). Critical nouns were provided only rarely as a completion (mean cloze = 6%, SD = 7%). This meant that, in the *expected* condition of the present study, the critical word was separated from any semantically related words in the prior discourse by a sentence boundary (see Van Petten et al. 1997 for evidence that purely associative priming between individual pairs of words is short-lived, and likely plays a minimal role in driving discourse contextual effects on the N400).

Second, we carefully matched the *unexpected plausible* and *implausible* conditions on the degree of "semantic relatedness" between the critical word and the full "bag of words" in the prior context, operationalized using Semantic Similarity Values, extracted using Latent Semantic Analysis (Landauer and Dumais 1997), see Table 1 ($t_{(248)} = 0.09$, $p = 0.93$, $d = 0.01$).

**** Insert Table 1 here ***

Each list contained 25 *expected*, 25 *unexpected plausible,* and 50 *implausible* scenarios. This ensured that each participant saw an equal proportion of plausible and implausible scenarios. Conditions were counterbalanced across lists such that, across all participants, (a) the same 25 discourse contexts appeared in all three conditions, but no individual saw the same context more than once, and (b) the same 25 critical words appeared in the 25 *unexpected plausible* scenarios and 25 of the *implausible* scenarios, but no individual saw this word more than once. We made the *a priori* decision to include all 50 implausible scenarios in our main analyses in order to maximize power to source-localized effects (see below). As a byproduct of counterbalancing the same high constraint contexts across the three conditions, the critical words in the *expected* scenarios had fewer letters, smaller orthographic neighborhoods and were more frequent than in the two other conditions (all *ts* > 5, *ps* < 0.001, *ds* > 0.54), see Supplementary Materials for an additional analysis that addresses potential concerns regarding these differences. However, all these lexical features were matched between the *unexpected plausible* and the *implausible* conditions (all *ts* < 1.74, *ps* > 0.08, *ds* < 0.16), see Table 1.

As discussed in Supplementary Materials, each list additionally included 100 scenarios with less constraining contexts (again, 50% plausible and 50% implausible). The 50 *low constraint implausible* scenarios served as fillers, and the 50 *low constraint plausible* scenarios served as a fourth experimental condition. In the present study, we focused on the three high constraint conditions, which were perfectly matched prior to the onset of the critical word. Analyses that directly contrast the *expected*, *high constraint plausible* and *low constraint plausible* words will be reported in a separate manuscript.

Overall procedure

Participants took part in two separate experimental sessions: one in which we simultaneously collected MEG/EEG data, and one in which we collected structural and functional MRI data. The present manuscript focuses on the EEG and MEG datasets. Below, we describe the acquisition and analysis of the MEG/EEG data, as well as the structural MRI data, which was used to constrain MEG source localization. A report of the fMRI dataset, together with detailed comparisons with the EEG/MEG dataset, will appear in a separate manuscript.

Participants

Thirty-three participants took part, but one MEG/EEG dataset was excluded because of technical problems. Here we report the results of 32 MEG/EEG datasets (16 females, mean age: 23.4; range: 18-35). All participants were native speakers of English (with no other language exposure before the age of 5), were right-handed, and had normal or corrected-to-normal vision. All were screened to exclude past or present psychiatric and neurological disorders, and none were taking medication affecting the Central Nervous System. The study was approved by the Mass. General Brigham Institutional Review Board (IRB), and written informed consent was obtained from all participants.

Stimuli presentation and task

Stimuli were presented using PsychoPy 1.83 software and projected on to a screen in white Arial font (size: 0.1 of the screen height) on a black background. On each trial, the first two sentences appeared in full (each for 3900ms, 100ms interstimulus interval, ISI), followed by a fixation (a white "++++"), which was presented for 550ms, followed by a 100ms ISI. Then the third sentence was presented word by word (each word for 450ms, 100ms ISI).

Participants' task was to judge whether or not the scenario "made sense" by pressing one of two buttons after seeing a "?", which appeared after each scenario (1400ms with a 100ms ISI). Response fingers were counterbalanced across participants. This task encouraged active coherence monitoring during online comprehension and was intended to ensure that comprehenders detected the anomalies, which is necessary to produce a late posterior positivity/P600 response (Sanford et al. 2011). In addition, following approximately 24/200 trials (distributed semi-randomly), participants answered a "YES/NO" comprehension question that appeared on the screen for 1900ms (100ms ISI). This encouraged participants to comprehend the scenarios as a whole, rather than focusing on only the third sentence in which the anomalies appeared.

Following each trial, a blank screen was presented with a variable duration that ranged from 100-500ms. This was then followed by a green fixation (++++) for a duration of 900ms, followed by an ISI of 100ms. These green fixations were used to estimate the noise covariance for the MEG source localization (see below). To ensure precise time-locking of stimuli, we used frame-based timing, which synced stimulus presentation to the frame refresh rate of the monitor (for example, a 450ms word presentation would be displayed for exactly 27 frames on our 60Hz monitor).

Stimuli were presented in eight blocks, each with 25 scenarios. Blocks were presented in random order in each participant. Participants took part in a short practice session before the formal experiment to gain familiarity with the stimulus presentation and tasks.

Data acquisition

MEG and EEG data acquisition

Participants sat inside a magnetically shielded room (IMEDCO AG, Switzerland). The MEG data were acquired with a Neuromag VectorView system (Elekta-Neuromag Oy, Finland) with 306 sensors — 102 triplets, with each triplet comprising two orthogonal planar gradiometers and one magnetometer. The EEG data were acquired at the same time using a 70-channel MEG-compatible scalp electrode system (BrainProducts, München), and referenced to an electrode placed on the left mastoid. An electrode was also placed on the right mastoid and a ground electrode was placed on the left collarbone. EOG data were collected with bipolar recordings: vertical EOG electrodes were placed above and below the left eye, and horizontal EOG electrodes were placed on the outer canthus of each eye. ECG data were also collected with bipolar recordings: ECG electrodes were placed a few centimeters under the left and right collarbones. Impedances were kept at <20kΩ at all scalp sites, at <10kΩ at mastoid sites, and at <30kΩ at EOG and ECG sites. Both MEG and EEG data were acquired with an online band-pass filter of 0.03-300Hz and were continuously sampled at 1000Hz.

To record the head position relative to the MEG sensor array for later co-registration of the MEG and MRI coordinate frames, the locations of three fiduciary points (nasion and two auricular), four head position indicator coils, all EEG electrodes, and at least 100 additional points, were digitized using a 3Space Fastrak Polhemus digitizer, integrated with the Vectorview system.

Structural MRI data acquisition

In order to create participant-specific head model for MEG source localization, we acquired structural MRIs using a 3T Siemens Trio scanner with a 32-channel head coil for all participants. T1-weighted high-resolution structural images were obtained using the following parameters: 1mm isotropic multi-echo magnetization-prepared rapid gradient-echo, MP-RAGE; Time to Repetition (TR): 2.53s; flip angle: 7 degrees; 4 echoes with TE: 1.69ms, 3.55ms, 5.41ms, and 7.27ms.

ERP Analysis

EEG data were analyzed using the Fieldtrip software package (Oostenveld et al. 2011) in the Matlab environment. EEG channels with excessive noise (7 out of the 70 channels, on average) were visually identified and marked as bad channels. We then applied a low band-pass filter (30Hz), down sampled the EEG data to 500Hz, and segmented the epochs from -2600ms to 1400ms, relative to the onset of the critical words. After that, we visualized the data in summary mode within the Fieldtrip toolbox to identify the trials that showed high variance across channels. These trials were then removed from subsequent analysis. We then carried out an Independent Component Analysis (ICA) to remove ICA components associated with eye-movement (one component on average was removed per participant). Finally, we visualized the artifact-corrected trials, and removed any additional trials with residual artifact. On average, 6.5% of trials were removed from each condition (equally distributed across the three conditions: $F(2,62) = 1.40$, $p = 0.26$, $\eta^2 = 0.048$), yielding, on average, 23 trials in the *expected* and *unexpected plausible* conditions, and 46 trials in the *implausible* condition. Finally, the data of bad channels were interpolated using spherical spline interpolation (Perrin et al. 1989). In each participant, at each site, we then calculated ERPs, time-locked to the onset of critical words, in each of the three conditions, applying a -100ms pre-stimulus baseline.

We next averaged these voltages across all time points and electrode sites within each of three spatiotemporal regions of interest (ROIs) that were selected, *a priori,* to capture the N400, the *late frontal positivity* and the *late posterior positivity/P600* ERP components (Kuperberg *et al.* 2020). The N400 was operationalized as the average voltage across ten electrode sites within a central region (Cz, C1, C2, C3, C4, CPz, CP1, CP2, CP3, CP4), averaged across all sampling points between 300-500ms; the *late frontal positivity* was operationalized as the average voltage across eight electrode sites within a prefrontal region (FPz, FP1, FP2, FP3, FP4, AFz, AF3, AF4), averaged across sampling points between 600-1000ms; the *late posterior positivity/P600* was operationalized as the average voltage across 11 electrode sites within a posterior region (Pz, P1, P2, P3, P4, POz, PO3, PO4, Oz, O1, O2), averaged across sampling points between 600-1000ms. We carried out planned statistical comparisons between each pair of conditions. We also report post-hoc ERP analyses across an earlier 200-300ms time-window in Supplementary Materials.

MEG Analysis

<u>MEG preprocessing, individual averaging and sensor-level visualization</u>

MEG data were analyzed using version 2.7.4 of the Minimum Norms Estimate (MNE) software package in Python (Gramfort et al. 2014). In each participant, in each run, MEG sensors with excessive noise were visually identified and removed from further analysis. This resulted in the removal of seven (on average) of the 306 MEG sensors. Signal-space projection (SSP) correction was used to correct for ECG artifact. Trials with eye-movement and blink artifacts were automatically removed (Gramfort *et al.* 2014). Then, after applying a band-pass filter at 0.1Hz to 30Hz, we segmented epochs from -100 to 1000ms, relative to the onset of the critical words. We removed epochs with additional artifact, as assessed using a peak-to-peak detection algorithm (the pre-specified cutoff for the maximal amplitude range was 4×10^{-10} T/m for the gradiometer sensors and 4×10^{-12} T for the magnetometer sensors). On average, 6.7% trials in each condition were removed (equally distributed across the three conditions: $F(2,62) = 0.87$, $p = 0.42$, $\eta^2 = 0.027$), yielding, on average, 21 artifact-free trials in the *expected* and *unexpected plausible* conditions, and 42 artifact-free trials in the *implausible* condition.

In each participant, in each block, at each magnetometer sensor and at each of the two gradiometers at each site, we calculated event-related fields (ERFs), time-locked to the onset of critical words in each of the three conditions, applying a -100ms pre-stimulus baseline. We averaged the ERFs across blocks in sensor space, interpolating any bad sensors using spherical spline interpolation (Perrin *et al.* 1989). We created gradiometer and magnetometer sensor maps to visualize the topographic distribution of ERFs across the scalp. In creating the gradiometer maps, we used the root mean square of the ERFs produced by the two gradiometers at each site.

MEG source localization in individual participants

Each participant's cortical surface was first reconstructed from their structural T1 MPRAGE image using the FreeSurfer software package developed at the Martinos Center, Charlestown, MA (http://surfer.nmr.mgh.harvard.edu). We used MNE-Python (Gramfort *et al.* 2014) to estimate the sources of the ERFs evoked by critical words in each of the three conditions, on each participant's reconstructed cortical surface using Minimum-Norm Estimation (MNE) (Hämäläinen and Sarvas 1989).

In order to calculate the inverse operator in each participant — the transformation that estimates the underlying neuroanatomical sources for a given spatial distribution of activity in sensor space — we first needed to construct a noise-covariance matrix of each participant's MEG sensor-level data, as well as a forward model in each participant (the model that predicts the pattern of sensor activity that would be produced by all dipoles within the source space).

To construct the noise covariance matrix in each participant, we used 650ms of MEG sensor-level data recorded during the presentation of the green inter-trial fixations (we used an epoch from 100-750ms, which cut off MEG data measured at the onset and offset of these fixations in order to avoid onset and offset evoked responses). We concatenated these fixations across blocks. To construct the forward model in each participant, we needed to (a) define the source space—the location, number and spacing of dipoles, (b) create a Boundary Element Model (BEM), which describes the geometry of the head and the conductivities of the different tissues, and (c) specify the MEG-MRI coordinate transformation—the location of MEG sensors in relation to the head surface.

The source space was defined on the white matter surface of each participant's reconstructed MRI and constituted 4098 vertices per hemisphere, with three orthogonally-orientated dipoles at each vertex (two tangential and one perpendicular to the cortical surface). We defined these vertices using a grid that decimated the surface into meshes, with a spacing of 4.9mm between adjacent locations (spacing: “oct6”). We created a single compartment BEM by first stripping the outer non-brain tissue (skull and scalp) from the pial surface using the watershed algorithm in FreeSurfer, and then applying a single conductivity parameter to all brain tissue bounded by the inner skull. We specified the location of the MEG sensors in relation to the head surface by manually aligning the fiducial points and 3D digitizer (Polhemus) data with the scalp surface triangulation created in FreeSurfer, using the mne_analyze tool (Gramfort *et al.* 2014).

We then calculated the inverse operator in each participant, setting two additional constraints. First, we set a loose constraint on the relative weighting of tangential and perpendicular dipole orientations within the source space (loose = 0.2). Second, we set a constraint on the relative weighting of superficial and deep neuroanatomical sources (depth = 0.8) in order to increase the likelihood that the minimum norm estimates would detect deep sources.

We then applied each participant’s inverse operator to the ERFs of all magnetometer and gradiometer sensors calculated within each block. We estimated activity at the dipoles that were orientated perpendicular to the cortical surface at each vertex (pick_ori = “normal”). Each of these perpendicular dipoles had both a positive and a negative value, which indicated whether the currents were outgoing or ingoing respectively. We chose to retain the two polarities of each estimated dipole for further analyses for two reasons. First, this approach allowed us to include all trials in each of our three conditions, thereby maximizing power without inflating our estimate of noise in the conditions with more trials (if we had chosen to simply estimate the magnitude of each dipole by squaring the positive and negative values to yield positively-signed estimates, we would have artificially inflated the noise estimates in the *implausible* conditions, which had twice as many trials as the *expected* and the *unexpected plausible* conditions). Second, by retaining this polarity information, we were able to determine whether any statistical differences between conditions were driven by differences in the magnitude and/or differences in the polarity of the dipoles evoked in each condition (see Supplementary Materials for further discussion).

Then, for each condition in each block, we computed noise-normalized dynamic Statistical Parametric Maps (dSPMs) (Dale *et al.* 2000) on each participant's cortical surface at each time point, and averaged these values across blocks within each participant. Finally, the source estimates for each participant were morphed on the FreeSurfer average brain, "fsaverage" (Fischl et al. 1999), for group averaging and statistical analysis.

Statistical analysis of MEG source-level evoked activity

To statistically analyze the source-localized evoked MEG responses, we carried out pairwise t-tests between each pair of conditions on the signed estimated dSPM values at each sampling point from 300-1000ms after critical word onset. Pairwise t-tests were calculated at each vertex in source space, over a large left-lateralized search region where we expected effects to be maximal (particularly the effects of lexical predictability, see Federmeier 2022). This search area included left lateral temporal cortex, left ventral temporal cortex, left medial temporal cortex, left lateral parietal cortex, left lateral frontal cortex, and left medial frontal cortex, and was defined on the Desikan-Killiany Atlas (Desikan et al. 2006), see Supplementary Figure 1A.

We used permutation-based cluster mass procedures (Maris and Oostenveld 2007) to correct for multiple comparisons in time and space. First, at each vertex and at each time point, any data points that exceeded a pre-set uncorrected significance threshold of 1% (i.e., $p \leq 0.01$) were -log10 transformed, and the rest were zeroed. The use of unsigned -log-transformed p-values allowed us to account for effects with the opposite polarities. A single neuroanatomical source that is located on one side of a sulcus can appear on the cortical surface as adjacent dipoles of opposite polarity (outgoing and ingoing currents) because of signal bleeding to the other side of the sulcus (Hämäläinen et al. 1993) (see Figures 6 and 7). The use of unsigned p-values therefore ensured that adjacent effects of opposite signs were treated as a single underlying source in the statistical analyses.

Second, to minimize multiple comparisons over time, we averaged the transformed *p*-values at each vertex within three *a priori* time-windows of interest: 300-500ms, corresponding to the N400 time-window, and 600-800ms and 800-1000ms, corresponding to the first and second halves of the time-window associated with late positive ERP effects. Then, to minimize multiple comparisons over space, these transformed *p*-values were further averaged across vertices within 140 spatial patches of approximately equal size (Khan et al. 2018), shown in Supplementary Figure 1B. This resulted in 140 separate test statistics in each time-window of interest, i.e. 140 spatiotemporally-clustered test statistics. This approach ensured that data points within the same time-window and spatial patch (i.e. each spatiotemporal cluster) were treated as originating from a single underlying source. We should note that, although this approach increases statistical power, it also constrains our statistical inference to the spatial resolution of each patch and to the temporal resolution of our *a priori* time-windows.

Finally, we computed a null distribution for our spatiotemporally-clustered test statistics, by carrying out exactly the same procedure as that described above, but this time randomly assigning condition labels within each participant, with 10,000 permutations. For each randomization, we took the largest value across all spatial patches as our cluster-mass statistic. To test our hypotheses, we compared each of the observed spatiotemporally-clustered test statistics against this null distribution. If an observed cluster-level statistic fell within the highest 5.0% of the distribution, we considered it to be significant. Note that this non-parametric cluster-based approach is robust to any differences in signal-to-noise resulting from different numbers of trials in the *expected*, *unexpected plausible* and *implausible* conditions.

When displaying the results, we show both the uncorrected and cluster-corrected results on the cortical surface. We projected the averaged -log10 transformed uncorrected p-values ($p < 0.05$) within each time window at each vertex on to the "fsaverage" brain (Fischl *et al.* 1999). We then grouped together all spatial patches that reached cluster-level significance into the neuroanatomical regions that are shown in Supplementary Figure 1A and listed in Supplementary Table 1 (defined using the Desikan-Killiany Atlas; Desikan *et al.* 2006). If one or more patches within a neuroanatomical region reached cluster-level significance, we circled the region in red.

We also carried out several additional analyses, which are reported in Supplementary Materials: (a) an analysis of a subset of the MEG data to address potential concerns regarding differences between the *expected* and *implausible* conditions in the number of trials per condition, and in the lexical properties of the critical words, (b) an analysis to determine whether there were effects in an earlier 200-300ms time-window, and (c) an analysis of an analogous search region within the right hemisphere.

Finally, in Supplementary Materials, we report an exploratory cross-ROI multivariate decoding analysis, which was intended to provide preliminary multivariate data to clarify or support certain points in our interpretation of the evoked effects, as discussed in the main manuscript.

Results

Behavioral

Participants correctly judged the plausibility of 89.42% scenarios (SD: 8.83%) and answered 80.12% (SD: 11.50%) of the comprehension questions correctly, suggesting that they were engaged in comprehension (see Supplementary Materials).

ERP

The ERP results (Figure 4, Table 2) replicate previous findings (Kuperberg *et al.* 2020). Between 300-500ms, the N400 amplitude increased across the three conditions. Between 600-1000ms, the *unexpected plausible* words produced a larger *late frontal positivity* than both the *expected* and *implausible* continuations, while the *implausible* words produced a larger *late posterior positivity/P600* than both the *expected* and *unexpected plausible* continuations.

**** Insert Figure 4 here ***

**** Insert Table 2 here ***

MEG

To facilitate a qualitative comparison between MEG sensor-level results and the scalp-recorded ERP results, we show the sensor-level MEG results in Figure 5 (we note, however, that MEG statistical analyses were carried out in source space). Between 300-500ms, the sensor-level MEG findings showed a similar graded increase in evoked activity, see Figure 5A (the evoked response to the *implausible* continuations was larger in MEG than in ERP, see Supplementary Materials for discussion). Between 600-1000ms, the *unexpected plausible* and *implausible* words produced larger responses than the *expected* words, and the topographic sensor maps revealed distinct patterns for each effect, see Figure 5B.

**** Insert Figure 5 here ***

Source-localized MEG: Evoked effects

300-500ms

As shown in Figure 6, between 300-500ms, *expected*, *unexpected plausible* and *implausible* words produced graded increases in activity within multiple regions of left lateral, ventral and medial temporal cortices. In medial temporal cortices, the effects were also driven by a dipole going in the opposite direction to the *expected* words. In addition, the *implausible* words produced a larger response than both the *expected* and *unexpected plausible* words within left inferior frontal cortex, as well as a larger response than the *expected* words within anterior cingulate cortex.

**** Insert Figure 6 here ***

600-1000ms

As shown in Figure 7, by 500ms, the activity produced by the *unexpected plausible* words within left temporal cortex had diminished. Between 600-1000ms, however, these continuations produced a response within left inferior frontal cortex, and re-activated the left middle temporal cortex, with a dipole going in the opposite direction to that produced in the N400 time-window.

The evoked activity produced by the *implausible* words was quite different. By 500ms, the response produced by these continuations within the left inferior frontal cortex had diminished, but the activity produced within left temporal and posterior fusiform cortices continued into the 600-1000ms time-window. In the latter half of this time-window, the *implausible* words also re-activated the left inferior frontal cortex, with a dipole going in the opposite direction to that produced between 300-500ms. Finally, throughout the 600-100ms window, the *implausible* words also produced a large dipole within left medial temporal cortex, again with the opposite polarity to that produced in the N400 window.

Pair-wise statistical comparisons showed that, within left inferior frontal cortex, there was a significant difference in comparing both the *unexpected plausible* and the *implausible* words with the *expected* words (600-800ms), with both effects being driven by dipoles going in opposite directions in the two conditions. A direct comparison between the *unexpected plausible* and *implausible* conditions, however, revealed no differences within left inferior frontal cortex. Within left lateral temporal cortex, there were significant differences in comparing the *unexpected plausible* words with both the *expected* (800-1000ms) and *implausible* words (600-800ms). Both these effects were driven by dipoles going in the opposite direction in each of the two conditions. In posterior fusiform cortex there was a significant difference in comparing the *implausible* words with both the *expected* (600-800ms) and *unexpected plausible* words *(600-800ms; 800-1000ms).* Similarly, within medial temporal cortex, the response produced by the *implausible* words differed from that produced by both other conditions (600-800ms; 800-1000ms).

**** Insert Figure 7 here ***

Discussion

We used MEG and EEG to track the spatiotemporal dynamics of evoked activity produced by *expected*, *unexpected plausible*, and *implausible* words during language comprehension. At the scalp surface, our ERP findings replicate previous studies by showing that, between 300-500ms, the three conditions produced progressively larger N400 responses (Kuperberg *et al.* 2020;

Nieuwland *et al.* 2020), and that, between 600-1000ms, *unexpected plausible* and *implausible* continuations produced two spatially-distinct late positivities (Van Petten and Luka 2012; DeLong *et al.* 2014; Brothers *et al.* 2020; Kuperberg *et al.* 2020). By simultaneously collecting MEG data and source-localizing the evoked response in both time windows, we were able to show, for the first time *where* in the brain these effects are produced.

Several previous MEG (e.g. Helenius *et al.* 1998; Halgren *et al.* 2002; Maess *et al.* 2006; Ihara *et al.* 2007) and intracranial (McCarthy et al. 1995) studies have reported effects of sentence context on the N400 within temporal and/or inferior frontal cortices. However, most of these previous studies have directly contrasted *expected* and *implausible* words, without dissociating the effects of lexical predictability and contextual plausibility. It has therefore been difficult to directly link activity produced at lower and higher levels of the fronto-temporal hierarchy to processing at different levels of representation (mapping between word-forms, semantic features and real-world knowledge). In addition, no previous MEG or intracranial study has source localized the evoked effects of semantic context beyond the N400 time window. By determining precisely when and how evoked activity across the fronto-temporal hierarchy is modulated following the onset of incoming words in both time windows, our findings shed new light on the cognitive architecture and neurobiology of online language comprehension.

300-500ms

The N400 effect of lexical predictability localizes to the left temporal cortex

In our ERP recordings, the N400 was smaller to the *expected* than to the *unexpected plausible* words, replicating many previous ERP studies (e.g. Kutas and Hillyard 1984; DeLong *et al.* 2005; Federmeier *et al.* 2007). MEG source-localization showed that this effect of lexical predictability localized to multiple regions within the left temporal cortex (lateral, ventral, and medial), suggesting that the process of mapping form onto meaning (lexico-semantic processing) was "easier" for *expected* than for *unexpected plausible* words. More specifically, in left ventral temporal regions, processing of *expected* (vs. *unexpected*) words was facilitated within the left posterior occipitotemporal fusiform cortex (orthographic-level facilitation, cf. Price and Devlin 2011) and the left mid-fusiform cortex (lexical-level facilitation, cf. Hirshorn et al. 2016; Woolnough et al. 2021). In left lateral temporal regions, processing was facilitated within the left superior temporal cortex (phonological-level facilitation; Solomyak and Marantz 2009; Vartiainen et al. 2009) and the left middle temporal cortex (lexical-level facilitation; Lau *et al.* 2008). Finally, the smaller dipole to the *expected* than the *unexpected* words in the left medial temporal cortex and in the anterior ventral temporal pole (bilaterally) may have reflected the reduced need to retrieve and "bind" distributed semantic features into distinct concepts (Lambon-Ralph et al. 2017).

In addition to producing a smaller dipole to the *expected* (versus *unexpected*) words, the medial temporal cortex also produced a larger dipole, with the opposite polarity, to the *expected* words. This is consistent with previous intracranial studies showing that, within medial temporal regions, both predictable and unpredictable words produce local field potentials in distinct population of neurons in the N400 time-window (McCarthy *et al.* 1995). We speculate that the dipole to the *expected* inputs indexed the recognition of an item-specific match (cf. Duncan et al. 2009) between the incoming semantic information and a pre-activated concept. The presence of two dipoles to *expected* and *unexpected* inputs, going in opposite directions, may explain why most previous MEG studies, which have typically used absolute (rather than signed) values for source localization, have failed to observe N400 effects in the medial temporal lobe (see Supplementary Materials for further discussion).

Critically, in this same N400 time window, we observed no effect of lexical predictability in the plausible sentences within inferior frontal regions. This result suggests that the effect of lexical predictability on the N400 largely reflects facilitation at the lower, lexical-semantic level (DeLong *et al.* 2005; Lau *et al.* 2016; Kuperberg *et al.* 2020).

The N400 effect of contextual plausibility localizes to both temporal and inferior frontal cortices

In ERPs, we also observed a smaller N400 to *unexpected plausible* words, relative to *implausible* words, again replicating previous findings (Kuperberg *et al.* 2020; Nieuwland *et al.* 2020). We note that this effect of contextual plausibility on the scalp-recorded N400 was much larger in MEG than in ERP, which has important implications for the functional interpretation of the N400 in the presence of an overlapping late posterior positivity/P600 (see Supplementary Materials for discussion). In MEG, the effect of contextual plausibility localized to *both* temporal and prefrontal cortices.

Within the left temporal cortex, the effect of contextual plausibility broadly localized to the same regions as the effects of lexical predictability, described above. Supplementary analyses also revealed some plausibility effects in homologous regions of the right temporal lobe (lateral and medial, see Supplementary Materials). Again, we interpret these smaller evoked responses within the temporal cortex as reflecting facilitated processing at the lexico-semantic level. However, in this case, instead of resulting from the pre-activation of a specific upcoming word, this facilitation resulted from the pre-activation of distributed features associated with a broad semantic category (i.e., animacy-based features; Wang et al. 2020). Thus, lexico-semantic processing was "easier" for the *plausible unexpected* inputs, where some semantic features were pre-activated, than for the *implausible* inputs where no semantic features were pre-activated (Paczynski and Kuperberg 2011, 2012; Kuperberg *et al.* 2020).

In the prefrontal cortex, the effect of contextual plausibility localized to the inferior portion of the left frontal and orbitofrontal cortices, as well as to their right hemisphere homologues (see Supplementary Materials). We interpret these smaller inferior frontal evoked responses to the *unexpected plausible* (vs. the *implausible*) continuations as reflecting the relative ease of mapping *plausible* event representations on to longer-term real-world knowledge.

Architectural implications: Towards a predictive coding framework

In the Introduction, we laid out four general frameworks of language comprehension. While each of the first three frameworks can explain a subset of our findings in the N400 time window, they cannot explain the full set of results across *both* temporal and inferior frontal cortices. For instance, both the *distributed state* and the *lexico-semantic facilitation* accounts can explain the graded N400 modulation across the three conditions within the temporal cortex. However, the *distributed state* account would have predicted similar graded effects within the inferior frontal cortex (Figure 1A), while the *lexico-semantic facilitation* account would have predicted *no* prefrontal N400 modulation at all (Figure 1B). The *prediction-integration* account *can* explain why the N400 in the inferior frontal cortex was selectively enhanced to *implausible* continuations. However, this framework would have incorrectly predicted *non-graded* N400 modulation across the three conditions within the temporal cortex, with an attenuation of the N400 only to the *expected* continuations (Figure 1C).

We argue that the *full* pattern of evoked activity produced between 300-500ms across the fronto-temporal hierarchy can be explained by the computational principles of hierarchical predictive coding (Figure 1D). We emphasize that this predictive coding framework shares several important features with the other three neurobiological models. Like these other frameworks, it assumes that language comprehension is both interactive and incremental (cf. Marslen-Wilson

1987; Altmann and Steedman 1988; Marslen-Wilson et al. 1988; MacDonald *et al.* 1994; Tanenhaus and Trueswell 1995), with the prior context influencing the initial feedforward sweep of evoked activity produced by incoming words between 300-500ms (the N400 response). Also similar to these other models, it assumes that, in the 300-500ms time window, information is continually transferred across temporal and inferior frontal cortices in both predictable and unpredictable sentences (see Baggio and Hagoort 2011, and see Lyu et al. 2019; Mamashli et al. 2019; Liu et al. 2020 for recent empirical evidence). Finally, similar to both the *prediction-integration* and the *lexico-semantic facilitation* accounts, predictive coding posits a hierarchical organization of representations across the fronto-temporal hierarchy. During discourse comprehension, we assume that comprehenders incrementally built a higher-level *event model* that was maintained over a relatively long time scale within the prefrontal cortex, and that interacted both with real-world knowledge, represented over a still longer time-scale, as well as with lower-level lexico-semantic representations, encoded within the temporal cortex at a shorter time-scale.

From a computational perspective, however, there is a key difference between predictive coding and other frameworks. While prior computational models of the N400 have assumed that evoked activity reflects changes in the "state" of neural activity at a given level of the cortical hierarchy (e.g. Brouwer *et al.* 2017; Rabovsky *et al.* 2018), predictive coding posits that the magnitude of the evoked response reflects the magnitude of prediction error, i.e. activity produced within a distinct set of "error units". At each level of the cortical hierarchy, these error units *only* encode information that is not suppressed (or "explained") by predictions produced by state units at the cortical level above (Rao and Ballard 1999; Friston 2005). Thus, predictive coding provides an intuitive biological explanation for the origin of the N400 neural response: Specifically, new unpredicted inputs trigger increased firing and post-synaptic potentials within these error units, which, in turn, results in a larger evoked N400 response.

Within the *temporal* cortex, predictive coding attributes the graded increases in N400 amplitude across the three conditions (*expected* < *unexpected plausible* < *implausible*) to graded increases in the magnitude of prediction error at the *lexico-semantic* level. In the case of *expected* continuations, lexico-semantic prediction error is fully suppressed by prior top-down predictions of the semantic features of specific upcoming words; in the case of *unexpected plausible* continuations, lexico-semantic prediction error is partially suppressed by prior top-down predictions of animacy-linked semantic features (see Wang *et al.* 2020), and, in the case of *implausible* continuations, lexico-semantic prediction error is not suppressed at all.

Within the *inferior frontal cortex*, predictive coding attributes the *non-graded* modulation of the N400 across the three conditions (*expected* = *unexpected plausible* < *implausible*) to differences in magnitude of *higher-level* prediction error produced at the level of the event model. According to this framework, there was no difference in the higher-level prediction error/evoked N400 activity to the *expected* and *unexpected plausible* words in inferior frontal cortex because, in both these conditions, information encoded in the plausible higher-level event model was explained/suppressed by predictions from still longer-term real-world knowledge. Despite the absence of an inferior frontal N400 effect for this contrast, unpredicted information was still shared between the left temporal and inferior frontal cortices in this time window, as evidenced by a small but significant cross-ROI decoding effect (see Supplementary Materials). In contrast, when the *implausible* continuations produced updates of the higher-level event model, these could not be explained/suppressed by predictions based on real-world knowledge, giving rise to a higher-level prediction error and an enhanced evoked response within this region (see Figure 2, right). This explanation follows directly from Rao and Ballard's discussion in their original description of hierarchical predictive coding in the visual system (Rao & Ballard, 1999): So long as newly encoded unpredicted input is consistent with the more general statistics of natural environmental inputs, it should *not* result in a large *increase* in activity within higher-level regions because any higher-level prediction error is continually suppressed. However, when "the statistics differ in certain drastic ways from natural statistics" (Rao & Ballard, 1999, page 84), then this will lead to a larger response within higher level regions of the cortical hierarchy.

Finally, predictive coding also posits that an inability to converge on a plausible interpretation in the prefrontal cortex should result in less accurate top-down predictions, which will fail to fully "switch off" lower-level lexico-semantic prediction error within the temporal cortex within the N400 time-window (see Figure 2, right). This failure to suppress lexico-semantic prediction error may have also contributed to the enhanced N400 response produced by the *implausible* (relative to the *unexpected plausible*) words within the temporal cortex, as well to a prolongation of this effect into the later time window.

600-1000ms

Replicating previous ERP findings, *unexpected plausible* and *implausible* words produced two distinct late positivities between 600-1000ms: a *late frontal positivity* to the *unexpected plausible* words (Federmeier *et al.* 2007; Van Petten and Luka 2012; DeLong *et al.* 2014; Brothers *et al.* 2020; Kuperberg *et al.* 2020), and a *late posterior positivity/P600* to the highly *implausible* continuations (Kuperberg 2007; van de Meerendonk *et al.* 2009; Van Petten and Luka 2012; DeLong *et al.* 2014; Brothers *et al.* 2020; Kuperberg *et al.* 2020; Brothers *et al.* 2021). Once again, our source-localized MEG results constrain our understanding of the functional significance of this late-stage evoked activity.

Late evoked effects to unexpected plausible words within left inferior frontal and middle temporal *cortices*

It has been proposed that the *late frontal positivity* evoked by *unexpected plausible* words reflects a successful high-level shift in the event model, together with feedback to the lower lexico-semantic level (Brothers *et al.* 2015; Brothers *et al.* 2020; Kuperberg *et al.* 2020). Our MEG findings are consistent with this account. Between 600-1000ms, *unexpected plausible,* relative to *expected,* continuations produced late evoked effects within both left inferior frontal and left middle temporal cortices.

Hierarchical predictive coding offers a computational-level explanation for this late-stage activity. Within this framework, these late evoked responses are attributed to "top-down error" — that is, activity produced by *new* top-down predictions at a given level of the hierarchy that cannot be explained by its prior state (Rao and Ballard 1997, 1999). More specifically, according to this framework, if an unpredicted incoming word leads the brain to retrieve new schema-relevant information from long-term memory (cf. Franklin *et al.* 2020, e.g. infer what trainees might be doing in a beach scenario), then this will result in the generation of *new* schema-relevant predictions (Kuperberg 2021) that are propagated down the cortical hierarchy. When these new top-down predictions activate the left inferior frontal cortex, they will produce top-down error at the level of the event model, explaining the larger late evoked response to the *unexpected plausible* continuations within this region, and when they reach the left temporal cortex they will produce top-down error at the lexico-semantic level, explaining the larger late evoked response within the left middle temporal cortex (see Figure 3, left). In the visual system, it has been proposed that, following an initial bottom-up sweep of activity, this type of feedback re-activation ensures that lower-level regions encode information that is consistent with global gestalt representations that are encoded in higher cortical areas (Lee and Mumford 2003).

Consistent with this interpretation, the late evoked effect within the left middle temporal cortex was driven by a dipole with the *opposite* polarity to the dipole produced within the left temporal cortex in the earlier N400 time-window. While the precise significance of this dipole reversal is unclear (see Supplementary Materials for further discussion), it provides evidence that this later evoked effect (“top-down” lexico-semantic error) was functionally distinct from the earlier stimulus-driven evoked N400 effects observed between 300-500ms (“bottom-up” lexico-semantic prediction error). This dipole reversal also provides evidence that this late evoked activity does not simply reflect a response to the subsequent word. Finally, an exploratory multivariate analysis revealed a small but significant above-chance cross-ROI decoding effect between left frontal and left temporal regions, suggesting that the unexpected information was indeed shared between these two brain regions within this late time window (see Supplementary Materials).

Late evoked effects to highly implausible words within posterior fusiform, inferior frontal and medial temporal cortices

In the present study, the *implausible* words were not simply implausible — they were also anomalous (e.g. “*cautioned the *drawers”*); that is, they *conflicted* with the state of the hierarchical generative model as a whole. This conflict may explain why, relative to *expected* words, the anomalies activated the anterior cingulate cortex in the earlier N400 time-window (Botvinick 2007; see also Ide et al. 2013).

It has been proposed that the *late posterior positivity/P600* produced by highly *implausible* continuations between 600-1000ms is linked to a conflict-driven reprocessing at lower levels of linguistic representation (van de Meerendonk *et al.* 2009; Brothers *et al.* 2020; Kuperberg *et al.* 2020; Brothers *et al.* 2021). Consistent with this theory, the *implausible* words produced a robust late evoked effect within the posterior (occipitotemporal) fusiform cortex – the so-called “visual word-form area" – that supports orthographic processing (Price and Devlin 2011; Heilbron, Richter*, et al.* 2020) (see Supplementary Materials for additional discussion on the relationship between the ERP and MEG evoked effects in this late time-window).

Predictive coding again provides a mechanistic account of this lower-level orthographic reprocessing (see Figure 3, right). Within this framework, the late evoked activity within the posterior fusiform cortex is attributed to the production of prediction error at a still lower orthographic level of representation (Price and Devlin 2011) because higher cortical levels failed to generate accurate top-down predictions that would have otherwise suppressed this low-level error within this late time window. Specifically, after inferring an anomalous event (e.g. <*lifeguards cautioned *drawers*>) between 300-500ms, it was not possible to shift the event model by retrieving new stored schemas from long-term memory. Therefore, between 600-1000ms, incorrect predictions based on the prior context and real-world knowledge would have continued to be propagated down the cortical hierarchy. Within the left middle temporal cortex, these top-down lexico-semantic predictions (e.g. *"swimmers"* <*animate*>) would have been incompatible with the lexico-semantic information that was inferred from the bottom-up input (e.g. *"drawers"* <*inanimate*>), resulting in a destabilization of the lexico-semantic state. As a result of this destabilization, the left middle temporal cortex would have produced noisy, inaccurate orthographic predictions. Upon reaching the posterior fusiform cortex, these predictions would have failed to suppress orthographic prediction error, leading to the enhanced evoked response within this region (orthographic reprocessing).

Predictive coding can also explain why, in this late time window, the anomalous continuations re-activated the left inferior frontal cortex, with a dipole of the opposite polarity to that produced by these continuations in the N400 time-window. Within this framework, this late inferior frontal evoked response reflected top-down error produced the level of the event model when the inaccurate predictions, based on real-world knowledge (<*lifeguards cautioned swimmers*>), failed to explain the implausible event (<*lifeguards cautioned drawers*>) that had been inferred within the N400 time-window. Indeed, an exploratory cross-ROI decoding analysis revealed *no* evidence of shared information between left inferior frontal and lateral temporal/fusiform regions within this later time window (see Supplementary Materials).

Finally, the *implausible* continuations also produced a dipole within the medial temporal cortex throughout the 600-1000ms time-window, again with the opposite polarity to that produced in the N400 time-window. We speculate that this medial temporal activity supported new learning/adaptation, which was triggered by the failure of the current generative model to explain the input — that is, to minimize error across the cortical hierarchy (see further below).

Open Questions

We have argued that the full time-course of evoked activity produced within temporal and inferior frontal cortices in response to *expected*, *unexpected plausible* and *implausible* words, can be understood within a hierarchical predictive coding framework of language comprehension. This interpretation leaves many open questions that will be important to address in future studies.

First, it will be important for future studies to *parametrically* manipulate lexical predictability and plausibility in order to determine precisely how these factors modulate evoked activity within left temporal and inferior frontal cortices at both early and later stages of processing. It will also be important to understand whether and how evoked activity in both time windows is modulated by the lexical and discourse constraint of the prior context. For example, according to predictive coding, between 300-500ms, evoked activity produced by *unexpected plausible* words within left temporal cortex should *not* be modulated by prior contextual constraint, consistent with scalp-recorded N400 findings (Kutas and Hillyard 1984; Federmeier *et al.* 2007; Kuperberg *et al.* 2020).

Second, consistent with a predictive coding architecture, we have assumed that, at each level of the cortical hierarchy, the amplitude of the evoked response largely reflects the magnitude of prediction error – activity within populations of error units that cannot be explained/suppressed by the level above (Friston 2005). A key claim of predictive coding is that these error units are computationally distinct from “state units”, which encode representational information, *regardless* of its predictability. As such, this framework predicts that, even if new information does *not* produce an increased evoked response, we should still be able to decode this information using multivariate methods, which are sensitive to representational information, regardless of its magnitude. It will therefore be important for future studies to test this hypothesis directly by using a combination of univariate and multivariate methods. For example, future studies should follow up our preliminary finding, reported in Supplementary Materials, that, within the left inferior frontal cortex, despite failing to produce a larger N400 response (higher-level prediction error) between 300-500ms, it was still possible to decode new *unexpected plausible* information within this region in this same time window. Similarly, future studies should test the hypothesis that, within the left temporal cortex, despite producing a very small evoked N400 response, it should be possible to decode *expected* information within the N400 time window (see Kok et al. 2012; Bell et al. 2016 for evidence of dissociations between univariate and multivariate activity to expected inputs in low-level visual perception).

A third and related set of questions concerns the nature of information *flow* across the cortical hierarchy. Like most other neurobiological frameworks of language comprehension, predictive coding assumes that, in plausible sentences, *both* unexpected and expected information are shared across multiple regions of the cortical hierarchy between 300-500ms (e.g. Baggio and Hagoort 2011, Lyu *et al.* 2019; Mamashli *et al.* 2019; Liu *et al.* 2020). What remains unclear is precisely *how* this information is transmitted between regions. For example, some researchers have hypothesized that different frequency bands of oscillatory neural activity carry top-down predictions (e.g. slow beta/alpha) and bottom-up prediction error (e.g. fast gamma) (Arnal and Giraud 2012; Bastos et al. 2015; Lewis and Bastiaansen 2015; Bastos et al. 2020). It will be important for future studies to test this hypothesis more explicitly.

Finally, an important set of open questions concerns the relationship between evoked neural activity and longer-term learning, particularly given the robust effects observed in the medial temporal cortex in both the 300-500ms and 600-1000ms time windows. In the present study, we focused on the role of evoked activity (prediction error) in relation to inference, i.e. the process of online language *comprehension*. However, prediction error is also thought to play a critical role in learning in both linguistic (Elman 1990; Dell and Chang 2014) and non-linguistic (Rescorla 1988) domains. Recent computational models, using standard connectionist architectures, have shown that N400 prediction errors can, in principle, drive incremental learning (see Rabovsky and McRae 2014; Rabovsky *et al.* 2018; Fitz and Chang 2019; see also footnote 2 in Introduction). Predictive coding offers a biologically plausible algorithm for instantiating this type of longer-term learning (Rao and Ballard 1999; Whittington and Bogacz 2019; see Nour Eddine *et al.* in press for discussion).

In addition to this relatively slow cortical adaptation, which may be linked to the N400, it is also possible that linguistic anomalies may trigger a more rapid form of learning that is often discussed in the P600 literature (e.g. Coulson et al. 1998; Hanulikova et al. 2012). Notably, in the present study, semantic anomalies produced a distinct reverse-dipole effect in the 600-1000ms time window within medial temporal cortices (see McClelland et al. 1995; O'Reilly and Rudy 2001 for more general discussion of distinct modes of medial temporal function in relation to slower *versus* faster learning). In future studies, it will be important to determine whether and how evoked activity in the medial temporal lobe, in both early and late time-windows, influences different types of longer-term learning.

Conclusions

Of course, no single study can provide definitive evidence for any single model of language comprehension. As we have noted, several of our individual findings, including the graded increase in N400 activity within the temporal cortex, and reprocessing of anomalies within posterior fusiform cortex, are also consistent with other psycholinguistic or neurobiological models. Here, we have interpreted the full pattern of findings within a single computational framework – predictive coding – which has been proposed as a unifying theory of brain function, in multiple domains of perception and cognition (Clark 2013), including lower-level aspects of language processing (Price and Devlin 2011; Blank and Davis 2016; Sohoglu and Davis 2020). Our findings suggest that the computational principles of predictive coding may also explain the time course of evoked activity produced across the fronto-temporal network that supports higher-level language comprehension.

Funding

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. This research was carried out at the Athinoula A. Martinos Center for Biomedical Imaging at the Massachusetts General Hospital, using resources provided by the Center for Functional Neuroimaging Technologies, P41EB015896, a P41 Biotechnology Resource Grant supported by the National Institute of Biomedical Imaging and Bioengineering (NIBIB), National Institutes of Health. This work also involved the use of instrumentation supported by the NIH Shared Instrumentation Grant Program, specifically, grant numbers S10RR014978 and S10RR021110.

Acknowledgements

We thank Nao Matsuda for her technical MEG support, as well as Edward Wlotko, Nate Delaney-Busch, Eric Fields, Allison Fogel and Arim Choi Perrachione for their assistance with data collection. We also thank Seppo Ahlfors for his help with data analysis, and Jasmine Falk and Rebeca Becdach for their help in data visualization. Finally, we thank Jeff Stibel for his support of Drs. Kuperberg and Wang.

References

Ahlfors SP, Han J, Belliveau JW, Hämäläinen MS. 2010. Sensitivity of MEG and EEG to source orientation. Brain Topogr. 23:227-232.

Altmann GT, Steedman M. 1988. Interaction with context during human sentence processing. Cognition. 30:191-238.

Arnal LH, Giraud AL. 2012. Cortical oscillations and sensory predictions. Trends in Cognitive Sciences. 16:390-398.

Baggio G, Hagoort P. 2011. The balance between memory and unification in semantics: A dynamic account of the N400. Lang Cogn Process. 26:1338-1367.

Bastos AM, Litvak V, Moran R, Bosman CA, Fries P, Friston KJ. 2015. A DCM study of spectral asymmetries in feedforward and feedback connections between visual areas V1 and V4 in the monkey. Neuroimage. 108:460-475.

Bastos AM, Lundqvist M, Waite AS, Kopell N, Miller EK. 2020. Layer and rhythm specificity for predictive routing. Proc Natl Acad Sci U S A. 117:31459-31469.

Bell AH, Summerfield C, Morin EL, Malecek NJ, Ungerleider LG. 2016. Encoding of stimulus probability in macaque inferior temporal cortex. Curr Biol. 26:2280-2290.

Blank H, Davis MH. 2016. Prediction errors but not sharpened signals simulate multivoxel fMRI patterns during speech perception. PLoS Biology. 14:e1002577.

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

Botvinick MM. 2007. Conflict monitoring and decision making: Reconciling two perspectives on anterior cingulate function. Cogn Affect Behav Neurosci. 7:356-366.

Brothers T, Swaab TY, Traxler MJ. 2015. Effects of prediction and contextual support on lexical processing: prediction takes precedence. Cognition. 136:135-149.

Brothers T, Wlotko EW, Warnke L, Kuperberg GR. 2020. Going the extra mile: Effects of discourse context on two late positivities during language comprehension. Neurobiol Lang. 1:135-160.

Brothers T, Zeitlin M, Choi Perrachione A, Choi C, Kuperberg G. 2021. Domain-general conflict monitoring predicts neural and behavioral indices of linguistic error processing during reading comprehension. J Exp Psychol Gen.

Brouwer H, Crocker MW, Venhuizen NJ, Hoeks JCJ. 2017. A neurocomputational model of the N400 and the P600 in language processing. Cogn Sci. 41 Suppl 6:1318-1352.

Brouwer H, Hoeks JC. 2013. A time and place for language comprehension: mapping the N400 and the P600 to a minimal cortical network. Front Hum Neurosci. 7:758.

Brysbaert M, Warriner AB, Kuperman V. 2014. Concreteness ratings for 40 thousand generally known English word lemmas. Behav Res Methods. 46:904-911.

Clark A. 2013. Whatever next? Predictive brains, situated agents, and the future of cognitive science. Behav Brain Sci. 36:181-204.

Coulson S, King JW, Kutas M. 1998. Expect the unexpected: Event-related brain responses to morphosyntactic violations. Lang Cogn Process. 13:21-58.

Dale AM, Liu AK, Fischl BR, Buckner RL, Belliveau JW, Lewine JD, Halgren E. 2000. Dynamic statistical parametric mapping: combining fMRI and MEG for high-resolution imaging of cortical activity. Neuron. 26:55-67.

Dell GS, Chang F. 2014. The P-chain: relating sentence production and its disorders to comprehension and acquisition. Philosophical Transactions of the Royal Society B: Biological Sciences. 369:20120394.

DeLong KA, Quante L, Kutas M. 2014. Predictability, plausibility, and two late ERP positivities during written sentence comprehension. Neuropsychologia. 61C:150-162.

DeLong KA, Urbach TP, Kutas M. 2005. Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. Nature Neuroscience. 8:1117-1121.

Desikan RS, Segonne F, Fischl B, Quinn BT, Dickerson BC, Blacker D, Buckner RL, Dale AM, Maguire RP, Hyman BT, Albert MS, Killiany RJ. 2006. An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. Neuroimage. 31:968-980.

Duncan K, Curtis C, Davachi L. 2009. Distinct memory signatures in the hippocampus: intentional States distinguish match and mismatch enhancement signals. J Neurosci. 29:131-139.

Elman JL. 1990. Finding structure in time. Cognitive Science. 14:179-211.

Federmeier KD. 2022. Connecting and considering: Electrophysiology provides insights into comprehension. Psychophysiology. 59:e13940.

Federmeier KD, Wlotko EW, De Ochoa-Dewald E, Kutas M. 2007. Multiple effects of sentential constraint on word processing. Brain Res. 1146:75-84.

Fischl B, Sereno MI, Dale AM. 1999. Cortical surface-based analysis. II: Inflation, flattening, and a surface-based coordinate system. Neuroimage. 9:195-207.

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

Franklin NT, Norman KA, Ranganath C, Zacks JM, Gershman SJ. 2020. Structured event memory: a neuro-symbolic model of event cognition. Psychol Rev. 127:327-361.

Friston KJ. 2005. A theory of cortical responses. Philos Trans R Soc Lond B Biol Sci. 360:815-836.

Furey ML, Tanskanen T, Beauchamp MS, Avikainen S, Uutela K, Hari R, Haxby JV. 2006. Dissociation of face-selective cortical responses by attention. Proc Natl Acad Sci U S A. 103:1065-1070.

Gershman SJ, Radulescu A, Norman KA, Niv Y. 2014. Statistical computations underlying the dynamics of memory updating. PLoS Computational Biology. 10:e1003939.

Geukes S, Huster RJ, Wollbrink A, Junghöfer M, Zwitserlood P, Dobel C. 2013. A large N400 but no BOLD effect–comparing source activations of semantic priming in simultaneous EEG-fMRI. PLoS One. 8:e84029.

Gramfort A, Luessi M, Larson E, Engemann DA, Strohmeier D, Brodbeck C, Parkkonen L, Hämäläinen MS. 2014. MNE software for processing MEG and EEG data. Neuroimage. 86:446-460.

Hagoort P, Hald L, Bastiaansen M, Petersson KM. 2004. Integration of word meaning and world knowledge in language comprehension. Science. 304:438-441.

Halgren E, Dhond RP, Christensen N, Van Petten C, Marinkovic K, Lewine JD, Dale AM. 2002. N400-like magnetoencephalography responses modulated by semantic context, word frequency, and lexical class in sentences. Neuroimage. 17:1101-1116.

Hämäläinen MS, Hari R, Ilmoniemi RJ, Knuutila JET, Lounasmaa OV. 1993. Magnetoencephalography—theory, instrumentation, and applications to noninvasive studies of the working human brain. Rev Modern Physics. 65:413-497.

Hämäläinen MS, Sarvas J. 1989. Realistic conductivity geometry model of the human head for interpretation of neuromagnetic data. IEEE Trans Biomed Eng. 36:165-171.

Hanulikova A, van Alphen PM, van Goch MM, Weber A. 2012. When one person's mistake is another's standard usage: The effect of foreign accent on syntactic processing. J Cogn Neurosci. 24:878-887.

Heilbron M, Armeni K, Schoffelen J-M, Hagoort P, de Lange FP. 2020. A hierarchy of linguistic predictions during natural language comprehension. In. bioRxiv.

Heilbron M, Richter D, Ekman M, Hagoort P, de Lange FP. 2020. Word contexts enhance the neural representation of individual letters in early visual cortex. Nat Commun. 11.

Helenius P, Salmelin R, Service E, Connolly J. 1998. Distinct time courses of word and context comprehension in the left temporal cortex. Brain. 121:1133-1142.

Hirshorn EA, Li Y, Ward MJ, Richardson RM, Fiez JA, Ghuman AS. 2016. Decoding and disrupting left midfusiform gyrus activity during word reading. Proc Natl Acad Sci USA. 113:8162-8167.

Ide JS, Shenoy P, Yu AJ, Li CS. 2013. Bayesian prediction and evaluation in the anterior cingulate cortex. J Neurosci. 33:2039-2047.

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

Khan S, Hashmi JA, Mamashli F, Michmizos K, Kitzbichler MG, Bharadwaj H, Bekhti Y, Ganesan S, Garel KA, Whitfield-Gabrieli S, Gollub RL, Kong J, Vaina LM, Rana KD, Stufflebeam SM, Hamalainen MS, Kenet T. 2018. Maturation trajectories of cortical resting-state networks depend on the mediating frequency band. Neuroimage. 174:57-68.

Kok P, Jehee JF, de Lange FP. 2012. Less is more: expectation sharpens representations in the primary visual cortex. Neuron. 75:265-270.

Kuperberg GR. 2007. Neural mechanisms of language comprehension: Challenges to syntax. Brain Res. 1146:23-49.

Kuperberg GR. 2021. Tea with milk? A Hierarchical Generative Framework of sequential event comprehension. Top Cogn Sci. 13:256-298.

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

Kuperberg GR, Jaeger TF. 2016. What do we mean by prediction in language comprehension? Lang Cogn Neurosci. 31:32-59.

Kuperberg GR, Sitnikova T, Caplan D, Holcomb PJ. 2003. Electrophysiological distinctions in processing conceptual relationships within simple sentences. Brain Res Cogn Brain Res. 17:117-129.

Kutas M, Federmeier KD. 2011. Thirty years and counting: Finding meaning in the N400 component of the event-related brain potential (ERP). Annu Rev Psychol. 62:621-647.

Kutas M, Hillyard SA. 1984. Brain potentials during reading reflect word expectancy and semantic association. Nature. 307:161-163.

Lambon-Ralph MA, Jefferies E, Patterson K, Rogers TT. 2017. The neural and computational bases of semantic cognition. Nat Rev Neurosci. 18:42-55.

Landauer TK, Dumais ST. 1997. A solution to Plato's problem: The Latent Semantic Analysis theory of acquisition, induction, and representation of knowledge. Psychological Review. 104:211-240.

Lau E, Nguyen E. 2015. The role of temporal predictability in semantic expectation: an MEG investigation. Cortex. 68:8-19.

Lau EF, Almeida D, Hines PC, Poeppel D. 2009. A lexical basis for N400 context effects: evidence from MEG. Brain Lang. 111:161-172.

Lau EF, Gramfort A, Hämäläinen MS, Kuperberg GR. 2013. Automatic semantic facilitation in anterior temporal cortex revealed through multimodal neuroimaging. J Neurosci. 33:17174-17181.

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

Lau EF, Weber K, Gramfort A, Hämäläinen MS, Kuperberg GR. 2016. Spatiotemporal signatures of lexico-semantic prediction. Cerebral Cortex. 26:1377-1387.

Lee TS, Mumford D. 2003. Hierarchical Bayesian inference in the visual cortex. J Opt Soc Am A. 20:1434.

Lerner Y, Honey CJ, Silbert LJ, Hasson U. 2011. Topographic mapping of a hierarchy of temporal receptive windows using a narrated story. J Neurosci. 31:2906-2915.

Lewis A, Bastiaansen M. 2015. A predictive coding framework for rapid neural dynamics during sentence-level language comprehension. Cortex. 68:155-168.

Liu Z, Shu S, Lu L, Ge J, Gao JH. 2020. Spatiotemporal dynamics of predictive brain mechanisms during speech processing: an MEG study. Brain Lang. 203:104755.

Lyu B, Choi HS, Marslen-Wilson WD, Clarke A, Randall B, Tyler LK. 2019. Neural dynamics of semantic composition. Proc Natl Acad Sci U S A. 116:21318-21327.

MacDonald MC, Pearlmutter NJ, Seidenberg MS. 1994. The lexical nature of syntactic ambiguity resolution. Psychological Review. 101:676-703.

Maess B, Herrmann CS, Hahne A, Nakamura A, Friederici AD. 2006. Localizing the distributed language network responsible for the N400 measured by MEG during auditory sentence processing. Brain Res. 1096:163-172.

Mamashli F, Khan S, Obleser J, Friederici AD, Maess B. 2019. Oscillatory dynamics of cortical functional connections in semantic prediction. Hum Brain Mapp. 40:1856-1866.

Marinkovic K, Dhond RP, Dale AM, Glessner M, Carr V, Halgren E. 2003. Spatiotemporal dynamics of modality-specific and supramodal word processing. Neuron. 38:487-497.

Maris E, Oostenveld R. 2007. Nonparametric statistical testing of EEG- and MEG-data. J Neurosci Methods. 164:177-190.

Marslen-Wilson WD. 1987. Functional parallelism in spoken word-recognition. Cognition. 25:71-102.

Marslen-Wilson WD, Brown C, Tyler LK. 1988. Lexical representations in spoken language comprehension. Lang Cogn Process. 3:1-17.

McCarthy G, Nobre AC, Bentin S, Spencer DD. 1995. Language-related field potentials in the anterior-medial temporal lobe: I. Intracranial distribution and neural generators. J Neurosci. 15:1080-1089.

McClelland JL, McNaughton BL, O'Reilly RC. 1995. Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. Psychol Rev. 102:419-457.

Mumford D. 1992. On the computational architecture of the neocortex. II. The role of cortico-cortical loops. Biol Cybern. 66:241-251.

Ness T, Meltzer-Asscher A. 2018. Predictive preupdating and working memory capacity: Evidence from event-related potentials. J Cogn Neurosci.1-23.

Nieuwland MS, Barr DJ, Bartolozzi F, Busch-Moreno S, Darley E, Donaldson DI, Ferguson HJ, Fu X, Heyselaar E, Huettig F, Matthew Husband E, Ito A, Kazanina N, Kogan V, Kohut Z, Kulakova E, Meziere D, Politzer-Ahles S, Rousselet G, Rueschemeyer SA, Segaert K, Tuomainen J, Von Grebmer Zu Wolfsthurn S. 2020. Dissociable effects of prediction and integration during language comprehension: evidence from a large-scale study using brain potentials. Philos Trans R Soc Lond B Biol Sci. 375:20180522.

Nobre AC, McCarthy G. 1995. Language-related field potentials in the anterior-medial temporal lobe: II. Effects of word type and semantic priming. J Neurosci. 15:1090-1098.

Nour Eddine S. 2021. Divide and Concur: A predictive coding account of the N400 ERP component. Medford, MA: Tufts University.

Nour Eddine S, Brothers T, Kuperberg GR. in press. The N400 in silico: A Review of Computational models. In: Federmeier K, editor. Psychology of Learning and Motivation Academic Press.

O'Reilly RC, Rudy JW. 2001. Conjunctive representations in learning and memory: principles of cortical and hippocampal function. Psychological review. 108:311.

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.

Paczynski M, Kuperberg GR. 2011. Electrophysiological evidence for use of the animacy hierarchy, but not thematic role assignment, during verb argument processing. Lang Cogn Process. 26:1402-1456.

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

Payne BR, Lee CL, Federmeier KD. 2015. Revisiting the incremental effects of context on word processing: Evidence from single-word event-related brain potentials. Psychophysiology. 52:1456-1469.

Perrin F, Pernier J, Bertrand O, Echallier JF. 1989. Spherical splines for scalp potential and current density mapping. Electroencephalogr Clin Neurophysiol. 72:184-187.

Price CJ. 2012. A review and synthesis of the first 20 years of PET and fMRI studies of heard speech, spoken language and reading. Neuroimage. 62:816-847.

Price CJ, Devlin JT. 2011. The interactive account of ventral occipitotemporal contributions to reading. Trends Cogn Sci. 15:246-253.

Qian T, Jaeger TF, Aslin RN. 2012. Learning to represent a multi-context environment: More than detecting changes. Front Psychol. 3:228.

Rabovsky M, Hansen SS, McClelland JL. 2018. Modelling the N400 brain potential as change in a probabilistic representation of meaning. Nat Hum Behav. 2:693-705.

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

Radvansky GA, Zacks JM. 2011. Event perception. Wiley Interdiscip Rev Cogn Sci. 2:608-620.

Rao RPN, Ballard DH. 1997. Dynamic model of visual recognition predicts neural response properties in the visual cortex. Neural Computation. 9:721-763.

Rao RPN, Ballard DH. 1999. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. Nature Neuroscience. 2:79-87.

Rescorla RA. 1988. Pavlovian conditioning: It's not what you think it is. Am Psychol. 43:151.

Sanford AJ, Leuthold H, Bohan J, Sanford AJS. 2011. Anomalies at the borderline of awareness: an ERP study. Journal of Cognitive Neuroscience. 23:514-523.

Smith NJ, Levy R. 2013. The effect of word predictability on reading time is logarithmic. Cognition. 128:302-319.

Sohoglu E, Davis MH. 2020. Rapid computations of spectrotemporal prediction error support perception of degraded speech. Elife. 9.

Solomyak O, Marantz A. 2009. Lexical access in early stages of visual word processing: A single-trial correlational MEG study of heteronym recognition. Brain Lang. 108:191-196.

Spratling MW. 2017. A review of predictive coding algorithms. Brain Cogn. 112:92-97.

Tanenhaus MK, Trueswell JC. 1995. Sentence comprehension. In: Miller JL, Eimas PD, editors. Speech, Language, and Communication 2 ed. San Diego, CA: Academic Press p 217-262.

van de Meerendonk N, Kolk HHJ, Chwilla DJ, Vissers CTWM. 2009. Monitoring in language perception. Lang Linguist Compass. 3:1211-1224.

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

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

Van Petten C, Luka BJ. 2012. Prediction during language comprehension: benefits, costs, and ERP components. Int J Psychophysiol. 83:176-190.

Van Petten C, Weckerly J, McIsaac HK, Kutas M. 1997. Working memory capacity dissociates lexical and sentential context effects. Psychological Science. 8:238-242.

Vartiainen J, Parviainen T, Salmelin R. 2009. Spatiotemporal convergence of semantic processing in reading and speech perception. J Neurosci. 29:9271-9280.

Wang L, Jensen O, Kuperberg GR editors. Representational Similarity Analysis reveals unique patterns associated with the fulfillment and violation of lexico-semantic prediction within the N400 time window: An MEG study, 25th Annual Meeting of the Cognitive Neuroscience Society; 2018; Boston, MA.

Wang L, Wlotko E, Alexander EJ, Schoot L, Kim M, Warnke L, Kuperberg GR. 2020. Neural evidence for the prediction of animacy features during language comprehension: Evidence from MEG and EEG Representational Similarity Analysis. J Neurosci. 40:3278-3291.

Whittington JCR, Bogacz R. 2019. Theories of error back-propagation in the brain. Trends Cogn Sci. 23:235-250.

Woolnough O, Donos C, Rollo PS, Forseth KJ, Lakretz Y, Crone NE, Fischer-Baum S, Dehaene S, Tandon N. 2021. Spatiotemporal dynamics of orthographic and lexical processing in the ventral visual pathway. Nat Hum Behav. 5:389–398.

Xiang M, Kuperberg G. 2015. Reversing expectations during discourse comprehension. Lang Cogn Neurosci. 30:648-672.

Table 1. Example of each experimental condition, together with stimuli characteristics.

Prior discourse-constraining context: *The lifeguards received a report of sharks right near the beach. Their immediate concern was to prevent any incidents in the sea. Hence, they cautioned the…*								
Scenario Type	**Critical word**	***Constraint**	****Cloze**	**+SSV**	**Length**	**++Freq.**	**^OLD**	**^^Conc.**
Expected	*swimmers…*	69% (14%)	69% (14%)	0.18 (.18)	5.69 (1.60)	1.53 (0.66)	1.93 (0.56)	4.30 (0.69)
Unexpected Plausible	*trainees…*	69% (14%)	0.1% (0.5%)	0.01 (.06)	7.46 (2.22)	0.61 (0.88)	2.61 (0.86)	4.15 (0.69)
Implausible	*drawers…*	66% (16%)	0% (0%)	0.01 (.05)	7.11 (2.04)	0.81 (0.85)	2.47 (0.81)	4.21 (0.66)

The critical words (underlined here although not in the experiment itself) were followed by three additional words, indicated here with three dots.

Means are shown with the standard deviations in parentheses.

*The lexical constraint of each discourse context was calculated by identifying the most common completion across participants who saw that context in a cloze norming study (see Supplementary Materials), and tallying the proportion of participants who provided this completion.

**Cloze probabilities of critical words were calculated based on the percentage of respondents in the cloze norming study who provided the critical noun.

+SSV: Semantic Similarity Values, quantifying the semantic relatedness between the critical words and the "bag of words" within the prior contexts, based on Latent Semantic Analysis.

Length: Number of letters.

++Freq.: Log Frequency values of critical words, retrieved from the English Lexicon Project.

^OLD: Orthographic Levenshtein Distance values of critical words, retrieved from the English Lexicon Project.

^^Conc.: Concreteness ratings of critical words (Brysbaert et al. 2014).

Table 2. ERP statistical results.

Spatiotemporal Region	**Contrast**	***t*-value** *df* (31)	***p*-value**	**Effect size (*d*)**
N400 (Central region, 300-500ms)	Unexpected > Expected	-5.31	< 0.001	0.94
	Implausible > Expected	-7.72	< 0.001	1.36
	Implausible > Unexpected	-4.63	0.001	0.82
Late frontal positivity (Prefrontal region, 600-1000ms)	Unexpected > Expected	3.03	0.005	0.53
	Implausible = Expected	1.51	0.14	0.24
	Unexpected > Implausible	2.52	0.018	0.42
Late posterior positivity/P600 (Posterior region, 600-1000ms)	Unexpected = Expected	1.91	0.07	0.34
	Implausible > Expected	7.65	< 0.001	1.12
	Implausible > Unexpected	5.75	< 0.001	0.99

Figure legends

Figure 1. Hypothesized patterns of evoked N400 modulation across *expected*, *unexpected plausible* and *implausible* words within temporal and inferior frontal cortices in four general frameworks of language comprehension. A. According to a *distributed state* account, the N400 reflects the degree of update induced by new inputs within a single state that is distributed across the fronto-temporal cortical hierarchy. It therefore predicts graded N400 modulation (*expected* < *unexpected plausible* < *implausible*) within *both* temporal and inferior frontal cortices. **B.** A *prediction-integration* account distinguishes between lexical facilitation resulting from lexical-level predictions in plausible sentences and event-level integration in implausible sentences. Within the temporal cortex, this framework predicts a reduced N400 only to *expected* words (*expected* < *unexpected plausible* = *implausible*), and, within the inferior frontal cortex, it predicts an enhanced N400 only to *implausible* words (*expected* = *unexpected plausible* < *implausible*). **C.** According to a *lexico-semantic facilitation* account, the N400 reflects the degree to which processing of an incoming word is facilitated at a lower lexico-semantic level of representation. It therefore predicts graded increases in the N400 response across the three conditions within the temporal cortex, but no differences in the N400 within inferior frontal cortex. **D.** *Predictive coding* posits that the amplitude of the N400 produced at any given level of the cortical hierarchy reflects the magnitude of prediction error, i.e. activity within error units that is not suppressed/explained by top-down predictions generated by state units at the level above. It predicts graded modulation of the N400 within the temporal cortex (*expected* < *unexpected plausible* < *implausible*), reflecting graded increases in the magnitude of lexico-semantic prediction error, and an enhanced N400 to *implausible* words within the inferior frontal cortex (*expected* = *unexpected plausible* < *implausible*), reflecting a failure to suppress higher-level prediction error produced at the level of

the event model.

Figure 2. 300-500ms (N400): Proposed predictive coding across the left fronto-temporal hierarchy. In rich, constraining discourse contexts, top-down predictions, based on stored real-world and linguistic knowledge, are propagated down the cortical hierarchy (faint blue diagonal arrows) before new bottom-up input becomes available. **Lexico-semantic level** (*pink*): Within regions of the left temporal cortex that support lexico-semantic processing, the lexico-semantic representation of the incoming word is inferred throughout the N400 time-window within "state units" (ST_{i-1}→ST_i). Within a distinct set of "error units", residual lexico-semantic information that is not suppressed (cannot be explained) by the top-down predictions (PE = max(0, ST– Pr), simplified in the figure as PE = ST – Pr), produces lexico-semantic prediction error, which manifests as the evoked N400 response. Lexico-semantic prediction error, and evoked activity within left temporal cortex is therefore smallest to *expected* words (e.g. "swimmers", not shown), larger to *unexpected plausible* words (e.g. "trainees", *left*), where some semantic features (e.g. <animate>) were predicted, and largest to *implausible* words (e.g. "drawers", *right*) where no semantic features were predicted. **Event model** (*green*): Throughout the N400 time-window, lexico-semantic prediction error passes up to left inferior frontal cortex (red arrows), where it induces updates of the higher-level event model. *Unexpected plausible* (*left*): These updates yield a plausible high-level event model that is congruent with predictions about the possibility/plausibility of real-world events. Therefore, higher-level *prediction error* is suppressed, and the evoked response within left inferior frontal cortex is no larger than to *expected* inputs. In addition, the newly-updated event model (ST_i) now generates correct top-down predictions (blue curly arrow) that "switch off"/suppress lower-level lexico-semantic prediction error, leading to the reduction of the evoked response within left temporal cortex at the end of the N400 time-window. *Implausible* (*right*): Updates of the event model yield a highly *implausible* interpretation (e.g.

<lifeguards cautioned drawers>) that cannot be explained by real-world knowledge predictions, resulting in a large higher-level *prediction error* at the level of the event model, and an enhanced evoked response within left inferior frontal cortex (red arrow). Moreover, because it is more difficult to converge on an implausible interpretation, top-down lexico-semantic predictions will be less accurate and less likely to suppress lower-level lexico-semantic prediction error (blue dotted curly arrow), further enhancing and prolonging the evoked response within left temporal cortex.

Figure 3. 600-1000ms: Proposed predictive coding across the left fronto-temporal hierarchy. Unexpected plausible (*left panel*). **A.** The plausible event inferred between 300-500ms (e.g. <lifeguards cautioned trainees>) is out of keeping with the comprehender's high confidence beliefs in the prior event model, previously inferred from the discourse context. This triggers the retrieval of new schema from long-term memory between 600-1000ms (gray curly arrow), which generate new predictions (blue diagonal arrow). **B.** *Event model:* In left inferior frontal cortex, residual information within the new predictions that cannot be explained by the prior event model produces *high-level top-down error* within the higher-level error units (max(0, Pr – ST) = TdE, simplified in the figure as Pr – ST = TdE), which manifests as a late evoked response in this region. This induces a large top-down shift of the event model within the state units ($ST_i \rightarrow ST_{i+1}$), which, in turn, generates new predictions (blue diagonal arrow). **C**. *Lexico-semantic level:* In left temporal cortex, residual information within the new top-down predictions that cannot be explained by the prior lexico-semantic state produces *top-down lexico-semantic error* within the error units, which manifests as a late evoked response in this region, and induces a top-down lexico-semantic shift within the state units. This produces orthographic predictions (blue curly arrow) that continue to suppress orthographic prediction error at the level below (**D**). **Implausible** (*right panel*). **A.** There are no stored schemas within memory that can explain the implausible/impossible input (black cross). Therefore, between 600-1000ms, incorrect top-down predictions continue to be generated based on the prior context (blue diagonal arrow). **B.** *Event model:* In left inferior frontal cortex, previous event predictions (e.g. <lifeguard cautioned swimmers>) fail to match the implausible event inferred in the N400 time-window (<lifeguards cautioned drawers>), producing *top-down event error* within the error units and a late evoked response within this region. This induces a top-down shift to a plausible event (<lifeguard cautioned swimmers>), which generates new lexico-semantic predictions. **C**. *Lexico-semantic level:* In left temporal cortex, these predictions (“swimmers”, <animate>) are incompatible with the lexico-semantic state inferred in the N400 time-window (“drawers”, <inanimate>), resulting in a destabilization of this lexico-semantic state (indicated with a “?”), and the production of inaccurate orthographic predictions that are passed down to the level below (blue curly dotted arrow). **D.** *Orthographic level:* Within left posterior fusiform cortex, the inaccurate orthographic predictions fail to suppress prediction error produced by the orthographic state (d-r-a-w-e-r-s), resulting in a late evoked response (reprocessing).

Figure 4. ERP results. A. Grand-averaged ERP waveforms elicited by critical words in each of the three conditions, shown at three representative electrode sites: Cz, FPz and Pz. *Expected*: solid black line; *Unexpected plausible*: solid red line; *Implausible*: dashed blue line. Negative voltage is plotted upwards. Dotted boxes are used to indicate the time-windows corresponding to the N400 (300-500ms), the *late frontal positivity* (600-1000ms) and the *late posterior positivity/P600* (600-1000ms) ERP components. **B. Voltage maps** show the topographic distributions of the ERP effects produced by contrasting *Expected*, *Unexpected plausible* and *Implausible* critical words between 300-500ms (left panel) and between 600-1000ms (right panel). Note that the N400 effects and the late positivity effects are shown at different voltage scales to better illustrate the scalp distribution of each effect.

Figure 5. MEG sensor-level results. A. 300-500ms. *Top:* Grand-averaged event-related magnetic fields produced by critical words in each of the three conditions, shown at a left temporal gradiometer sensor (MEG0242+0243). The 300-500ms (N400) time-window is indicated using a dotted box. *Bottom:* MEG Gradiometer (Grad.) and Magnetometer (Mag.) sensor maps show the topographic distributions of the MEG N400 effects produced by contrasting the *Expected*, *Unexpected plausible* and *Implausible* critical words between 300-500ms. In all contrasts, the distribution of the MEG N400 effect was maximal over temporal sites, particularly on the left. **B. 600-1000ms.** MEG Gradiometer (Grad.) and Magnetometer (Mag.) sensor maps show the topographic distributions of the MEG effects produced by contrasting the *Expected*, *Unexpected plausible* and *Implausible* critical words in the first half (600-800ms) and the second half (800-1000ms) of the late time-window of interest. In order to better illustrate the scalp distribution of these late effects, these sensor maps are shown at a different scale from that used for the 300-500ms sensor maps. The contrasts between the *Unexpected plausible* and *Expected* critical words and the contrast between the *Implausible* and *Expected* critical words reveal somewhat distinct spatial distributions of sensor-level activity.

Figure 6. MEG source-level activity produced by the expected, unexpected plausible and implausible critical words in the 300-500ms (N400) time-window. A. Signed dynamic Statistical Parametric Maps (dSPMs) produced by *Expected* (*top*), *Unexpected plausible* (*middle*), and *Implausible* (*bottom*) critical words are shown at 100ms intervals from 200 until 500ms. All dSPMs are displayed on the FreeSurfer average surface, "fsaverage" (Fischl *et al.* 1999), thresholded at 0.15, with red indicating outgoing currents (positive dSPM values) and blue indicating ingoing currents (negative dSPM values). Ingoing and outgoing currents that are directly adjacent to one another are interpreted as reflecting a single underlying dipole (neuroanatomical source). This is because if an underlying dipole/source is situated on one side of a sulcus, the signal can bleed into the other side, leading to the appearance of an adjacent dipole in the opposite direction (Hämäläinen *et al.* 1993). The full dynamics of these source activations (at all sampling points) are shown as videos in Supplementary Materials. **B.** Statistical maps contrasting *Unexpected plausible* and *Expected* words (*top*), *Implausible* and *Expected* words (*middle*), and *Implausible* and *Unexpected plausible* words (*bottom*) within the 300-500ms (N400) time-window. Averaged -log10 transformed uncorrected p-values ($p < 0.05$) at each vertex are shown on the "fsaverage" surface. We grouped together all spatial patches that reached cluster-level significance into the neuroanatomical regions that are shown in Supplementary Figure 1A and listed in Supplementary Table 1 (defined using the Desikan-Killiany Atlas; Desikan *et al.* 2006). If one or more patches within a neuroanatomical region reached cluster-level significance, we indicate the region using a red circle. Within left temporal cortex, effects reached cluster-level significance, within (i) superior temporal gyrus, extending anteriorly towards the temporal pole and extending posteriorly into the supramarginal gyrus, (ii) the mid-portion of the superior temporal sulcus/middle temporal cortex (for contrasts involving the *Implausible* words), (iii) left

ventral temporal cortex (mid- and posterior fusiform gyrus), and (iv) left medial temporal cortex (parahippocampal and entorhinal). The *Implausible* words additionally produced cluster-level effects in the left inferior frontal cortex (relative to both other conditions), and in the anterior cingulate cortex (relative to the *Expected* words). See Supplementary Materials for analyses over the right hemisphere (Supplementary Figures 2, 3 and 4).

Figure 7. MEG source-level activity produced by the expected, unexpected plausible and implausible critical words in the late 600-1000ms time-window. A. Signed dynamic Statistical Parametric Maps (dSPMs) produced by *Expected* (*top*), *Unexpected plausible* (*middle*), and *Implausible* (*bottom*) critical words are shown at 100ms intervals from 500 until 1000ms. All dSPMs are displayed on the FreeSurfer average surface, “fsaverage” (Fischl *et al.* 1999), thresholded at 0.15, with red indicating outgoing currents (positive dSPM values) and blue indicating ingoing currents (negative dSPM values). Ingoing and outgoing currents that are directly adjacent to one another are interpreted as reflecting a single underlying dipole (neuroanatomical source). This is because, if an underlying dipole/source is situated on one side of a sulcus, the signal can bleed into the other side, leading to the appearance of an adjacent dipole in the opposite direction (Hämäläinen *et al.* 1993). The full dynamics of these source activations (at all sampling points) are shown as videos in Supplementary Materials. **B.** Statistical maps contrasting *Unexpected plausible* and *Expected* words (*top*), *Implausible* and *Expected* words (*middle*), and *Implausible* and *Unexpected plausible* words (*bottom*) are shown between 600-800ms (*left*) and between 800-1000ms (*right*). Averaged -log10 transformed uncorrected p-values ($p < 0.05$) at each vertex are shown on the “fsaverage” surface. We grouped together all spatial patches that reached cluster-level significance into the neuroanatomical regions that are shown in Supplementary Figure 1A and listed in Supplementary Table 1 (defined using the Desikan-Killiany Atlas; Desikan *et al.* 2006). If one or more patches within a neuroanatomical region reached cluster-level significance, we indicate the region using a red circle. The *Unexpected plausible* versus *Expected* contrast revealed significant clusters within left middle temporal and inferior frontal cortices (driven by dipoles going in opposite directions in the two conditions). The *Implausible* versus *Expected* contrast revealed significant clusters within left posterior (occipitotemporal) fusiform cortex

(driven by a dipole to the *Implausible* words), left inferior frontal cortex (driven by dipoles going in opposite directions in the two conditions), and within the medial temporal cortex (driven by a dipole to the *Implausible* words). See also Supplementary Materials for analyses over the right hemisphere (Supplementary Figures 2, 3 and 4).