This paper was submitted for peer review on February 8, 2021. Based on our past experience, we expect that it will change significantly from this date until the time it is published, based on our reviewers' comments, comments of colleagues, and additional work by the authors. Thus, although we provide the manuscript for open access at this early stage, we ask interested readers to bear in mind that this version is NOT final, and, to check back to find the latest version. We will provide the final peer- reviewed version of the manuscript on our website once it is accepted for publication.

Dynamic predictive coding across the left fronto-temporal language hierarchy: Evidence from MEG, EEG and fMRI

*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

Abstract

Predictive coding has been proposed as a unifying theory of brain function. However, few studies have examined this theory during complex cognitive processing across multiple time-scales and levels of abstraction. We used MEG, EEG and fMRI to ask whether dynamic, hierarchical predictive coding can account for the timecourse of evoked activity at multiple cortical levels during language comprehension. Unexpected words produced increased activity in left temporal cortex (lower-level prediction error). Critically, violations of high-precision event predictions produced additional activity within left inferior frontal cortex (higher-level prediction error). Furthermore, the successful resolution of higher-level prediction error led to later feedback to temporal cortex (top-down sharpening), while a failure to resolve these errors led to sustained activity at still lower levels (reanalysis). These findings suggest that fundamental principles of dynamic hierarchical predictive coding — suppression of prediction error, precision-weighting, delayed top-down sharpening — can explain the dynamics of neural activity during human language comprehension.

Introduction

The process by which we make sense of the world around us can be understood as *probabilistic inference* — the use of prior knowledge, encoded within a *generative model*, to infer the underlying higher-level representation that best "explains" the bottom-up input[1]. It has been proposed that the brain approximates probabilistic inference using an algorithm known as *predictive coding*[2-5]. According to this algorithm, more abstract representations, encoded at higher levels of the cortical hierarchy, are used to predict or "reconstruct" information at lower cortical levels. Any new bottom-up information that is not present within these top-down reconstructions produces *prediction error*, which is used to *update* the higher-level representations, allowing them to produce more accurate reconstructions that switch off the lower-level prediction error. Inference is complete when prediction error is minimized across all levels of the generative hierarchy.

Although predictive coding is sometimes interpreted as a unifying theory of brain function[4,5], most of its supporting evidence comes from studies examining neural activity at lower cortical levels within a single time window[6]. For example, predictable inputs have been shown to evoke less activity than unpredictable inputs within lower-level temporal and occipital regions across multiple tasks, including auditory processing[7], speech perception[8] and visual word recognition[9]. These findings are consistent with the suppression of lower-level prediction error by top-down reconstructions that match expected perceptual inputs.

There has been far less study of whether this theory can account for the timecourse of neural activity evoked at both lower *and* higher levels of the cortical hierarchy in complex and rapidly changing environments. According to *dynamic hierarchical predictive coding*[10,11], the abstract representation that generates top-down reconstructions is a dynamic "state" that (a) is iteratively updated by new unpredicted input (lower-level prediction error) as it becomes available

over time, and (b) receives top-down reconstructions from still higher levels of the cortical hierarchy. In a changing environment, the brain continually tracks its confidence in this state. If a lower-level prediction error induces an update that is inconsistent with either a prior high-certainty state, or with reconstructions received from the higher cortical level, then this will result in a *higher-level prediction error*[11], which can trigger a "model shift"[12] at the highest level of the generative hierarchy. Successful high-level shifts will then generate new reconstructions that provide feedback to lower levels of cortex, enhancing activity over consistent representations and reducing activity over inconsistent representations at later stages of processing (top-down "sharpening")[4,13].

Language comprehension provides an excellent model system for testing this dynamic hierarchical predictive coding framework. This is because it requires us to transform rapidly unfolding sequences of words into a dynamic, high-level state that encodes our understanding of the events being communicated — an *event model*[14]. Within this framework[15], the event model receives top-down reconstructions from long-term schema-relevant knowledge, represented at the highest level of the hierarchy. At any given time, the brain continually estimates its confidence in the current high-level event representation, based on the prior context, while also generating top-down probabilistic reconstructions of upcoming lexico-semantic information (top-down pre-activation[16,17]). When a new word is encountered, any new lexico-semantic information that is not contained in these lower-level reconstructions — *lower-level lexico-semantic prediction error*— is passed back up and used to update the event model.

According to dynamic predictive coding, incremental updates to the event model, induced by lower-level prediction error, are usually sufficient to explain the bottom-up input. Importantly, however, if an update leads the comprehender to infer an event that is either inconsistent with a

prior high-certainty event representation, or that falls outside the range of plausible events reconstructed by the higher-level schema, then it will produce a *higher-level event prediction error*. This event prediction error induces a shift away from the current schema at the highest level of the hierarchy[15]. If there is a new schema stored within long-term memory that can better explain the input, then it will be retrieved[18], resulting in the production of new reconstructions that provide feedback to lower cortical levels, enhancing activity over schema-consistent lexico-semantic representations[15]. If, however, the newly inferred event is completely anomalous, with no pre-stored schema that can explain it, then this will result in a failure to switch off prediction error at still lower levels of the hierarchy (reanalysis), and may trigger new learning in order to explain the input[18,19].

Studies using scalp-recorded event-related potentials (ERPs) have uncovered some evidence that the brain does indeed differentiate between unpredictable words that do, or do not, violate higher-level contextual constraints. In plausible sentences, contextually unexpected words generate a larger evoked response between 300-500ms than expected words (the N400 effect), regardless of the constraint of the prior context[20-22]. However, only words that violate higher-level contextual constraints produce additional late activity between 600-1000ms, with different scalp distributions depending on whether they yield plausible or anomalous interpretations[22,23]. Within a predictive coding framework, evoked (phase-locked) neural responses reflect the magnitude of prediction error[4]. These findings therefore provide some evidence for a temporal distinction between lower-level (lexico-semantic) and higher-level (event) prediction error during language comprehension. To date, however, it remains unknown whether this temporal distinction is accompanied by a neuroanatomical dissociation across the left lateralized fronto-temporal hierarchy that is classically associated with language processing.

While numerous previous fMRI and MEG studies have established clear effects of top-down context on activity within this fronto-temporal network during sentence comprehension[24], none has been able to address this question directly. This is because most of these studies contrasted implausible and plausible words, without independently manipulating predictability and contextual constraint. Moreover, fMRI lacks the temporal resolution to dissociate evoked activity at earlier and later stages of processing, while MEG studies have rarely reported activity in later time-windows. Finally, no previous study of sentence comprehension has examined the time course and spatial localization of neural activity produced using all three neuroimaging methods in the same participants. Given that these techniques are sensitive to different aspects of underlying neural activity, such direct comparisons are critical for integrating the large ERP, MEG and fMRI literatures examining the influences of context on language processing.

We therefore undertook a comprehensive multimodal neuroimaging study (MEG, EEG and fMRI) that examined the timecourse and spatial localization of neural responses evoked by incoming words as comprehenders read four types of multi-sentence discourse scenarios (Table 1). We compared neural activity evoked by *expected* critical words and three different types of unpredictable critical words: plausible words in low constraint contexts (*low constraint unexpected*), plausible words that violated high constraint contexts (*high constraint unexpected*), and words that yielded impossible interpretations (*anomalous*).

Insert Table 1 here

We expected that both the *low constraint unexpected* and the *high constraint unexpected* words would produce a larger evoked response between 300-500ms (a larger N400) than *expected* words[20-22]. If, as posited by predictive coding, this effect reflects *lower-level lexico-semantic*

prediction error, then it should localize to lower levels of the language cortical hierarchy (left temporal cortex). We also expected that only *high constraint unexpected* words would additionally evoke activity in a later time-window 600-1000ms (a *late frontal positivity* ERP effect[21,22]). According to the dynamic hierarchical predictive coding framework outlined above, this late activity should reflect the production of a *higher-level event prediction error* that is produced when a newly inferred event violates a prior high precision estimate of a different event. As such, it should localize to higher regions of the language cortical hierarchy (left inferior frontal cortex). Moreover, it should be accompanied by a re-activation of lower-level regions (temporal cortex), reflecting feedback activation of new schema-relevant lexico-semantic information (top-down "sharpening").

Finally, we predicted that, relative to the expected words, *anomalous* words that were incompatible with prior event reconstructions would produce a larger evoked response within the left inferior frontal cortex (an early higher-level event prediction error), as well as a larger response within the temporal cortex (due to a failure to switch off lower-level lexico-semantic prediction error). The failure to retrieve a new schema from long-term memory to explain the input should also result in a different pattern of activity in the later time window (600-1000ms), corresponding to the *late posterior positivity/P600* ERP effect[22,25,26]. This late activity may reflect a failure to switch off prediction error ("reanalysis") within regions that support still lower-level orthographic processing (e.g. the posterior fusiform cortex[9]), and/or activity within regions implicated in longer-term learning (e.g. the medial temporal lobe[27]).

To test these hypotheses, we collected MEG and EEG data in the same session. A distributed source localization analysis of the MEG data, which is relatively undistorted by the conductivities of the skull and scalp, allowed us to track the time course and localization of evoked

activity produced by the incoming words. The simultaneous collection of EEG data enabled us to link this source-localized activity to ERP effects reported in the prior literature. Finally, in a separate session, we collected fMRI in the same participants, which allowed us to examine similarities and differences between source-localized MEG activity and the hemodynamic response across our four conditions.

Results

Behavioral results

Participants correctly judged the plausibility of the discourse scenarios in 85.5% (SD: 6.3%) of trials on average. They answered 82.4% (SD: 10.1%) of the comprehension questions correctly, suggesting that they were engaged in comprehension. See Supplementary Materials section 2 for a detailed report.

ERP results

Plausible unexpected *vs.* expected

The N400 evoked by the *expected* critical words was significantly smaller (less negative) than that evoked by the *low constraint unexpected* ($t(31) = -8.53$, $p < 0.001$) and the *high constraint unexpected* critical words ($t(31) = -5.31$, $p < 0.001$), see Figure 1a.

Insert Figure 1 here

Between 600-1000ms, the contrast between the *low constraint unexpected* and *expected* critical words did not reveal any effects (prefrontal region: $t(31) = -0.78$, $p = 0.44$; posterior region: $t(31) = -0.26$, $p = 0.79$). However, the contrast between the *high constraint unexpected* and *expected* critical words produced a late frontal positivity effect (prefrontal region: $t(31) = 3.03$, p

= 0.005), but no late posterior positivity/P600 effect (posterior region: $t(31) = 1.91$, $p = 0.07$), see Figure 1b.

Anomalous *vs.* expected

This contrast again revealed an N400 effect ($t(31) = -7.72$, $p < 0.001$). Between 600-1000ms, it additionally revealed a late posterior positivity/P600 effect (posterior region: $t(31) = 7.65$, $p < 0.001$), but no late frontal positivity effect (prefrontal region: t: $t(31) = 1.51$, $p = 0.14$), see Figure 1b.

These findings replicate our previous ERP study using overlapping stimuli in a different group of participants[22].

MEG results

The sensor-level findings are shown in Figure 2. The MEG N400 was smaller to *expected* critical words than to all three types of unpredictable critical words. Between 600-1000ms, the topographic sensor maps contrasting the two types of plausible *unexpected* with the *expected* critical words show similar patterns of activity, but the magnetometer maps suggest that the effect was larger for the contrast between *high constraint unexpected* and *expected* words. The *anomalous* versus *expected* contrast revealed a spatially distinct pattern of activity.

Insert Figure 2 here

Source localized MEG activity: Plausible unexpected *vs.* expected

300-500ms: Figure 3a (left) depicts the signed dSPMs produced by the *low constraint unexpected*, *expected,* and *high constraint unexpected* critical words at 100ms intervals, from 200ms until 500ms. Figure 3a (right) shows statistical maps contrasting the two types of plausible *unexpected* critical words with the *expected* words within the 300-500ms window of interest. Both

contrasts reveal significantly more activity to the *unexpected* than the *expected* critical words within the left lateral temporal cortex (superior temporal gyrus, extending anteriorly towards the temporal pole, and posteriorly into the supramarginal gyrus, and the mid-portion of the superior temporal sulcus/middle temporal cortex), and the left ventral temporal cortex (mid and posterior fusiform gyrus). They also revealed effects within the left medial temporal cortex (parahippocampal and entorhinal), which were driven both by a dipole to the *unexpected* critical words (outgoing) and a dipole in the opposite direction (ingoing) to the *expected* critical words.

Insert Figure 3 here

600-1000ms: Figure 3b (middle panel) presents the signed dSPMs produced by the critical words in the same three conditions at 100ms intervals, from 500 until 1000ms, and the statistical maps for both contrasts between 600-800ms (left panel) and 800-1000m (right panel). The contrast between the *low constraint unexpected* and *expected* critical words showed no significant effects in either time window (although it did reveal non-significant activity within the anterior inferior frontal gyrus throughout the 600-1000ms window, and within the left lateral temporal cortex between 800-1000ms). The contrast between the *high constraint unexpected* and *expected* critical words, however, revealed effects within the left anterior inferior frontal cortex and within the left middle temporal cortex, which reached cluster-level significance within the 800-1000ms time window, and were driven by dipoles going in opposite directions in the two conditions. Of note, the dipoles within the left middle temporal cortex were of the opposite polarity to those observed within the same region in the 300-500ms time window.

Source localized MEG activity: Anomalous *vs.* expected

Figure 4 shows the signed dSPMs produced by the *anomalous* and the *expected* critical words at 100ms intervals from 200ms until 1000ms, and the statistical contrasts between the two conditions for the 300-500ms, 600-800ms and 800-1000ms time windows of interest.

Insert Figure 4 here

300-500ms: The *anomalous* words produced effects within left lateral, ventral and medial temporal cortices that appeared qualitatively similar, but stronger than the effects produced by the *unexpected* plausible (versus *expected*) critical words, described above. In addition, this contrast revealed significantly more activity to the *anomalous* than the *expected* critical words within the left inferior frontal and anterior cingulate cortex.

600-1000ms: In this later time window, the *anomalous* vs. *expected* contrast revealed effects within the posterior portion of the left temporal fusiform cortex (significant between 600-800ms, driven by increased activity to the *anomalous* words), within the anterior inferior frontal gyrus (significant between 800-1000ms, driven by dipoles going in opposite directions to the *anomalous* and *expected* words), and within the left parahippocampal gyrus (significant across the whole 600-1000ms window, driven by a large ingoing dipole to the *anomalous* words, which was of the opposite polarity to that observed during the 300-500ms time window).

We report the results of exploratory analyses over the right hemisphere in Supplementary Figures 2, 3 and 4. We also illustrate the dynamics of source activation in each of the four experimental conditions as "movies" in Supplementary materials.

fMRI results

Regions showing significantly greater hemodynamic responses to the unpredictable critical words (*low constraint unexpected*, *high constraint unexpected* and *anomalous*) than to the *expected* critical words are shown in Figure 5, alongside a summary of the MEG source-localized results (reported above) for comparison.

Insert Figure 5 here

Low constraint unexpected *vs.* expected

This contrast revealed a significant hemodynamic effect within the left inferior frontal cortex, but no significant effect within the left temporal cortex (Table 2A). This qualitatively mirrored the pattern of MEG activity detected in the 600-1000ms time window, but the MEG frontal effect was smaller and, as noted above, it did not reach significance.

High constraint unexpected *vs.* expected

This contrast revealed significant hemodynamic effects within the left inferior frontal cortex and the mid-portion of the left superior temporal sulcus. Again, this was qualitatively similar to the MEG effects observed between 600-1000ms, but again the left inferior frontal effect was more extensive in fMRI than in MEG. In addition, fMRI revealed clusters within the left inferior parietal lobule, and left lateral and medial middle/superior frontal cortices (Table 2B).

Anomalous vs. expected

Again, this contrast revealed hemodynamic effects that mirrored the late MEG effects: activity within the left inferior frontal cortex (again more extensive than in MEG) and within the left fusiform gyrus (Table 2C).

Insert Table 2 here

The results of an exploratory whole brain fMRI analysis are reported in Supplementary Figure 5 and Supplementary Table 2.

Discussion

We used multiple neuroimaging techniques to ask whether the principles of dynamic hierarchical predictive coding can explain the location and timing of evoked neural activity produced by expected, unexpected and anomalous words during language comprehension. We showed that, relative to predicted continuations, words carrying unpredicted lexico-semantic information produced larger evoked responses at lower levels of the left fronto-temporal language hierarchy (left temporal cortex), while words that additionally violated higher-order contextual constraints produced activity at higher levels of the hierarchy (left inferior frontal cortex). In a later time window, prediction violations also activated different parts of the temporal cortex depending on whether they resulted in plausible or anomalous interpretations. We first describe the pattern of MEG and ERP effects for each contrast of interest. We then turn to the pattern of activity revealed by fMRI across the four conditions, discussing both its divergence and convergence with the source-localized MEG effects.

Lower-level lexico-semantic prediction error within left temporal cortex is produced by incoming words, regardless of contextual constraint

Consistent with many previous ERP studies[20-22], contextually unexpected words produced a larger N400 between 300-500ms at the scalp surface than expected words. A key claim of predictive coding is that differences in evoked activity between expected and unexpected inputs are driven by the top-down suppression of prediction error to *expected* inputs at *lower levels* of the cortical

hierarchy (*expectation suppression*[6,7]). Our MEG findings support this claim. The evoked effect between 300-500ms localized to multiple regions within left temporal cortex that are known to support lexical and semantic processing. These included left *anterior* temporal cortices (ventral and superior/middle temporal), which function to "bind" widely distributed *semantic* features into distinct concepts[28], and left *mid- temporal* cortices (mid-superior/middle temporal[29,30] and mid-fusiform[31]), which function to *map* orthographic and phonological representations onto meaning (lexical processing).

Previous MEG[32] and intracranial studies[33] have also reported increased activation in temporal cortex to unexpected (*versus* expected) words in the N400 time window. However, in these earlier studies, the unexpected words were often implausible or they violated strong contextual constraints. Using plausible sentences, we showed that, between 300-500ms, the activity evoked by unexpected words within the temporal cortex was very similar in low constraint and high constraint contexts. This provides strong evidence that, instead of reflecting an enhanced response to implausible continuations, or the costs of inhibiting incorrect lexico-semantic predictions, these differences were driven by the top-down *facilitation* of expected lexico-semantic information within the temporal cortex. Specifically, we suggest that, in high constraint contexts, comprehenders incrementally built an event model[14] that generated top-down lexico-semantic reconstructions of expected upcoming words. These reconstructions immediately suppressed the lexico-semantic prediction error produced by new expected inputs.

In addition to these expectation suppression effects within left anterior and mid-temporal cortices, we also observed an MEG effect in the left medial temporal cortex within the same 300-500ms time window, consistent with previous intracranial studies[33]. This medial temporal effect, however, was not only driven by a dipole to the *unexpected* critical words, but also by a dipole in

the opposite direction to the *expected* critical words. We suggest that the dipole to the *unexpected* words reflected a functional role of the left medial temporal cortex (along with anterior lateral temporal regions) in retrieving and binding the semantic features associated with the incoming word[28], possibly supported by "pattern completion" within the hippocampus itself[27]. The dipole to the *expected* words may have reflected a neural "resonance"[34] within medial temporal subpopulations that were already pre-activated prior to encountering the new bottom-up input[35]. The presence of two dipoles going in opposite directions may explain why previous MEG studies have failed to detect effects within the medial temporal cortex within the N400 time window. This is because most MEG studies have used unsigned, rather than signed, dipole values for source localization, and the absolute values of two dipoles going in opposite directions are likely to cancel out.

Higher-level prediction error within left inferior frontal cortex is produced only by words that violate high certainty predictions

A key assumption of the account outlined above is that the top-down lexico-semantic reconstructions that suppress lower-level prediction error are informed by long-term schema knowledge that is relevant to the current message being communicated. Within this hierarchical framework, these schemas are represented at the highest level of the generative hierarchy, and they themselves generate reconstructions that constrain the current event model[15]. During real-world language comprehension, however, messages can change rapidly. In order to continue predicting effectively, comprehenders must be able to recognize event boundaries[36] so that they can rapidly shift the event model by retrieving new high-level schemas[15,18]. Dynamic hierarchical predictive coding makes two important claims regarding these high-level shifts. First, they are triggered by *higher-level prediction error*, which is produced whenever new inputs violate a high confidence

prior belief in the higher-level state[11]. Second, they result in the generation of new top-down reconstructions that provide retroactive feedback to lower levels of the cortical hierarchy, enhancing activity over consistent representations (top-down "sharpening"[4,13]).

Our findings support both these claims. Replicating previous ERP studies[21,22], we found that, relative to expected words, unexpected words produced a *late frontal positivity* ERP effect between 600-1000ms only in high constraint contexts. In MEG, the same contrast revealed activity within the left inferior frontal cortex in this late time window. This was accompanied by a re-activation of the left middle temporal cortex. No late frontal or temporal effects were observed when contrasting expected words with unexpected words in low constraint contexts.

We suggest that in both the low and high constraint contexts, the lower-level lexico-semantic prediction error led comprehenders to infer a new plausible event, resulting in the production of reconstructions that switched off the lower-level lexico-semantic prediction error, thereby attenuating the evoked response within the left temporal cortex at the end of the N400 time window. However, in the high constraint context, this newly inferred event violated a prior high-certainty belief in a different event that had previously been inferred from the context[37,38]. This increased the gain on the new event information, resulting in a *higher-level event prediction error* within the left inferior frontal cortex in the later 600-1000ms time window. This higher-level prediction error initiated the retrieval of a new schema from long-term memory[18], enabling comprephenders to successfully shift their event model, and resolve the error[22,39]. The updated event model, in turn, provided retroactive feedback to the left temporal cortex, enhancing activity over schema-consistent lexico-semantic representations, while reducing activity over incorrectly predicted lexico-semantic information[15]. The top-down nature of this feedback enhancement may explain why, within this late time window, the dipoles within the temporal cortex were of the

opposite polarity to those produced by the bottom-up prediction error within the 300-500ms time window. This account is also consistent with the well-known role of the left inferior frontal cortex in top-down suppression and selection[40].

A breakdown of predictive coding to anomalous words

This hierarchical predictive coding framework posits that higher-level prediction error should also be produced if a newly updated state is inconsistent with prior reconstructions received from a still higher cortical level. Critically, however, if this higher-level prediction error cannot be resolved because the input is *incompatible* with the constraints of the generative model, or with alternative models stored in long-term memory, then the late retrieval and top-down sharpening mechanisms described above should break down. For example, after encountering a semantic anomaly, it is impossible to retrieve a new schema that can explain the input, and so the conflict between the top-down reconstructions produced by the current schema and the bottom-up lexico-semantic prediction error cannot be resolved. This will therefore lead to (a) a failure to switch off prediction error at even lower levels of the cortical hierarchy (perceptual reanalysis), and/or (b) new learning in order to explain the input[18,19].

Our findings are broadly consistent with this account. First, at the scalp surface, the anomalous words produced an N400 that was larger than that produced by the plausible unexpected continuations (this difference was less prominent in ERP than in MEG, see Supplementary Materials section 3). MEG localized the activity within this 300-500ms time window not only to the left temporal cortex, but also to the left inferior frontal and anterior cingulate cortices. We suggest that the inferior frontal activity reflected the production of an early event prediction error (because the impossible event fell outside the range of event reconstructions generated by the current schema), and that the enhanced activity within the temporal cortex

resulted from a failure to settle on a higher-level interpretation within this time window, and therefore to switch off lower-level lexico-semantic prediction error. The surprising failure to minimize prediction error within the N400 time window may have led to the early recruitment of the anterior cingulate cortex[41].

Second, within the late time window (600-1000ms), the semantic anomalies also produced a *late posterior positivity/P600* ERP effect, which is often triggered by high-level linguistic conflict[22,25,26], and thought to reflect a lower-level reanalysis of the input[22,25,39]. Consistent with this proposal, in MEG we observed sustained activity within posterior fusiform cortex, which supports sub-lexical orthographic processing[9]. We suggest that this "orthographic reanalysis" arose because the brain failed to settle on a single lexico-semantic representation, and therefore failed to produce reconstructions that switched off orthographic prediction error at this still lower level of the linguistic hierarchy.

Finally, throughout the 600-1000ms window, semantic anomalies also produced an effect within the medial temporal cortex. This region is highly interconnected with the hippocampus, which plays a major role in detecting associative and contextual novelty[42], primarily through a "comparator function" that tracks the magnitude of prediction violations[43], thereby paving the way towards new learning[44]. Consequently, this medial temporal activation may have indirectly supported updates in the parameters of the generative model that allowed comprehenders to adapt to anomalous inputs (consistent with known links between the *late posterior positivity/P600* and adaptation[45]). Alternatively, it may have supported the learning of new schemas from the novel anomalous inputs[18,27,46]. Both of these interpretations are consistent with the important computational role of prediction error in bridging comprehension and learning[38].

Convergence and divergence between fMRI and MEG/EEG

A second goal of this study was to understand how hemodynamic activity, recorded using fMRI, converged and diverged from the pattern of ERP and source-localized MEG effects produced in the same paradigm and in the same group of participants.

The clearest discrepancy between the fMRI and MEG/EEG data was that fMRI failed to detect the ERP and MEG effects observed in the N400 time window (300-500ms). For example, even though the contrast between the *low constraint unexpected* and *expected* critical words revealed significant MEG effects within left lateral, ventral and medial temporal cortices (corresponding to the N400 effect), the same contrast in fMRI showed no significant differences within the temporal cortex. The contrast between *high constraint unexpected* and *expected* critical words did reveal some hemodynamic activity within the left middle temporal cortex, and the contrast between *anomalous* and *expected* words revealed activity within the fusiform cortex. However, both these effects can be explained by later MEG/EEG activity, from 600-1000ms.

Although striking, this insensitivity of the hemodynamic response to N400 activity is not altogether surprising. Others have noted that MEG is more likely to localize top-down contextual effects to the temporal lobe than fMRI[30]. In addition, multimodal neuroimaging studies of semantic priming report fMRI effects that are much smaller and less robust than MEG N400 effects[47,48]. A likely reason for these discrepancies is that, while MEG and EEG are highly sensitive to brief, time-locked activity[49], fMRI is relatively blind to transient responses that are associated with the initial stages of feedforward activity[50,51].

Conversely, because the hemodynamic response integrates activity across multiple successive time windows, the signal is dominated by activity at later stages of processing. Indeed, the clearest pattern of convergence between fMRI effects and source-localized MEG effects was within the 600-1000ms time window. Both techniques revealed effects within the left

frontal/middle temporal cortex to *high constraint unexpected* (versus *expected*) critical words, and within the left frontal/fusiform cortex to *anomalous* (versus *expected*) critical words. Consistent with previous studies[50,52], activity within the prefrontal cortex was more robust and extensive in fMRI than MEG (note that the left frontal effect to *low constraint unexpected* versus *expected* critical words was significant in fMRI but not in MEG). This may be because MEG is insensitive to radial sources from gyri, and because tangential sources on opposing sides of sulci can cancel out[53]. It is also possible that the hemodynamic response was less time-locked to the critical words, and that it detected activity past 1000ms. Nonetheless, given the challenges of solving the inverse problem, the qualitative similarity between the MEG activity detected within the late time window and the hemodynamic response in the same contrasts provides independent corroborating evidence for the late MEG source-localized effects.

Conclusion

By tracking the timecourse and localization of evoked neural activity to incoming linguistic information, we showed clear dissociations in the production of prediction error at different levels of the left fronto-temporal cortical hierarchy. Consistent with classic predictive coding frameworks, lower-level prediction error, produced by the lexico-semantic features of individual words, was localized to lower levels of the hierarchy (left temporal cortex). Critically, as predicted by *hierarchical* and *dynamic* predictive coding, higher-level prediction error, produced by whole events, was observed at higher levels of the hierarchy (left inferior frontal cortex), and was modulated by prior certainty of the higher-level event representation (precision-weighting). Finally, when comprehenders were able to resolve this high-level error by shifting to a new plausible interpretation, this led to feedback activation of the temporal cortex at a later stage of processing (top-down "sharpening"). Taken together, these findings provide strong evidence that

a basic computational principle – the minimization of prediction error – can explain the functional dynamics of feedforward and feedback activity during human language comprehension.

Methods

Materials

Participants read four types of three-sentence scenarios, each with a critical noun in the third sentence, see Table 1. In the *expected* scenarios, the critical word was predictable following a high constraint context. In each of the three other conditions, the critical word was unpredictable, but each for a different reason. In the *low constraint unexpected* scenarios, the critical word was plausible but unpredictable because it followed a low constraint context. In the *high constraint unexpected* scenarios, the critical word was plausible but unpredictable because it violated a high constraint context. In the *anomalous* scenarios, the critical word followed a high constraint context and violated the animacy selectional constraints of the preceding verb (which constrained either for animate or inanimate nouns).

The stimuli were based on those used in a recent ERP study[22]. A full description is provided there as well as in Supplementary Materials, section 1. Briefly, in each scenario, the discourse context was either high constraint (average cloze probability of the most probable word: 68%), or low constraint (average cloze: 22%), as quantified in a cloze norming study that was carried out in participants recruited through Amazon Mechanical Turk (see Supplementary Materials section 1 for details). 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 constituted 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. The verb was always relatively non-constraining in minimal contexts (cloze probability of the most probable word was below 24%, as quantified in another cloze norming study in which participants recruited through Amazon Mechanical Turk were presented with just a proper name, the verb, and a determiner, see Supplementary Materials section 1 for details).

To create the *expected* scenarios, each *high constraint* context was paired with the noun with the highest cloze probability for that context. To create the *high constraint unexpected* scenarios, each *high constraint* 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 *low constraint unexpected* scenarios, the same unexpected noun was paired with the *low constraint* context, again ensuring that it was plausible in relation to this context. To create the *anomalous* scenarios, each *high constraint* context was paired with a noun that violated the animacy selectional constraints of the verb. In all scenarios, the critical noun was followed by three additional words to complete the sentence. This gave rise to our four conditions of interest.

Table 1 shows the stimulus characteristics of the critical nouns in each of the four scenario types. Critical words in the *expected* scenarios had fewer letters, smaller orthographic neighborhoods and were more frequent than in the unpredictable scenarios (all ts > 5, ps < 0.001). However, all these values were matched between the three types of unpredictable scenarios (all ts < 1.5, ps > 0.10). In addition, the semantic relatedness between the critical words and their prior contexts (operationalized using Semantic Similarity Values, SSVs, extracted using Latent Semantic Analysis, LSA (http://lsa.colorado.edu/, *term-to-document* with default settings) were matched between the three types of unpredictable scenarios (all ts < 1, ps > 0.10 for all pairwise comparisons). As expected, these values were greater in the *expected* scenarios than in the three types of unpredictable scenarios (all ts > 8, ps < 0.001).

Each participant read 150 experimental scenarios: 25 *expected*, 50 *low constraint unexpected*, 25 *high constraint unexpected*, and 50 *anomalous*. In addition, each participant read an additional 50 filler anomalous scenarios with low constraint contexts. This overall list composition ensured that each participant viewed 50% plausible and 50% anomalous scenarios, and that critical words were just as likely to be plausible following high constraint and low constraint contexts. Counterbalancing worked so that the combination of the verbs and critical words in all four conditions appeared across four lists, and, within each list, no participant read the same combination of verb and critical word more than once.

Overall procedure: MEG/EEG and MRI sessions

Participants took part in two separate experimental sessions: one for simultaneous MEG/EEG recordings, and one for structural and functional MRI (s/fMRI) recordings. We took several steps to minimize any confounds due to repetition of stimuli across sessions: (1) At least two weeks intervened between MEG/EEG and s/fMRI session; (2) The order of participation was fully counterbalanced across sessions (participants' gender was included in this counterbalancing scheme); (3) Each participant viewed a different list in the MEG/EEG and fMRI session, which reduced repetition of contexts or critical words. For any contexts that did repeat across the two sessions, we constructed versions of the lists that changed proper names and small details of these contexts (such that they had minimal impact on cloze probability/lexical constraint). Full details of these modified stimuli and the counterbalancing scheme are provided in Supplementary Materials section 1.

Participants

All participants were native speakers of English (with no other language exposure before the age of 5), 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. Written consent was obtained before participation following the guidelines of the Massachusetts General Hospital Institutional Review Board.

Here we report the results of 32 separate MEG/EEG datasets (16 females, mean age: 23.4; range: 18-35) and 31 separate fMRI datasets (16 females, mean age: 23.5; range: 18-35). Twenty-nine of these participants (15 females, mean age: 23.5; range: 18-35) participated in both the MEG/EEG and fMRI sessions.

Originally, a total of thirty-five participants were recruited for the study. Thirty-three took part in both the MEG/EEG and the fMRI sessions (17 females, mean age: 24.5 years; range: 18-35 years); the remaining two participants participated in fMRI but failed to return for the MEG/EEG session. Of the 33 participants who took part in the MEG/EEG session, we excluded one dataset because of technical problems. Of the 35 participants who took part in the fMRI session, we excluded four participants due to technical problems, termination of the experiment by the participant, or excessive movement (for further details of cutoff criterion, see fMRI analysis below).

Stimuli presentation and task

In both the MEG/EEG and fMRI sessions, stimuli were presented using PsychoPy 1.83 software[54] 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 in the MEG/EEG session, and for 350ms in the fMRI session, followed by a 100ms ISI in both sessions. Then the third sentence was presented word by word (each word for 450ms, 100ms ISI).

In both the MEG/EEG and the fMRI session, participants' task was to judge whether or not the scenario "made sense" by pressing one of two buttons (response fingers were counterbalanced across participants) after seeing a "?", which appeared after each scenario (1400ms with a 100ms ISI). This task encouraged active coherence monitoring during online comprehension and was intended to prevent participants from completely disregarding the anomalies (see[55] for evidence that detecting anomalies is necessary to produce a neural response). In addition, following approximately 24/200 trials (semi-randomly distributed across runs), 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.

In the MEG/EEG session, 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).

In the fMRI session, between each trial, a green fixation (++++) was presented for a duration that ranged from 2-18 seconds (average 6.2 seconds). This was to optimize the deconvolution of the event-related hemodynamic response function, as determined using the OptSeq2 algorithm (see https://surfer.nmr.mgh.harvard.edu/optseq). In order to keep the stimulus presentation synced with the scanner, preventing an accumulation of timing errors as a result of waiting for the screen to refresh during stimulus presentation, we used a "non-slip" routine timer in PsychoPy.

In both the MEG/EEG and the fMRI session, stimuli were presented over eight runs, each with 25 scenarios. Runs were presented in random order in each participant. Participants took part in a short practice session before both sessions 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 2 orthogonal planar gradiometers and 1 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 <20 kΩ at all scalp sites, at <10 kΩ at mastoid sites, and at <30 kΩ 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 and functional MRI data acquisition

Structural and functional MRIs were acquired using a 3T Siemens Trio scanner with a 32-channel head coil. Each of the eight fMRI runs lasting for approximately eight minutes, and in each run, we acquired 238 functional volumes (87 axial slices, 1.5mm slice thickness, 204mm field of view, in-plane resolution of 136mm) with a gradient-echo sequence (Time to Repetition, TR: 2s; echo time: 30ms; echo spacing: 0.65ms; flip angle: 80 degrees). The acquisition angle was raised to approximately 20 degrees off the AC-PC line in order to reduce distortion along the anterior temporal lobe. Slices were acquired, three at a time, using an interleaved simultaneous multi-slice protocol with a GeneRalized Autocalibrating Partial Parallel Acquisition (GRAPPA) acceleration factor of 2. Following the eight functional runs, we acquired T1-weighted high-resolution structural images (1mm isotropic multi-echo magnetization-prepared rapid gradient-echo, MP-RAGE; TR: 2.53s; flip angle: 7 degrees; 4 echoes with TE: 1.69ms, 3.55ms, 5.41ms, and 7.27ms).

Preprocessing and initial data analysis

EEG preprocessing and individual averaging

EEG data were analyzed using the Fieldtrip software package[56] in the Matlab environment[57]. 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), downsampled 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% of trials were removed from each condition (equally distributed across the four conditions: $F_{(3,93)} = 0.88$, $p = 0.45$, $\eta^2 = 0.028$), yielding, on average, 23 trials in the *expected* and *high constraint unexpected* conditions, and 46 trials in the *low constraint unexpected* and *anomalous* conditions. Finally, the data of bad channels were interpolated using spherical spline interpolation[58]. In each participant, at each site, we then calculated ERPs, time-locked to the onset of critical words, in each of the four conditions, applying a -100ms pre-stimulus baseline. After that, we averaged these voltages across all time points and electrode sites within each of three spatiotemporal regions of interest to carry out statistical analyses, as described below.

MEG preprocessing, individual averaging and sensor-level visualization

MEG data were analyzed using version 2.7.4 of the Minimum Norms Estimate (MNE) software package in Python[59]. 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[59]. 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, 16% trials in each condition were removed (equally distributed across the four conditions: $F_{(3,93)} = 1.54$, $p = 0.21$, $\eta^2 = 0.047$), yielding, on average, 21 artifact-free trials in

the *expected* and *high constraint unexpected* scenarios, and 42 artifact-free trials in the *low constraint unexpected* and *anomalous* scenarios.

In each participant, in each run, 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 four conditions, applying a -100ms pre-stimulus baseline. We averaged the ERFs across runs in sensor space, interpolating the bad sensors using spherical spline interpolation[58]. 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[59] to estimate the sources of the ERFs evoked by critical words in each of the four conditions, on each participant's reconstructed cortical surface using Minimum-Norm Estimation (MNE[60]).

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 runs. 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[59].

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 run. We chose to estimate 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 dipole for further analyses for two reasons. First, this approach allowed us to include all trials in each of our four 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 *low constraint unexpected* and the *anomalous* conditions, which had twice as many trials as the *expected* and the *high constraint unexpected* 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.

Then, for each condition in each run, we computed noise-normalized dynamic Statistical Parametric Maps (dSPMs[61]) on each participant's cortical surface at each time point. The obtained dSPM values were then averaged across runs within each participant. Finally, the source estimates of each participant were morphed on the FreeSurfer average brain "fsaverage"[62] for group averaging and statistical analysis, as described below.

<u>fMRI preprocessing and individual analysis</u>

Functional volumes were preprocessed using SPM12 in the Matlab environment[57]. In each participant, the first volume of each run was realigned to the first volume of the first run, and all images within a given run were realigned to the first image of that run. The resulting images were slice-time corrected and the mean of the functional images was co-registered with the individual's structural MPRAGE image. The structural images were segmented into grey and white matter, and

the functional images were spatially normalized to the standard Montreal Neurological Institute (MNI) template. Images were smoothed with an 8mm full width at half maximum (FWHM) Gaussian kernel.

We next used the Artifact Detection Toolbox[63] to calculate the percentage of time points/volumes (across all runs) in which the composite motion was greater than 1.5mm. If more than 5% of the volumes in any run met this criterion, we excluded that run. This resulted in the exclusion of one participant (all runs met this criterion), the exclusion of two runs in a second participant, and the exclusion of a single run in a third participant. For all the remaining runs that were included in the analysis, we used the same toolbox to create nuisance regressors associated with any volume in which the composite motion was greater than 1.5mm. Over these remaining runs, less than 0.22% of volumes/time points per participant, on average, were "marked" by one of these extra regressors.

At the first level of analysis, each run was modeled with a design matrix that included regressors for each condition. These were modeled as epochs from the onset of the critical word in the third sentence until the offset of the sentence-final word. Additional regressors were included for the contexts (from the onset of the first sentence until the onset of the critical word, not differentiating between conditions), and for the question mark events (from the onset of the question mark until the onset of the inter-trial fixations). All these regressors were convolved with a canonical hemodynamic response function (HRF). The model also included the additional nuisance regressors created using the Artifact Detection Toolbox[63], as described above. First level contrasts were defined to take to the second level for random effects group analysis: each condition (contrast value of 1) versus an implicit baseline (contrast value of 0).

Group-level statistical analysis and hypothesis testing

Planned comparisons for all three methods

For ERP, MEG and fMRI analyses, we carried out planned *a priori* statistical comparisons between neural activity evoked by each type of unpredictable critical word (*low constraint unexpected*, *high constraint unexpected*, *anomalous*) and the *expected* critical words.

Statistical analysis of ERP data

To analyze the ERP data, our planned comparisons (paired t-tests) were carried out on voltages that were averaged across all time points and electrode sites within each of three spatiotemporal regions of interest. These regions were selected, *a priori,* to capture the N400, the late frontal positivity and the late posterior positivity/P600 ERP components. They were the same as those used in our previous ERP study using overlapping stimuli in a different group of participants[22]. 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 all 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 all sampling points between 600-1000ms.

Statistical analysis of MEG source-level data

To analyze the source-level MEG data, we carried out our planned statistical comparisons over a large left-lateralized search region that included classic language-related areas as well as other regions of interest (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). This search area was defined on the Desikan-Killiany Atlas[64] and is illustrated in Figure 6. The correspondence between names of the anatomical regions given in Figure 6 (as well as in Figure 5 and Table 2) and the nomenclature of the Desikan-Killiany regions is given in Supplementary Table 1. Within this search region, we examined activity within three 200ms 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 positivity ERP effects. To account for multiple comparisons, we tested hypotheses using permutation-based cluster mass procedures based on[65] and modified as described next.

Insert Figure 6 here

For each contrast of interest, within each time window of interest, we carried out pairwise t-tests on the signed estimated dSPM values at each vertex and at each time point. Instead of using the resulting signed t-values to compute our cluster-level statistic, we used unsigned -log-transformed p-values. This is because a single neuroanatomical source that is located on one side of a sulcus can appear on the cortical surface as adjacent groups of dipoles of opposite polarity (outgoing and ingoing) because of signal bleeding to the other side of the sulcus[66]. This is clearly apparent in the activation maps that show the signed dSPM values at each location in each condition (see Figures 3 and 4): positive dSPM values, corresponding to outgoing dipoles (shown in red), and negative dSPM values, corresponding to ingoing dipoles (shown in blue), often appear on either side of a sulcus. The use of unsigned p-values therefore ensured that adjacent effects of opposite signs were treated as a single cluster/single underlying source. Within each time window

of interest, 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.

In order to account for multiple spatial comparisons across the search area, we subdivided it into 140 equal-sized patches[67], shown in Supplementary Figure 1. Within each patch, we took the average of the -log-transformed p-values across all time points within each time window of interest (300-500ms, 600-800ms, 800-1000ms) as our cluster statistic. We then carried out exactly the same procedure as that described above, but this time we randomly assigned dSPM values between the two conditions for a given contrast. This was repeated 10,000 times. For each randomization, we took the largest cluster mass statistic across all spatial patches, and in this way created a null distribution for the cluster mass statistic. To test our hypotheses at each spatial patch in each time window of interest, we compared the observed cluster-level statistic for that patch against the null distribution. If our observed cluster-level statistic fell within the highest 5.0% of the distribution, we considered it to be significant. Note that this cluster-based method allowed us to account for temporal and spatial discontinuities in effects (resulting from noise). However, it constrains any statistical inference to the spatial resolution of each patch and to the temporal resolution of our *a priori* time windows.

In order to illustrate the results, we projected the averaged uncorrected -log10 transformed p-values ($p < 0.05$) at each vertex on to the "fsaverage" brain. We use circles to indicate any spatial patches in which we observed a significant cluster, grouping these areas by the anatomical regions shown in Figure 6 and listed in Supplementary Table 1.

Finally, in addition to carrying out these tests over our *a priori* left-lateralized search region of interest, we also carried out more exploratory analyses using the same procedure over an

analogous search region over the right hemisphere. We report these results in Supplementary Figures 2, 3 and 4.

Statistical analysis of fMRI data

At the group (second) level of analysis, we constructed a repeated measures ANOVA model that included the within-subject effects (31 regressors) and one regressor for every condition (versus implicit baseline). We used this model to create Statistical Parametric Maps (SPMs) of the t-statistics for each contrast of interest.

We report the results of directional t-tests for regions that showed more hemodynamic activity to each type of unpredictable critical word than to the *expected* critical words (*low constraint unexpected* > *expected*; *high constraint unexpected* > *expected*; *anomalous* > *expected*) within the same *a priori* left lateralized search region of interest as that used in the MEG analysis. For the fMRI analysis, this search region was defined in Montreal Neurological Institute (MNI) volume space, using the AAL atlas[68]. The correspondence between the names of the anatomical regions illustrated in Figure 6 and the nomenclature of the Tzourio-Mazoyer regions is given in Supplementary Table 1.

To account for multiple comparisons, we set an initial voxel-level threshold of $p < 0.001$ (whole brain), and we inferred significance if clusters within the search region reached a cluster-level family-wise error-corrected (FWE) threshold of $p < 0.05$, using a small volume correction (SVC)[69]. We report the size and the p-value of each cluster (as a whole), as well as the z-scores and uncorrected p-values of the individual peaks within that cluster. All coordinates reported are in MNI space. Although statistical analysis was carried out in MNI volume space, for maximal comparability to the MEG results, we converted the t-maps to right-anterior-superior (RAS) space and plotted the results on the "fsaverage" brain surface[62].

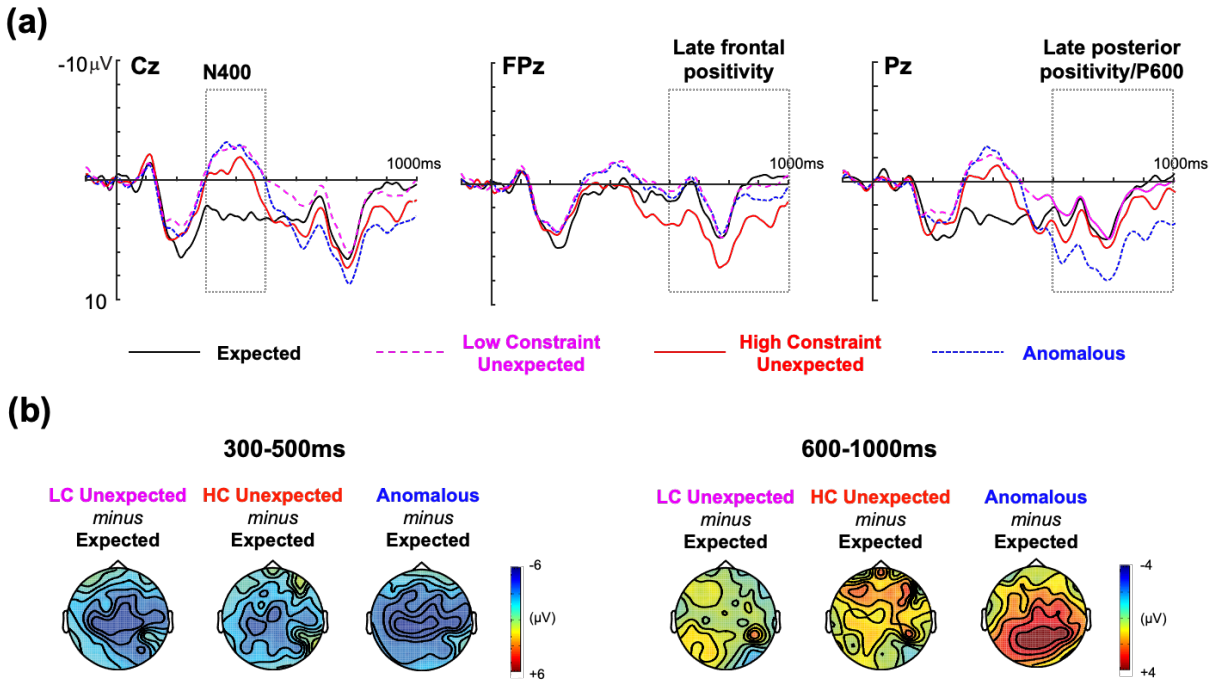
In addition to carrying out analyses over our *a priori* left-lateralized search region of interest, we also carried out more exploratory whole brain analyses that included all brain regions. We report these results in Supplementary Figure 5 and Supplementary Table 2.

Acknowledgements

This work was funded by the National Institute of Child Health and Human Development (R01 HD082527 to G.R.K.). We thank Ross Mair for his help with fMRI sequences. 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, Jasmine Falk and Rebeca Becdach for their help in making figures. 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.

Figures

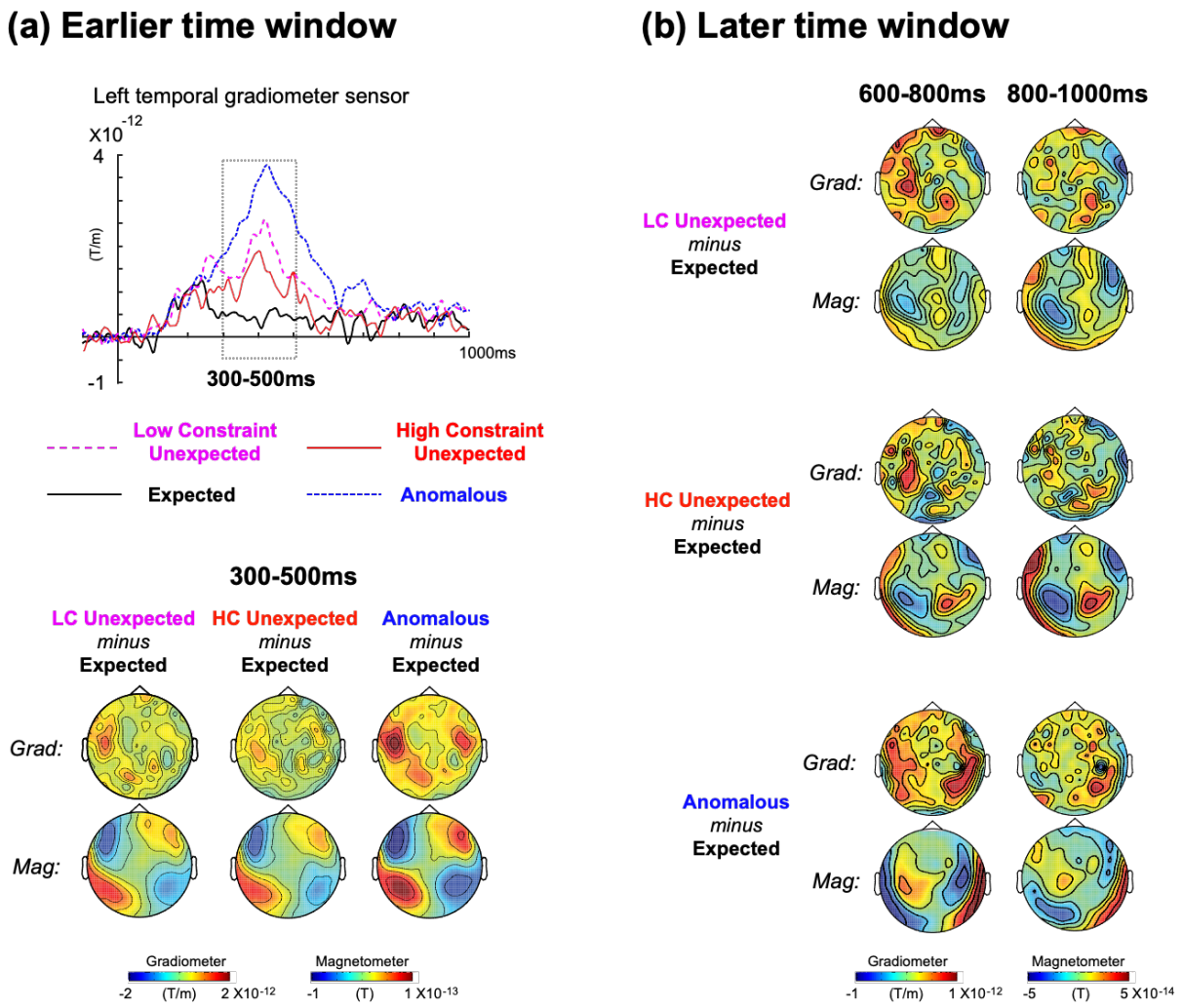
Figure 1. ERP results.



(a) Grand-averaged ERP waveforms elicited by critical words in each of the four conditions, shown at three representative electrode sites: Cz, FPz and Pz. *Expected*: solid black line; *Low Constraint Unexpected*: dashed magenta line; *High Constraint Unexpected*: solid red line; *Anomalous*: dashed blue line. Negative voltage is plotted upwards. 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 are indicated using dotted boxes.

(b) Voltage maps show the topographic distributions of the ERP effects produced by contrasting each of the three types of unpredictable critical words with the *expected* 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.

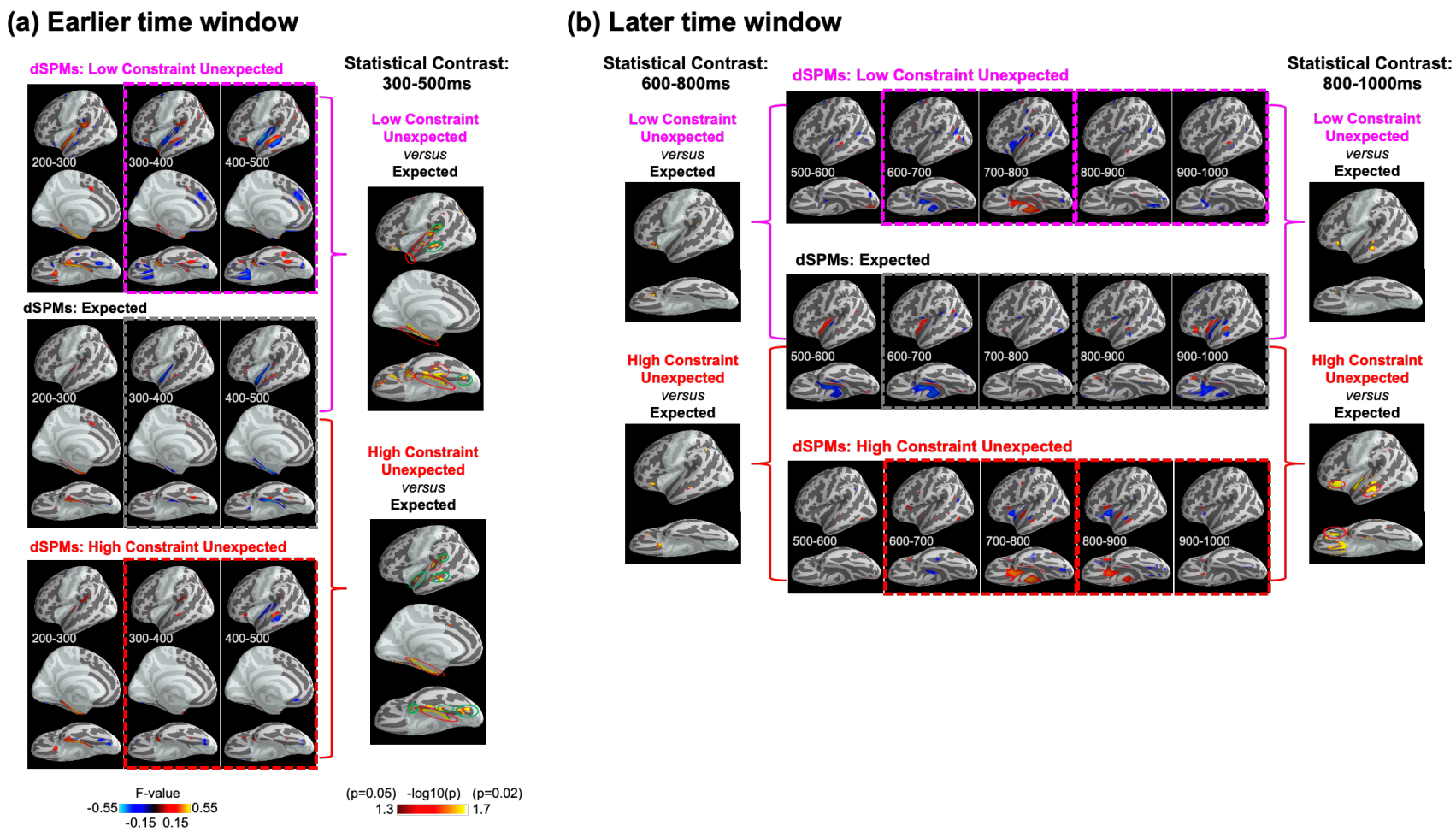
Figure 2. MEG sensor-level results.



(a) Earlier time window. *Top:* Grand-averaged event-related magnetic fields produced by critical words in each of the four 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 each of the three types of unpredictable critical words with the *expected* critical words between 300-500ms. In all three contrasts, the distribution of the MEG N400 effect was maximal over temporal sites, particularly on the left.

(b) Later time window. MEG Gradiometer (Grad) and Magnetometer (Mag) sensor maps show the topographic distributions of the MEG effects produced by contrasting each of the three types of unpredictable critical words with the *expected* 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 each type of plausible unexpected word and the *expected* critical words produced magnetic fields with similar spatial distributions, but the Magnetometer maps suggest that the effect was stronger for the contrast between the *high constraint unexpected* and the *expected* critical words, than for the contrast between the *low constraint unexpected* and the *expected* critical words. The contrast between the *anomalous* and *expected* critical words revealed the strongest effects, with a somewhat distinct spatial distribution of sensor-level activity.

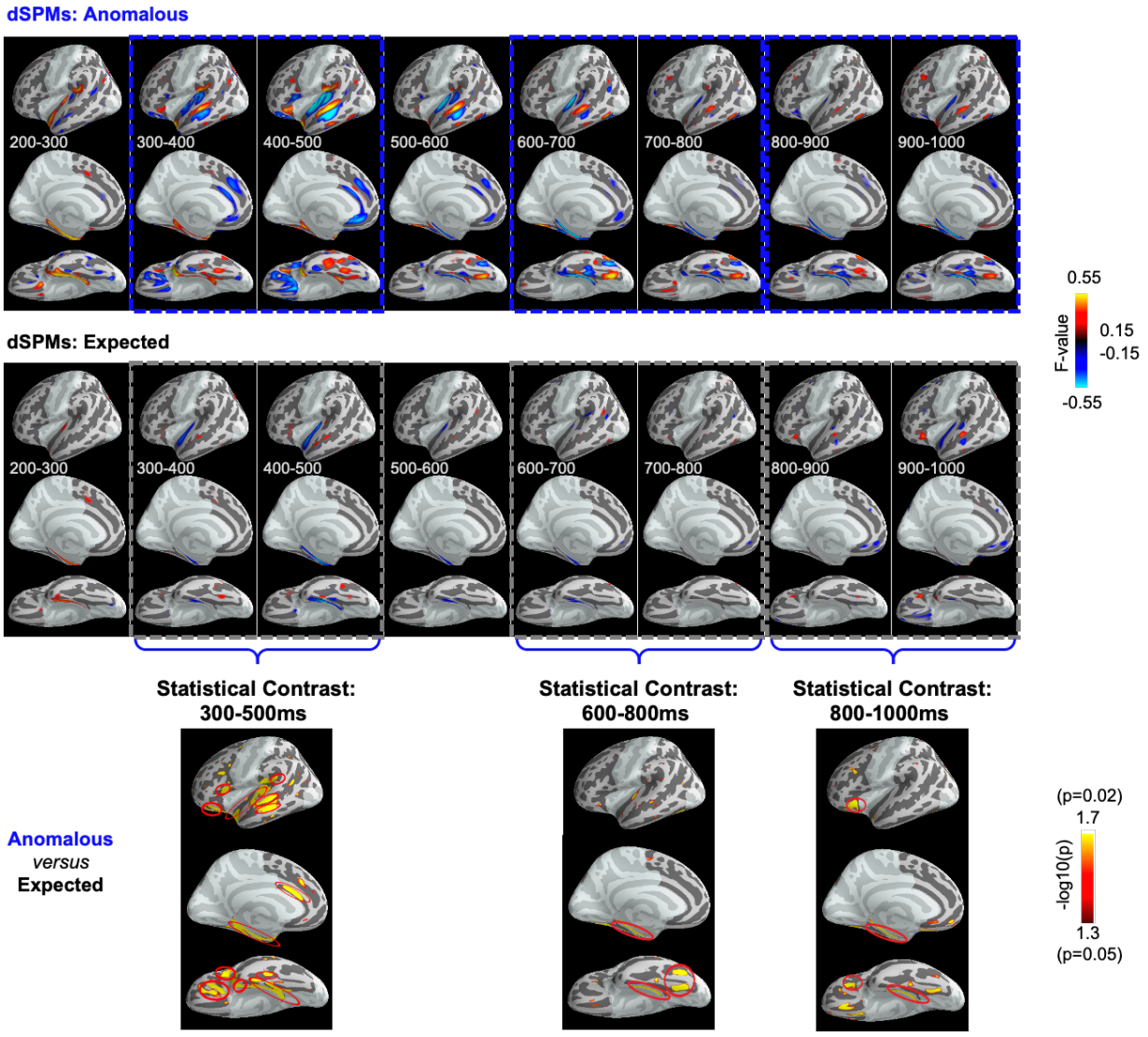
Figure 3. MEG source-level activity produced by the unexpected plausible and the expected critical words.



(a) Earlier time window. *Left:* Signed dSPMs produced by the *low constraint unexpected*, the *expected*, and the *high constraint unexpected* critical words, shown at 100ms intervals from 200 until 500ms. *Right:* Statistical maps contrasting the *low constraint*

unexpected and *expected* critical words, and the *high constraint unexpected* and *expected* critical words within the 300-500ms (N400) time window of interest. Red circles indicate activity that reached cluster-level significance for each contrast individually. Green circles indicate activity that reached significance in an analysis that combined the two types of plausible unexpected critical words and contrasted the resulting activity with that produced by the *expected* critical words. Because previous ERP work had consistently shown that these two contrasts produce similar effects within the N400 time window[20-22], we carried out this analysis in order to increase power. **(b) Later time window:** Signed dSPMs produced by the *low constraint unexpected*, the *expected*, and the *high constraint unexpected* critical words, shown at 100ms intervals from 500 until 1000ms. Statistical maps contrasting each type of plausible unexpected critical word with the *expected* critical words are shown between 600-800ms (left) and between 800-1000ms (right). Red circles indicate regions that reached cluster-level significance in each contrast. All dSPMs are thresholded at 0.15, with red indicating outgoing dipoles and blue indicating ingoing dipoles. Both dSPMs and contrast maps are displayed on the FreeSurfer average surface, "fsaverage"[62].

Figure 4. MEG source-level activity produced by the anomalous and the expected critical words.

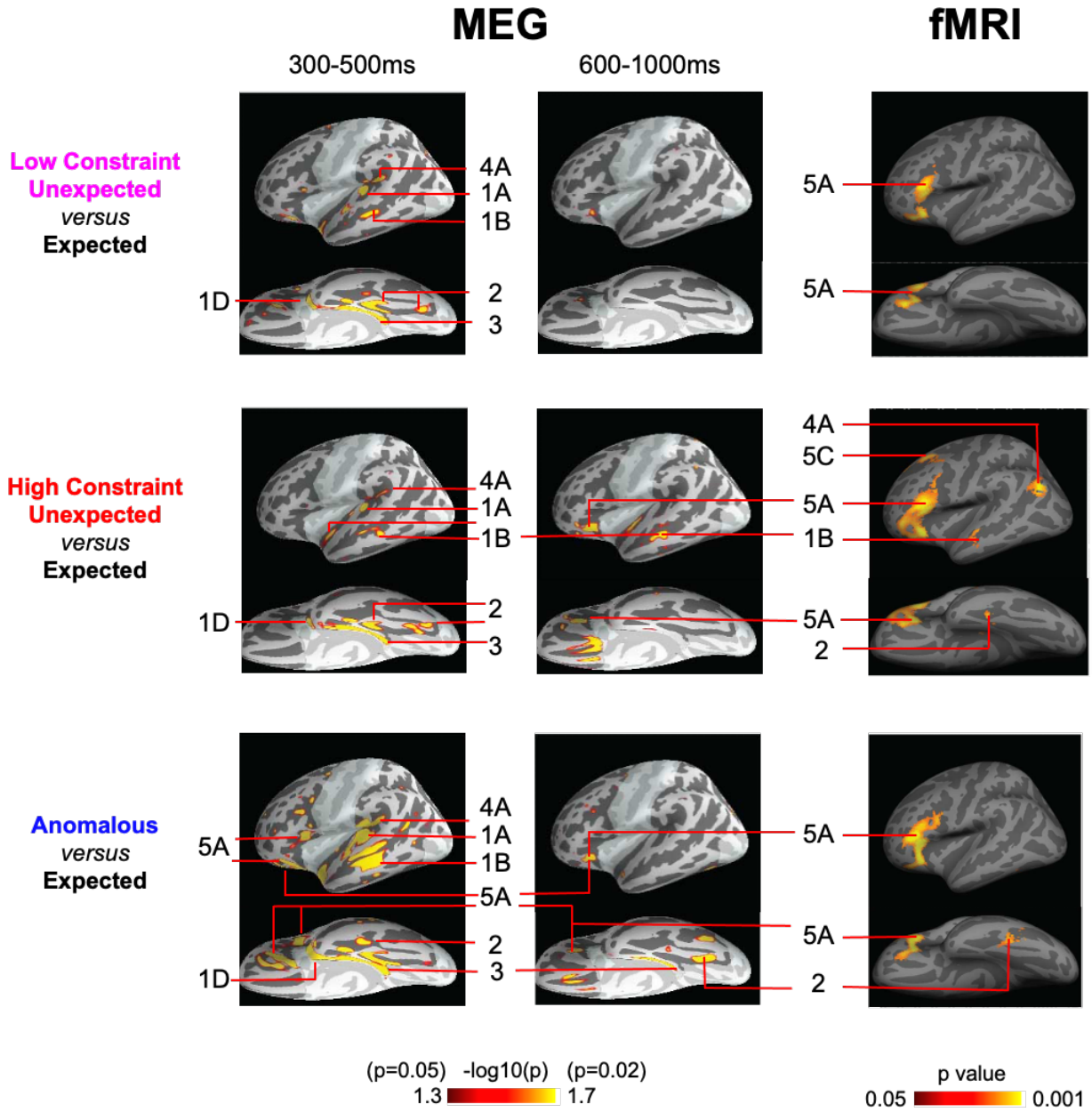


Top and middle rows: Signed dSPMs produced by the *anomalous* and *expected* critical words, shown at 100ms intervals from 200ms until 1000ms.

Bottom row: Statistical maps contrasting the *anomalous* and *expected* critical words within our three *a priori* time windows of interest: 300-500ms, 600-800ms and 800-1000ms. Red circles indicate regions that reached cluster-level significance. All dSPMs are thresholded at 0.15, with

red indicating outgoing dipoles and blue indicating ingoing dipoles. Both dSPMs and contrast maps are displayed on the FreeSurfer average surface, “fsaverage”[62].

Figure 5. FMRI results, together with summarized MEG source-localized effects for each contrast.

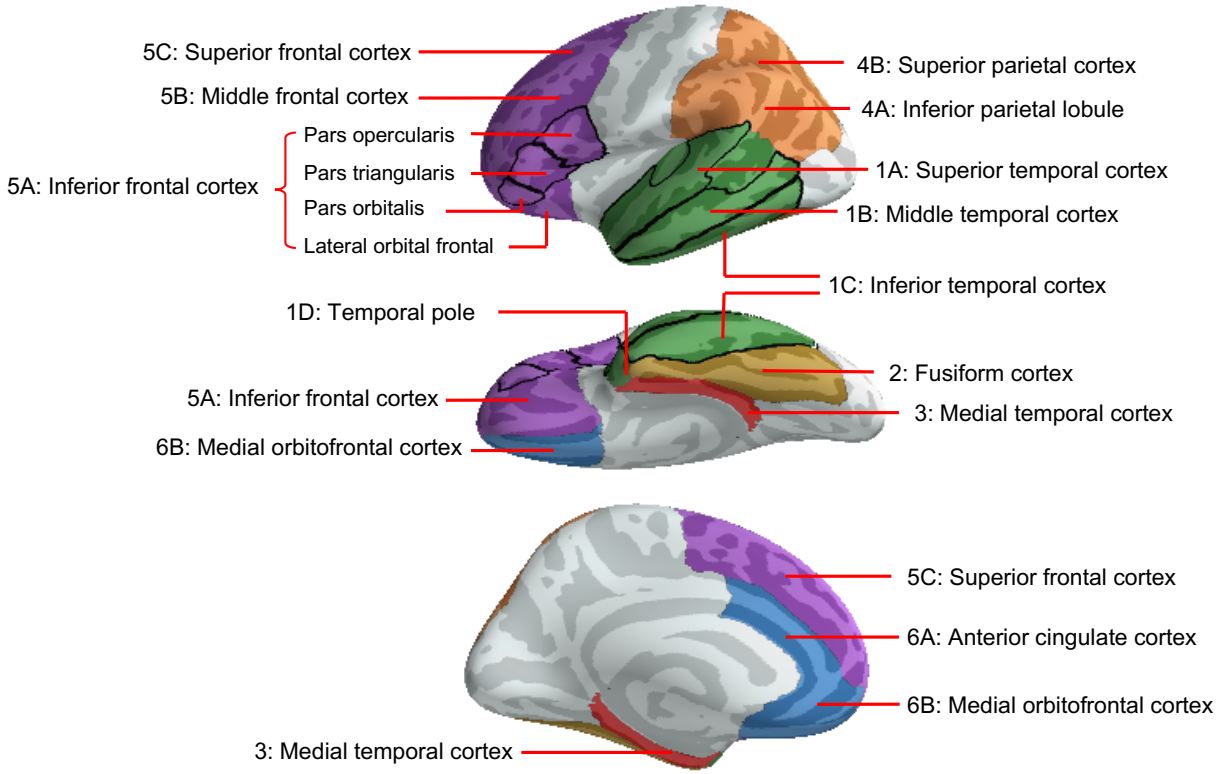


Right column: FMRI statistical maps showing hemodynamic activity that was significantly greater to critical words in each of the three unpredictable conditions (*low constraint unexpected*, *high constraint unexpected*, *anomalous*) than to critical words in the *expected* condition. All activity indicated reached a cluster-level significance threshold after family-wise error (FWE) correction

of $p < 0.05$, small volume corrected (SVC)[69] over the search region of interest (shown in Figure 6). The numbers correspond to the numbering of the regions shown in Figure 6 and in Supplementary Table 1. They also correspond to the regions listed in Table 2, which provides full details of the fMRI results. Although fMRI analyses were carried out in MNI volume space, the results are plotted on the left lateral and ventral FreeSurfer average surfaces ("fsaverage"[62]) to facilitate direct comparisons with the MEG results.

Left and middle columns: To facilitate comparisons between the fMRI results and the source-localized MEG results, the MEG source-localized effects between 300-500ms (left column) and between 600-1000ms (right column) are shown for each contrast of interest, displayed with a vertex-wise threshold of $p \leq 0.05$ (p-values: -log10 transformed). The full presentation of these MEG results is given in Figures 3 and 4. The patterns of fMRI activity were qualitatively similar to the patterns of MEG activity within the late time window, although, within the prefrontal cortex, the hemodynamic effects were more extensive and robust than the effects detected by MEG.

Figure 6. Left-lateralized search region used to carry out MEG and fMRI statistical analysis.



For the MEG statistical analysis, these regions were defined on the "fsaverage" FreeSurfer surface[62] using the Desikan–Killiany atlas[64]. For the fMRI analysis, they were defined in Montreal Neurological Institute (MNI) volumetric space using the automated anatomical labeling (AAL) atlas[68]. In this figure, all regions are displayed on the fsaverage surface. Supplementary Table 1 lists the correspondence between the names of the regions indicated here, and the nomenclature of the equivalent regions in the Desikan–Killiany and AAL atlases.

Tables

Table 1. Examples of the four experimental conditions together with stimuli characteristics.

Scenario Type	Example	*Lexical constraint	**Cloze	+SSV	No. of Letters	++Frequency	^Orthographic neighborhood	^^Concreteness
1. *Expected*	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 swimmers...	69% (14%)	69% (14%)	0.18 (.18)	5.69 (1.60)	1.53 (0.66)	1.93 (0.56)	4.30 (0.69)
2. *Low Constraint Unexpected*	Eric and Grant received the news late in the day. They mulled over the information, and decided it was better to act sooner rather than later. Hence, they cautioned the trainees...	19% (9%)	0.6% (1%)	0.01 (.05)	7.41 (2.33)	0.64 (0.86)	2.58 (0.89)	4.08 (0.72)
3. *High Constraint Unexpected*	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 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)
4. *Anomalous*	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 drawer...	67% (15%)	0% (0%)	0.01 (.05)	7.11 (2.04)	0.81 (0.85)	2.47 (0.81)	4.21 (0.65)

Scenarios were created around the same verb (here, "cautioned"). The critical word in each of the example sentences is underlined (although this was not the case in the experiment itself). The final sentence continued with three additional words, as indicated by the 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 the cloze norming study (see Supplementary Materials, section 1), and tallying the proportion of participants who provided this completion.
**Cloze probabilities of critical words were calculated based on the percentage of respondents providing the critical noun used in the experiment.
+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 (LSA).
++Log Frequency values, retrieved from the English Lexicon Project.
^Orthographic Levenshtein Distance values, retrieved from the English Lexicon Project.
^^Concreteness ratings, retrieved from [70].

Table 2. FMRI results. Clusters showing significantly more hemodynamic activity to the unpredictable than the expected critical words within the left-lateralized search region of interest.

	^No.	Voxel p-value	z-score	MNI (x, y, z)	*Size	+Cluster p-value
A. *Low Constraint Unexpected > Expected*						
Inferior frontal gyrus (pars triangularis)	5A	<0.0001	5.17	-48, 24, 12	1054	<0.0001
Inferior frontal gyrus (pars orbitalis)		<0.0001	5.16	-38, 28, -10		
Middle cingulate cortex	6A	<0.0001	4.16	2, 24, 38	241	<0.005
Supplementary motor area		<0.0001	4.02	-2, 18, 46		
Anterior cingulate cortex		<0.0001	3.78	-6, 32, 28		
B. *High Constraint Unexpected > Expected*						
Fusiform cortex (temporal)	2	<0.0001	4.49	-40, -24, -18	189	<0.02
Middle temporal cortex (anterior)	1C	<0.0001	3.77	-56, -16, -8		
Inferior temporal cortex	1B	<0.0001	3.47	-48, -18, -20		
Inferior parietal lobule (angular gyrus)	4A	<0.0001	4.81	-42, -64, 26	430	<0.0005
Inferior parietal lobule (other)		<0.0005	3.58	-32, -82, 44		
Inferior frontal gyrus (pars orbitalis)	5A	<0.0001	6.41	-36, 24, -8	2328	<0.0001
Inferior frontal gyrus (pars triangularis)		<0.0001	5.36	-48, 22, 18		
Inferior frontal gyrus (pars opercularis)		<0.0001	4.04	-38, 8, 34		
Middle frontal cortex	5B	<0.0001	4.65	-20, 20, 46	855	<0.0001
Superior frontal cortex (medial)	5C	<0.0001	4.55	-6, 38, 40		
Superior frontal cortex (lateral)		<0.0001	3.89	-12, 40, 50		
Supplementary motor area		<0.0005	3.67	-2, 20, 52		
C. *Anomalous > Expected*						
Fusiform	2	<0.0001	4.56	-44, -48, -20	168	<0.02
Inferior frontal gyrus (pars triangularis)	5A	<0.0001	5.86	-46, 28, 8	1902	<0.0001
Inferior frontal gyrus (pars orbitalis)		<0.0001	5.83	-36, 28, -8		
Inferior frontal gyrus (pars opercularis)		<0.0001	4.33	-40, 8, 22		

We only report regions that reached a cluster-level significance threshold after family-wise error (FWE) correction of $p < 0.05$, small volume corrected (SVC) over the search region [69].
Anatomical locations and Montreal Neurological Institute (MNI) template coordinates correspond to the p-values and z-scores of representative peaks within each cluster. We used the automated anatomical labeling (AAL) atlas to define the anatomical regions reported. Only one peak per anatomical region is reported.
^No.: The numbering and names of each region correspond to those shown in Figures 5 and 6. Supplementary Table 1 lists the correspondence between the names of the regions indicated here and the names of the regions from the AAL atlas.
*Size of cluster: the number of contiguous voxels within each cluster.
+Cluster p-value: the cluster-level significance after FWE correction of $p < 0.05$, SVC over the search region.

References

1 Griffiths, T. L., Kemp, C. & Tenenbaum, J. B. in *The Cambridge Handbook of Computational Psychology* (ed R. Sun) 59-100 (Cambridge University Press, 2008).

2 Mumford, D. On the computational architecture of the neocortex. II. The role of cortico-cortical loops. *Biological Cybernetics* **66**, 241-251, doi:10.1007/BF00198477 (1992).

3 Rao, R. P. N. & Ballard, D. H. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nat Neurosci* **2**, 79-87, doi:10.1038/4580 (1999).

4 Friston, K. J. A theory of cortical responses. *Philos Trans R Soc Lond B Biol Sci* **360**, 815-836, doi:10.1098/Rstb.2005.1622 (2005).

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

6 de Lange, F. P., Heilbron, M. & Kok, P. How do expectations shape perception? *Trends Cogn Sci* **22**, 764-779, doi:10.1016/j.tics.2018.06.002 (2018).

7 Todorovic, A. & de Lange, F. P. Repetition suppression and expectation suppression are dissociable in time in early auditory evoked fields. *J Neurosci* **32**, 13389-13395, doi:10.1523/JNEUROSCI.2227-12.2012 (2012).

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

9 Price, C. J. & Devlin, J. T. The interactive account of ventral occipitotemporal contributions to reading. *Trends Cogn Sci* **15**, 246-253, doi:10.1016/J.Tics.2011.04.001 (2011).

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

11 Friston, K. J. Hierarchical models in the brain. *PLoS Comput Biol* **4**, e1000211, doi:10.1371/journal.pcbi.1000211 (2008).

12 Baldi, P. & Itti, L. Of bits and wows: A Bayesian theory of surprise with applications to attention. *Neural Netw* **23**, 649-666, doi:10.1016/j.neunet.2009.12.007 (2010).

13 Lee, T. S. & Mumford, D. Hierarchical Bayesian inference in the visual cortex. *Journal of the Optical Society of America A* **20**, 1434, doi:10.1364/josaa.20.001434 (2003).

14 Radvansky, G. A. & Zacks, J. M. Event perception. *Wiley Interdiscip Rev Cogn Sci* **2**, 608-620, doi:10.1002/wcs.133 (2011).

15 Kuperberg, G. R. Tea with milk? A Hierarchical Generative Framework of Sequential Event Comprehension. *Topics in Cognitive Science*, doi:10.1111/tops.12518 (2020).

16 DeLong, K. A., Troyer, M. & Kutas, M. Pre-processing in sentence comprehension: sensitivity to likely upcoming meaning and structure. *Lang Linguist Compass* **8**, 631-645, doi:10.1111/lnc3.12093 (2014).

17 Kuperberg, G. R. & Jaeger, T. F. What do we mean by prediction in language comprehension? *Lang Cogn Neurosci* **31**, 32-59, doi:10.1080/23273798.2015.1102299 (2016).

18 Franklin, N. T., Norman, K. A., Ranganath, C., Zacks, J. M. & Gershman, S. J. Structured event memory: a neuro-symbolic model of event cognition. *Psychol Rev* **127**, 327-361, doi:10.1037/rev0000177 (2020).

19 Gershman, S. J., Monfils, M. H., Norman, K. A. & Niv, Y. The computational nature of memory modification. *Elife* **6**, doi:10.7554/eLife.23763 (2017).
20 Kutas, M. & Hillyard, S. A. Brain potentials during reading reflect word expectancy and semantic association. *Nature* **307**, 161-163, doi:10.1038/307161a0 (1984).
21 Federmeier, K. D., Wlotko, E. W., De Ochoa-Dewald, E. & Kutas, M. Multiple effects of sentential constraint on word processing. *Brain Research* **1146**, 75-84, doi:10.1016/j.brainres.2006.06.101 (2007).
22 Kuperberg, G. R., Brothers, T. & Wlotko, E. A tale of two positivities and the N400: Distinct neural signatures are evoked by confirmed and violated predictions at different levels of representation. *J Cogn Neurosci* **32**, 12-35, doi:10.1162/jocn_a_01465 (2020).
23 Van Petten, C. & Luka, B. J. Prediction during language comprehension: benefits, costs, and ERP components. *Int J Psychophysiol* **83**, 176-190, doi:10.1016/j.ijpsycho.2011.09.015 (2012).
24 Hagoort, P. & Indefrey, P. The neurobiology of language beyond single words. *Annu Rev Neurosci* **37**, 347-362, doi:10.1146/annurev-neuro-071013-013847 (2014).
25 van de Meerendonk, N., Kolk, H. H. J., Chwilla, D. J. & Vissers, C. T. W. M. Monitoring in language perception. *Lang Linguist Compass* **3**, 1211-1224, doi:10.1111/j.1749-818X.2009.00163.x (2009).
26 Kuperberg, G. R. Neural mechanisms of language comprehension: Challenges to syntax. *Brain Research* **1146**, 23-49, doi:10.1016/j.brainres.2006.12.063 (2007).
27 McClelland, J. L., McNaughton, B. L. & O'Reilly, R. C. 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, doi:10.1037/0033-295X.102.3.419 (1995).
28 Lambon-Ralph, M. A., Jefferies, E., Patterson, K. & Rogers, T. T. The neural and computational bases of semantic cognition. *Nat Rev Neurosci* **18**, 42-55, doi:10.1038/nrn.2016.150 (2017).
29 Wydell, T. N., Vuorinen, T., Helenius, P. & Salmelin, R. Neural correlates of letter-string length and lexicality during reading in a regular orthography. *J Cogn Neurosci* **15**, 1052-1062, doi:10.1162/089892903770007434 (2003).
30 Lau, E. F., Phillips, C. & Poeppel, D. A cortical network for semantics: (De)constructing the N400. *Nature Rev Neurosci* **9**, 920-933, doi:10.1038/nrn2532 (2008).
31 Woolnough, O. *et al.* Spatiotemporal dynamics of orthographic and lexical processing in the ventral visual pathway. *Nature Human Behaviour*, doi:10.1038/s41562-020-00982-w (2020).
32 Halgren, E. *et al.* N400-like magnetoencephalography responses modulated by semantic context, word frequency, and lexical class in sentences. *NeuroImage* **17**, 1101-1116, doi:10.1006/nimg.2002.1268 (2002).
33 McCarthy, G., Nobre, A. C., Bentin, S. & Spencer, D. D. Language-related field potentials in the anterior-medial temporal lobe: I. Intracranial distribution and neural generators. *J Neurosci* **15**, 1080-1089 (1995).
34 Carpenter, G. A. & Grossberg, S. Normal and amnesic learning, recognition and memory by a neural model of cortico-hippocampal interactions. *Trends in Neurosciences* **16**, 131-137 (1993).

35 Wang, L., Kuperberg, G. & Jensen, O. Specific lexico-semantic predictions are associated with unique spatial and temporal patterns of neural activity. *Elife* **7**, doi:10.7554/eLife.39061 (2018).
36 Zacks, J. M., Tversky, B. & Iyer, G. Perceiving, remembering, and communicating structure in events. *Journal of Experimental Psychology: General* **130**, 29-58, doi:10.1037/0096-3445.130.1.29 (2001).
37 Altmann, G. T. & Mirkovic, J. Incrementality and prediction in human sentence processing. *Cogn Sci* **33**, 583-609, doi:10.1111/j.1551-6709.2009.01022.x (2009).
38 McRae, K., Brown, K. S. & Elman, J. L. Prediction-based learning and processing of event knowledge. *Topics in Cognitive Science* **(this issue)** (2020).
39 Brothers, T., Wlotko, E. W., Warnke, L. & Kuperberg, G. R. Going the extra mile: Effects of discourse context on two late positivities during language comprehension. *Neurobiology of Language* **1**, 135-160, doi:10.1162/nol_a_00006 (2020).
40 Thompson-Schill, S. L., D'Esposito, M., Aguirre, G. K. & Farah, M. J. Role of left inferior prefrontal cortex in retrieval of semantic knowledge: a reevaluation. *Proceedings of the National Academy of Sciences* **94**, 14792-14797 (1997).
41 Alexander, W. H. & Brown, J. W. The role of the anterior cingulate cortex in prediction error and signaling surprise. *Top Cogn Sci* **11**, 119-135, doi:10.1111/tops.12307 (2019).
42 Kumaran, D. & Maguire, E. A. Match mismatch processes underlie human hippocampal responses to associative novelty. *J Neurosci* **27**, 8517-8524, doi:10.1523/JNEUROSCI.1677-07.2007 (2007).
43 Gray, J. A. The contents of consciousness: A neuropsychological conjecture. *Behavioral and Brain Sciences* **18**, 659-676, doi:10.1017/s0140525x00040395 (1995).
44 Lisman, J. E. & Grace, A. A. The hippocampal-VTA loop: controlling the entry of information into long-term memory. *Neuron* **46**, 703-713, doi:10.1016/j.neuron.2005.05.002 (2005).
45 Coulson, S., King, J. W. & Kutas, M. Expect the unexpected: Event-related brain responses to morphosyntactic violations. *Lang Cogn Process* **13**, 21-58, doi:10.1080/016909698386582 (1998).
46 Sanders, H., Wilson, M. A. & Gershman, S. J. Hippocampal remapping as hidden state inference. *Elife* **9**, doi:10.7554/eLife.51140 (2020).
47 Geukes, S. *et al.* A large N400 but no BOLD effect–comparing source activations of semantic priming in simultaneous EEG-fMRI. *PloS One* **8**, e84029 (2013).
48 Lau, E. F., Weber, K., Gramfort, A., Hämäläinen, M. S. & Kuperberg, G. R. Spatiotemporal signatures of lexico-semantic prediction. *Cerebral Cortex* **26**, 1377-1387, doi:10.1093/cercor/bhu219 (2016).
49 Salmelin, R. & Kujala, J. Neural representation of language: activation versus long-range connectivity. *Trends Cogn Sci* **10**, 519-525, doi:10.1016/j.tics.2006.09.007 (2006).
50 Vartiainen, J., Liljestrom, M., Koskinen, M., Renvall, H. & Salmelin, R. Functional magnetic resonance imaging blood oxygenation level-dependent signal and magnetoencephalography evoked responses yield different neural functionality in reading. *J Neurosci* **31**, 1048-1058, doi:10.1523/JNEUROSCI.3113-10.2011 (2011).
51 Furey, M. L. *et al.* Dissociation of face-selective cortical responses by attention. *Proc Natl Acad Sci U S A* **103**, 1065-1070, doi:10.1073/pnas.0510124103 (2006).

52 Liljestrom, M., Hulten, A., Parkkonen, L. & Salmelin, R. Comparing MEG and fMRI views to naming actions and objects. *Human Brain Mapping* **30**, 1845-1856, doi:10.1002/hbm.20785 (2009).
53 Ahlfors, S. P. *et al.* Cancellation of EEG and MEG signals generated by extended and distributed sources. *Human Brain Mapping* **31**, 140-149 (2010).
54 Peirce, J. W. PsychoPy--Psychophysics software in Python. *Journal of Neuroscience Methods* **162**, 8-13, doi:10.1016/j.jneumeth.2006.11.017 (2007).
55 Sanford, A. J., Leuthold, H., Bohan, J. & Sanford, A. J. S. Anomalies at the borderline of awareness: an ERP study. *J Cogn Neurosci* **23**, 514-523 (2011).
56 Oostenveld, R., Fries, P., Maris, E. & Schoffelen, J.-M. FieldTrip: Open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Comput Intell Neurosci* **2011**, 1, doi:10.1155/2011/156869 (2011).
57 Matlab 2014b (The MathWorks, Inc., Natick, Massachusetts, United States, 2014).
58 Perrin, F., Pernier, J., Bertrand, O. & Echallier, J. F. Spherical splines for scalp potential and current density mapping. *Electroencephalography and Clinical Neurophysiology* **72**, 184-187 (1989).
59 Gramfort, A. *et al.* MNE software for processing MEG and EEG data. *NeuroImage* **86**, 446-460, doi:http://dx.doi.org/10.1016/j.neuroimage.2013.10.027 (2014).
60 Hämäläinen, M. S. & Sarvas, J. Realistic conductivity geometry model of the human head for interpretation of neuromagnetic data. *IEEE Transactions on Biomedical Engineering* **36**, 165-171, doi:10.1109/10.16463 (1989).
61 Dale, A. M. *et al.* Dynamic statistical parametric mapping: combining fMRI and MEG for high-resolution imaging of cortical activity. *Neuron* **26**, 55-67 (2000).
62 Fischl, B., Sereno, M. I. & Dale, A. M. Cortical surface-based analysis. II: Inflation, flattening, and a surface-based coordinate system. *NeuroImage* **9**, 195-207, doi:10.1006/nimg.1998.0396 (1999).
63 Artifact Detection Toolbox (ART) (Gabrieli Laboratory, MIT, 2011).
64 Desikan, R. S. *et al.* An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *NeuroImage* **31**, 968-980, doi:10.1016/j.neuroimage.2006.01.021 (2006).
65 Maris, E. & Oostenveld, R. Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods* **164**, 177-190, doi:10.1016/j.jneumeth.2007.03.024 (2007).
66 Hämäläinen, M. S., Hari, R., Ilmoniemi, R. J., Knuutila, J. E. T. & Lounasmaa, O. V. Magnetoencephalography—theory, instrumentation, and applications to noninvasive studies of the working human brain. *Reviews of Modern Physics* **65**, 413-497, doi:10.1103/RevModPhys.65.413 (1993).
67 Khan, S. *et al.* Maturation trajectories of cortical resting-state networks depend on the mediating frequency band. *Neuroimage* **174**, 57-68, doi:**10.1016/j.neuroimage.2018.02.018** (2018).
68 Tzourio-Mazoyer, N. *et al.* Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *NeuroImage* **15**, 273-289, doi:10.1006/nimg.2001.0978 (2002).
69 Worsley, K. J. *et al.* A unified statistical approach for determining significant signals in images of cerebral activation. *Human Brain Mapping* **4**, 58-73, doi:10.1002/(sici)1097-0193(1996)4:1<58::aid-hbm4>3.0.co;2-o (1996).

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