Semantics vs. World Knowledge in Prefrontal Cortex

Liina Pylkkänen, Bridget Oliveri, and Andrew J. Smart

Department of Linguistics and Department of Psychology, New York University

KEYWORDS: semantic processing, word knowledge, MEG, Anterior Midline Field, ventromedial prefrontal cortex, left inferior prefrontal cortex

Please address correspondence to
Liina Pylkkänen
Department of Psychology
New York University
6 Washington Place, Room 870
New York, NY 10003
Tel: (212) 998-8386
liina.pylkkanen@nyu.edu

Abstract

Humans have knowledge about the properties of their native language at various levels of representation; sound, structure and meaning computation constitute the core components of any linguistic theory. Although the brain sciences have engaged with representational theories of sound and syntactic structure, the study of the neural bases of sentence-level semantic computation has so far focussed on manipulations that mainly vary knowledge about the world, and not necessarily linguistic knowledge about meaning, as defined by formal semantics. In this MEG study, we vary both semantic and world knowledge in the same experiment, and show that semantic violations, but not world knowledge violations, elicit an effect in ventromedial prefrontal cortex (vmPFC), while both types of violations engage the left inferior prefrontal cortex. In our previous work, we have shown that the vmPFC is also sensitive to various types of "coercions," i.e., operations that repair semantic type-mismatch. Together, these results suggest that the vmPFC is involved in the composition of complex meaning, but not in the evaluation of whether an expression fits one's knowledge of the world.

INTRODUCTION

Imagine you are sitting in a windowless office. Your colleague walks in and says: "Please close your window." The sentence is obviously English – you have no trouble understanding it – but since the world around you does not meet the presupposition of the request, i.e., that you have a window, the request comes across as inappropriate. Compare this to the following utterance: "Please estimate your students." Now you are no longer even sure if your colleague is speaking English. Although you can imagine making sense of similar requests, such as "Please estimate your students' grades," this particular sentence is somehow ill-formed. Why? Because it violates a semantic constraint having to do with the possibilities of combining a verb such as *estimate*, which most naturally takes a full sentence as its direct object (e.g., *I estimated what my students knew*), with just a noun phrase. Specifically, such a combination only works if the noun describes a relation, such as *price*, *value*, or *weight* (Caponigro & Heller, 2007, Nathan, 2006) Students are individuals, not relations, and therefore they cannot occur as the object of *estimate*. Although your knowledge that students are individuals is knowledge about the world, your knowledge that *estimate* only accepts relational nouns as its object is knowledge about the semantics of the English language.

This contrast between the contextually inappropriate and the semantically ill-formed expressions illustrates a fundamental distinction in language. On the one hand, the grammar of our language constrains the range of possible expressions, including constraints such as the one on *estimate*. However, humans never produce the majority of possible expressions of their language. This is because we generally use language to communicate about the world, and only a small subset of possible expressions make sense given the way the world actually is. Thus in the

local context of the windowless office, the utterance "please close your window" is unlikely to occur. Many other possible expressions are unlikely ever to occur, simply because they do not describe situations that fit any easily imaginable state of affairs. For example, you'll probably never hear the sentence "my pet cloud ate a rock," since clouds, as we know them, are not possible pets and rocks in our world are not edible. But the sentence is nevertheless grammatical English.

The past twenty years of cognitive neuroscience has seen hundreds of studies on the brain responses elicited by sentences that violate world knowledge. This body of research stemmed from Kutas and Hillyard's original discovery that sentences such as *he spread the warm bread with socks* elicit an increased N400 amplitude (Kutas & Hillyard, 1980), which is today perhaps the most widely replicated finding in the cognitive neuroscience of language. In the context of linguistic theory, i.e., the formal study of linguistic representations, this sentence is a typical world knowledge violation: it follows all the semantic rules of English, and the reason it sounds odd is simply that socks do not have the right chemical make-up to function as a spread. However, in cognitive neuroscience, violations of this sort are generally called semantic violations, revealing a terminological difference between cognitive neuroscience and linguistics. Consequently, phenomena that are considered "semantic" in linguistic theory have gone almost completely unstudied. This means that we know next to nothing about the neural bases of one of the core components of grammar. In this article, we will follow the terminology of linguistics: "semantics" refers to the composition operations that serve to construct the meaning of an expression and "world knowledge" to our non-linguistic knowledge about the world that, for example, determines whether an utterance describes a plausible situation or not. In this type theory, a semantic violation is a situation where composition rules such as functional application

or predicate modification (for a review, see, Pylkkänen & McElree, 2006) are unable to apply because the expressions does not provide the appropriate input for the rules. A world knowledge violation (without a semantic violation) is a situation where semantic composition succeeds but the resulting meaning describes an unlikely or impossible event in the world.

In this work we used magnetoencephalography (MEG) to test whether semantic violations would elicit neural responses distinct from world knowledge violations, as most theories of linguistic representation would predict. Specifically, we were interested in two regions: the ventromedial prefrontal cortex (vmPFC) and the left inferior prefrontal cortex (LIPC), or "Broca's area." Focusing on the vmPFC was motivated by recent MEG findings that sentences that are semantically well-formed but hard to compose elicit increased amplitudes in the Anterior Midline Field (AMF), an MEG response component whose neural generator localizes in the vmPFC (Pylkkänen & McElree, 2007; Brennan & Pylkkänen, 2008, Pylkkänen et al., 2009). This establishes the AMF component as a candidate neural correlate of semantic composition in the linguistic sense. We predicted that if the AMF is indeed related to the process of composing complex meanings from simpler ones, and if encountering a semantic violation involves increased composition effort, then semantic violations should elicit increased AMF amplitudes. Alternatively, if the AMF is related specifically to *successful* semantic composition, then semantic violations might elicit less AMF activity than well-formed controls. Most importantly though, we were interested in whether the AMF generator, i.e., the vmPFC, would show functional specificity for semantic processing, i.e., an effect of semantic violations, but not of world knowledge violations.

The LIPC was treated as a second region of interest due to results by Hagoort et al. (2004), who identified the LIPC as sensitive to two different types of world knowledge violations.

Specifically, the LIPC showed increased activation both for expressions that described impossible situations as well as for sentences that were plausible but false. The expressions describing impossible situations were statements such as *Dutch trains are sour*, where a taste-describing adjective is predicated of an inedible object. Following the cognitive neuroscience tradition, these expressions were labelled semantic violations. However, like the Kutas & Hillyard stimuli discussed above, in terms of linguistic theory, the sentence *Dutch trains are sour* is, in fact, semantically well-formed – in the compositional system both *train* and *sour* describe properties of individuals and thus they can combine without a problem to form the complex property of being both a train and sour (Heim & Kratzer, 1998). However, once we have performed such a composition, our world knowledge dictates that the complex property is a highly strange one, given that sour is a taste and trains are vehicles and not foods. Thus, although it would not be impossible to build the mismatch between trains and sourness into their linguistic representations (cf., Pustejovsky, 1995), most linguistic theories would treat this incompatibility as world knowledge. Expressions of the *Dutch trains are sour* type were contrasted with sentences that described plausible but non-existent states of affairs, such as *Dutch trains are white*, which is false given that trains in Holland are yellow. When compared with well-formed controls, both types of violations elicited an N400 effect in ERPs, and increased LIPC activation in fMRI. The authors concluded that semantic and world knowledge are integrated simultaneously and by the same neural structures. But, as reviewed above, the *Dutch trains are sour* type violations do not necessarily tap onto semantic composition in the linguistic sense. Our aim was to establish whether the LIPC would be sensitive to semantic violations that clearly depend on semantic knowledge about language and not the world.

To construct a direct contrast between semantic and world knowledge violations, we took advantage of the semantic constraints of verbal un-prefixation in English (Andrews, 1986; Bowerman; 1982; Clark, Carpenter, & Deutsch, 1995; Funk, 1988; Horn, 2002; Li, 1992; Kemmerer & Wright, 2002; Marchand, 1969; Sawada, 1995). This choice was inspired by a previous aphasia study on the same phenomenon (Kemmerer & Wright, 2002), which showed that a deficit in verbal un-prefixation can dissociate from more general conceptual problems. Crucially, the English un-prefix is ambiguous between a verbal and an adjectival use, and it is the verbal use that exbihits the special semantic constraints. When the prefix *un-* attaches to adjectives, it yields a rather straightforwardly negative meaning: *unhealthy*, *unhappy* and *unethical* all negate the meanings of their adjectival stems. But when *un-* attaches to verbs, its semantic impact is quite different. *Bill unbuttoned his shirt* does not mean that Bill did not button his shirt. Rather, the meaning is reversative, roughly paraphrasable as 'Bill undid the result of buttoning his shirt' (Horn, 2002; Marchand, 1969).

Reversative un-prefixation is semantically constrained in several ways. For example, *un-* generally requires its stem to describe a so-called "accomplishment" (Dowty, 1979), i.e., an event that has a complex structure, consisting of a process that leads up to a change of state. Thus although it is possible to uncurl one's hair in the morning, it is not possible to "unleave for work," even though the reverse of leaving is easy to imagine – think of forgetting your keys and having to return, for example. One reason for this is that *leave* describes a change of state without a lead-up process, and thus does not have the requisite aspectual structure for *un-*. Verbal *un-* exhibits additional even subtler restrictions. Specifically, un-verbs are most natural with stems that describe actions which put something "into a more marked or specialised state" (Covington, 1981, p. 34). The derived un-verb then signifies return to more "normal"

circumstances, or entropy (Horn, 2002). It is easy to see how this applies to a large class of frequent un-verbs. Folding creates a special configuration, unfolding releases it; similarly for *braid – unbraid*, *button – unbutton*, *buckle – unbuckle*, and so forth. Sometimes the properties of the direct object may also be relevant for the notion of entropy. For example, although uncrossing one's arms is perfectly natural, uncrossing the street sounds strange (Kemmerer & Wright, 2002). This contrast is explained by the entropy-constraint: while crossing one's arms creates a more marked situation or configuration, crossing the street does not. Strikingly, *un-* can have a vacuous meaning if the verbal stem already describes an event that brings about entropy. For example, although *freeze* and *unfreeze* have clearly different meanings, *thaw* and *unthaw* have exactly the same meaning, given that *thaw* already describes a release-type event (of becoming unfrozen) (Horn, 2002).

Crucially, although understanding how the world works is sufficient to block one from uttering the sentence "Dutch trains are sour," world knowledge by itself is not sufficient for mastering verbal un-prefixation. To correctly use verbal *un-*, one needs to acquire the subtle semantic constraints that restrict its distribution. If it is indeed the case that such semantic constraints are qualitatively different from world knowledge constraints on language usage, then violations of the two types of knowledge should engage different neural circuits. With magnetoencephalography (MEG), we sought to assess the detailed spatial and temporal characteristics of the brain responses elicited by semantic and world knowledge violations, with the specific aim of evaluating whether the vmPFC and the LIPC are sensitive to these violations. Differences between the violation conditions and the well-formed controls were evaluated both in a hypothesis-driven region of interest (ROI) analysis focussed on the vmPFC and the LIPC as well as in a global whole brain analysis.

METHODS

Participants

Fifteen right-handed native English speakers (4 male) participated in the study. All were graduate or undergraduate students at New York University (ages 18 - 34) and were paid for their participation.

Materials

Subjects were presented with three types of sentences: semantic violations, world knowledge violations, and well-formed control expressions. The semantic stimuli violated the semantic constraints on verbal un-prefixation, whereas the world knowledge violations combined well-formed unprefixed verbs with implausible objects. In order to have the semantic and world knowledge violations occur in the same syntactic position in each sentence, all stimuli were passives and the target item was the unprefixed verbal participle. To force a verbal reading of the participle, the sentences occurred in the progressive, which blocks the adjectival reading of passive participles (Dowty, 1979). The un-prefixed participles served as the target items in the MEG data analysis.

(1) a. Well-formed: … the wine was being uncorked…

b. Semantic violation: … the wine was being unchilled…

c. World violation: … the thirst was being uncorked…

(2) a. Well-formed: … the toilet was being unclogged…

b. Semantic violation: … the toilet was being unflushed…

c. World violation: … the towel was being unclogged…

(3) a. Well-formed: … the lightbulb was being unscrewed…

b. Semantic violation: … the lightbulb was being unswitched…

c. World violation: … the shadow was being unscrewed…

To increase the naturalness of the stimuli, sentence fragments such as those illustrated in (1-3) were further embedded into larger structures, as shown in Table 1. Altogether, the critical stimuli consisted of 23 triplets of the form shown in Table 1. Crucially, all the impossible un-prefixations in the semantic violations were well-formed adjectival participles (e.g., *unchilled wine* for 1b, *unflushed toilet* for 2b, and *unswitched lightbulb* for 3b), so it was not the case that the semantic violations involved nonwords and the other two conditions existent words. Rather, the ill-formed un-prefixations were, in fact, matched with the well-formed un-prefixations in length ($t(22) = 0.89$, $p = 0.38$), lexical frequency of the verbal stem ($t(22) = -1.18$, $p = 0.25$; HAL corpus) as well as surface frequency ($t(22) = 1.62$, $p = 0.12$; HAL corpus). All lexical statistics are summarized in Table 2.

To assure that any effects of the two violations would not be explainable in terms of semantic priming, the un-prefixed participles were matched in semantic relatedness to the passivized subjects across the three conditions. As a measure of semantic relatedness, we used Latent Semantic Analysis (LSA), which showed that the co-occurance of the subjects and the unprefixed participles did not differ across conditions ($F(2, 22) = 1.28$, $p = 0.29$). Co-occurance

with the subject was also estimated with respect to the verbal stem of the unprefixed form, and this co-occurance was matched across conditions as well ($F(2, 22) = 0.92$, $p = 0.56$).

The critical subject-copula-participle phrases (underlined in Table 1) were also normed for well-formedness by 45 New York University undergraduates. Subjects rated each sentence on a scale from 1 to 7, where 7 was completely well-formed and 1 completely ill-formed. There was a highly reliable main effect of condition on the ratings ($F(2, 22) = 48.62$, $p = 0.0001$), but crucially, this was driven solely by the higher ratings for the well-formed stimuli (M = 5.3, SD = 2.09), while the semantic violations (M = 2.8, SD = 1.9) and the world knowledge violations (M = 2.4, SD = 1.83) were judged as equally ill-formed ($p = 0.36$). The LSA co-occurrence measures and the well-formedness judgment data are summarized in Table 2.

As is common in MEG studies of sentence processing, the design was within subjects, since this allows for a larger number of trials and better signal to noise ratio for source analysis. Various measures were taken to reduce the sense of repetition for our subjects. First, the critical materials were embedded in a large number fillers, as described below. Second, we introduced some carefully constructed lexical and syntactic variation into the non-critical regions of the triplets, to further decrease the repetitiveness of the stimuli. In the matrix clause preceding the critical region – the critical region is underlined in Table 1 – we varied whether the subject was definite, quantificational, or introduced by a cleft structure (*it was the judge who saw that …*). Crucially, this variation never altered the lexical semantic field of the matrix clause and thus was unlikely to affect its semantic relatedness to our un-prefixations. Finally, the lexical material of the last two words in each triplet was also varied, but the two words following the critical unprefixed participle were always kept constant.

It should be mentioned that although the semantic violations were ill-formed, as evidence by the well-formedness norming data, it is possible that subjects might in some cases be able to coerce the verbal stems into describing events that do accept the un-prefix. But crucially, English does not have any productive rule for this type of coercion, which is why these stimuli are treated here as cases of unresolvable mismatch. Our hypotheses required us to only focus on trials where the interpretation of the semantic violations did fail, as intended. To diagnose subjects' intuitions about the well-formedness of each stimulus, we had them perform off-line sensicality judgments at the end of each sentence during the MEG measurement.

Filler materials were created to ensure that the well-formedness of the target items could not be predicted on the basis of the progressive auxiliary (*being*) or the un-prefix. 46 filler sentences were structurally identical to the critical items except they did not occur in the progressive and thus they always allowed a well-formed adjectival reading (e.g., *the wine was unchilled*). Another 46 fillers were structurally identical to the critical items except that they involved a non-prefixed participle and were also well-formed (e.g., *the wine was being corked*). These materials were further mixed with 69 sentences that were structurally similar to the test items and which also balanced the amount of repetition of the well-formed and ill-formed verbal un-prefixations. Altogether, each subject saw both the ill-formed and the well-formed verbal un-prefixations three times. The order of presentation was pseudorandom in such a way that the effect of repetition was counterbalanced across conditions. Finally, the materials included 145 filler sentences which were structurally different from the critical stimuli but approximately of similar length. Altogether, the subjects read 375 sentences, 46% of the materials were ill-formed/anomalous. As already mentioned, each sentence was followed by a yes/no sensicality judgment.

Recording Procedures

During the experiment, subjects lay in a dimly-lit magnetically-shielded room. The stimuli were projected onto a screen at a distance of 17 cm from the subject, in white non-proportional Courier font (font size = 90) against a black background. Trials began with a fixation cross in the center of the screen, at which the sentence presentation was initiated with the subjects' button press. Each sentence appeared on the screen one word at a time, each word displayed for 300 ms followed by an intervening 300 ms of black screen. At the end of the sentence, the question "Make sense?" appeared on the screen, prompting the subject to indicate their sensicality judgment with their left middle or left index finger. After completing each quarter of the experiment, the subjects were allowed to take a break before beginning the next trial; at this instruction screen they were also reminded to try to avoid blinking during trials.

Neuromagnetic fields were collected using a 275-channel whole head gradiometer (CTF, Vancouver, Canada) sampling at 600Hz in a band between 0.1Hz and 200Hz. Recording was locked to a 1300 ms interval surrounding the critical target verb: 300ms prior to its onset, and 1000ms afterwards. The complete session lasted about an hour and fifteen minutes.

For four subjects, T1-weighted magnetic resonance (MR) full-brain anatomical images had been acquired in the context of another experiment using a 3-T Siemens Allegra (Siemens Medical, Malvern, PA). In order to examine the localization of our ventromedial effect (see Results) on these MRIs, each of the four subject's cortical surface was reconstructed using BrainVoyager QX software (Brain Innovation B.V. Maastricht, The Netherlands). To allow for precise co-registration of MEG and MRI data, vitamin E tablets were placed on the preauricular and nasion points (the locations of the MEG fiducial coils) prior to the MRI scans. MEG sensor

and MRI coordinate systems were co-registered using BESA and BrainVoyager. The digitized head surface points from each subject were aligned with the MRI head surface using a warping algorithm. Once the head surface points were aligned with the head surface, the reconstructed cortical surfaces were imported into BESA. For each subject, MNEs, were computed using approximately 2500-4500 locations on the individual brain surface defined by the gray-white-matter boundary. The distributed source solutions were then overlaid onto the reconstructed cortical surfaces.

MEG data analysis

The MEG responses to the un-prefixed participles were analyzed using distributed source analysis. MEG data were first cleaned of artifacts and averaged according to stimulus category. Data from any trials with either incorrect responses or misses (defined as response times longer than 4 seconds) were excluded from the averages. Trials with excessively long reaction times were excluded in case the subject in those situations might have performed the sensicality judgment by recalling the sentence after the question mark, as opposed to incrementally processing stimulus. The artifact threshold varied between 2500 and 400fT, depending on the general amplitude range of the subject. Overall, 22.7% of the trials were excluded due to behavioral errors or artifacts. Prior to source analysis, the data were high-pass at 1Hz and low-pass filtered at 40Hz.

MEG data were analyzed as distributed sources using L2 Minimum Norm Estimates (MNEs), calculated in BESA 5.1. Each MNE was based on the activity of 1426 regional sources evenly distributed in two shells 10% and 30% below a smoothed standard brain surface.

Regional sources in MEG can be regarded as sources with two single dipoles at the same location but with orthogonal orientations. The total activity of each regional source was computed as the root mean square of the source activities of its two components. The minimum norm images were depth weighted as well as spatio-temporally weighted, using a signal subspace correlation measure introduced by Mosher & Leahy (1998).

Activity in the vmPFC and the LIPC was assessed both in an ROI and a whole brain analysis. The time-window for the ROI-analysis was established by examining the time-course of grandaveraged minimum norm activity (across subjects and conditions) in the ventromedial and left inferior ROIs. Both peaked between 200 and 300ms and thus the ROI analysis was performed for the 200-350ms interval. Fig. 1 plots the ventromedial and left inferior regions that were used in the ROI analysis. Activity within these regions was averaged sample by sample and significant differences between the violation conditions and the well-formed controls were assessed with the cluster-based nonparametric permutation test (Maris & Oostenveld, 2007). For each ROI, sample by sample t-statistics were first computed for both comparisons (i.e., semantic vs. well-formed, and world knowledge vs. well-formed). Samples that that showed a difference at the .05 level were then grouped into clusters on the basis of temporal adjacency. This resulted in three vmPFC clusters (246- 263ms, 291- 310, 330-342ms) and two LIPC clusters (268-273ms, 301-330ms) for the semantics vs. well-formed comparison and three LIPC clusters for the world knowledge vs. well-formed comparison (205-208ms, 298-312ms, 316-335ms). Cluster-level t-statistic was derived by summing the absolute values of the t-statistics within each cluster. Finally, the clusters were evaluated for significance using a permutation test (10000 permutations).

For robustness, we also used a whole brain analysis to test for vmPFC and LIPC effects. We compared the MNEs of the activity elicited by the experimental conditions sample by sample in two pairwise analyses, one between the semantic violations and the well-formed stimuli and the other between the world knowledge violations and the wellformed stimuli. A difference was considered significant if it remained reliable ($p < 0.05$) for at least 10 samples (15ms) and was observed in at least 10 adjacent sources.

RESULTS

Behavioral data

At the end of each sentence, participants were presented with a question mark prompting them to judge whether the sentence made sense or not. For inclusion in the MEG data analysis, we required above-chance performance on all conditions. Eleven out of the fifteen subjects who participated in the experiment passed this criterion, suggesting that some subjects either had trouble with our stimuli or did not pay sufficient attention. For the eleven participants, accuracy on the sensicality judgment task averaged at 77% (SD = 1.3%) for the well-formed controls, 74% (SD = 1.4%) for the semantic violations, and 88% (SD = 1%) for the world knowledge violations. Although the world knowledge violations had a higher rate of accuracy than the semantic violations, these two types of violations did not differ in ill-formedness in a separate judgment experiment employing a 1-7 scale instead of a binary judgment (see Methods: Materials). Subjects could use as much time as they wished to perform the sensicality judgements. The resulting mean reaction times were long: 7017ms (SD = 3104ms) for the well-

formed stimuli, 6601ms (SD = 1406ms) for the semantic violations and 5707ms (SD = 1739ms) for the world knowledge violations. The long reaction times are likely to reflect the rather high representational complexity of our stimuli. For inclusion in the MEG data analysis, we required that the subject respond correctly within four seconds, as described above.

VmPFC and LIPC ROI-analysis

Our ROI analysis focused on the vmPFC and the LIPC. As explained above, significant differences between the violation conditions and the well-formed controls were assessed with a cluster-based nonparametric permutation test (Maris & Oostenveld, 2007). As shown in Fig. 1., this procedure identified three significant clusters (Monte Carlo $p < 0.05$, corrected). First, the vmPFC showed a significant increase of activation for semantic violations as compared to controls at 290-310ms. Second, the semantic violations also elicited increased activity in the LIPC at 300-330ms. Finally, amplitudes in the LIPC were also increased for world knowledge violations in the same time-window as the LIPC effect of semantic violations. Thus the ROI analysis yielded evidence for a semantics specific effect in the vmPFC, as predicted by the hypothesis that this region is involved in semantic composition, but not necessarily in the evaluation of the real-world plausibility of an expression. The LIPC results replicate Hagoort et al.'s (2004) fMRI findings for world knowledge violations and further show that violations of language internal semantic constraints also affect the LIPC. In other words, the LIPC is not a region that differentiates between semantics and world knowledge.

Whole brain analysis

Figure 2 plots all effects that passed the significance criterion for pair-wise comparisons of activation elicited by the two violations as compared to the well-formed controls. As the visualization of ventral prefrontal activity reveals, a clear effect of semantic violations is observed at 225-300ms and at 325-350ms, with no corresponding effects for world knowledge violations, consistent with the ROI-analysis. Results pertaining to the LIPC were similarly consistent with the ROI-analysis: as shown in the bottom panel of Fig. 2, both semantic and world knowledge violations elicited increased amplitudes in the LIPC at 300-350ms.

Additional effects in the whole brain analysis included a left occipital effect of semantic violations at 275-325ms and inferior temporal effects bilaterally at 225-250ms for semantic violations and right-laterally at 250-300ms for world knowledge violations. Thus it is possible that these structures also participate in the processing of the two types of violations; however, as they did not pertain to our hypotheses regarding the vmPFC and the LIPC, we refrain from speculating about them further.

Finally, we assessed whether the vmPFC effect of semantics would also be observed in a direct contrast with the world knowledge violations. As shown in Fig 3., such an effect is observed, but the localization is more anterior than the one observed in the comparison with the well-formed controls. In the raw minimum norms, the ventromedial activity elicited by the semantic violations was rather broadly distributed along the medial line of ventral orbitofrontal cortex. Thus it is unclear what, if any, conclusions follow from this slight difference in localization; most likely the center of the effect is simply somewhat different depending on

whether the semantic violations are compared to the well-formed controls or the world knowledge violations.

The vmPFC effect of semantics on individual MRIs

MRIs were available for four subjects. To further gauge the localization of the semantics-specific vmPFC effect, the cortices of these four brains were segmented and the Minimum Norm Estimates of these four subjects were recalculated, using approx. 2500-4500 locations on the individual brain surface defined by the gray-white-matter boundary. Fig. 4 shows for each individual, a ventral view of the current densities associated with the experimental conditions at the time of the significant vmPFC cluster (300-330ms). The general finding of increased ventromedial activity for the semantic violations was clearly observable in each individual's data. Further, for each individual, the activity was clearly medial, centered on the gyrus rectus. There was, however, no generalization regarding laterality, one subject showing an effect bilaterally (S2), two subjects left-laterally (S1, S4), and one right-laterally (S3). Thus there may be substantial individual differences.in the lateralization of the ventromedial effect, although it must also be kept in mind that the minimum norm localization method has limited ability to differentiate the hemispheres for this type of medial activity.

DISCUSSION

In this research we investigated an aspect of linguistic representation that has received little attention in cognitive neuroscience, namely semantic constraints that are not reducible to world

knowledge. In particular, we sought to assess whether the vmPFC would show specific sensitivity to language-internal semantic knowledge, given that previous studies have identified activity in this region as a potential neural correlate of semantic composition (Pylkkänen & McElree, 2007; Brennan & Pylkkänen, 2008; Pylkkänen et al., 2008). MEG responses to violations of the semantic constraints on verbal un-prefixation indeed showed reliably enhanced frontal activity localizing to the vmPFC. Violations of world knowledge, on the other hand, did not affect activity in this region. Thus our results show that the brain distinguishes between linguistic semantics and world knowledge, and that the vmPFC participates in the composition of complex meanings from the elementary building blocks of language.

Semantic violations also elicited increased amplitudes in the LIPC, but rather than being uniquely semantic, this effect was shared with world knowledge violations. This result is exactly parallel to Hagoort et al.'s (2004) data, although, as explained above, their semantic violations would be considered world knowledge violations in most representational theories. Together, these findings suggest that the LIPC is sensitive to violations of various types, which fits the general observation that this region, or "Broca's area," participates in a broad range of language related processes, including lexical-semantic processing, aspects of syntax, and speech production (Grodzinsky & Amunts, 2006). Thus it is quite possible that the LIPC effects elicited by the semantic and world knowledge violations are functionally distinct. Our LIPC findings show some consistency with the aphasia data of Kemmerer and Wright (2002), who found patients with LIPC damage to show both specific deficits in verbal-unprefixation as well as more general lexical semantic problems. However, given that the aim of our study was to identify neural regions sensitive to semantics but not to world knowledge, our discussion will focus on the vmPFC effect elicited by the semantic violations.

The vmPFC and language processing

The finding that semantic violations, but not world knowledge violations, affect vmPFC activity is interesting both for models of language processing as well as for our understanding of vmPFC function more generally. For language processing, the critical question is whether the vmPFC participates in comprehension in a central way, or whether its function is limited to environments where there is some type of semantic mismatch. So far studies in our lab have employed either resolvable or, as in the current study, unresolvable mismatch. The motivation for this has been to vary semantic composition while keeping the syntax constant. For resolvable mismatch, we have demonstrated that the AMF response, localizing to the vmPFC, shows increased amplitudes both for expressions such as *begin the book,* where a verb that semantically selects for events (as in *begin writing a book*) combines with an entity-denoting object (Pylkkänen & McElree, 2007; Pylkkänen et al., 2009), as well as for expressions such as *the clown jumped for ten minutes*, where a verb describing a punctual event combines with a durative adverb, yielding a repetitive reading of the verb (Brennan & Pylkkänen, 2008). Both of these types of expressions involve operations, often called "coercions," that repair an initial mismatch between the meaning of a predicate and its argument or modifier (for reviews, see Pylkkänen & McElree, 2006; Pylkkänen, 2008). The current study shows that in addition to resolvable mismatch, the vmPFC shows increased amplitudes for semantic mismatch that cannot be repaired and thus results in ill-formedness. This suggest that the vmPFC effect is not related to successful composition of a well-formed representation, but rather to attempts at composition, which may or may not ultimately succeed. Interestingly, the vmPFC effect of semantic violations occurred about 100ms earlier than our previously obtained coercion effects (Pylkkänen & McElree, 2007; Brennan &

Pylkkänen, 2008), suggesting that the mismatch between the verbal *un*- and its stem in the violation condition can be detected very rapidly, perhaps because the prefix *un*-, as a closed class morpheme, is accessed very quickly and imposes very strong constraints on its stem. The observed latency variation of vmPFC effects of semantic variables will be an important question for future studies.

The above set of MEG results on the vmPFC can be explained both by the hypothesis that the vmPFC performs basic semantic composition, as well as the hypothesis that vmPFC houses mechanisms that are specific to mismatch resolution. While we cannot definitively distinguish between these two options with the present data, the former, more general hypothesis does lend itself better to explaining several extant finding about language processing and the vmPFC. For example, in an auditory fMRI study, Maguire and colleagues found that ventromedial orbitofrontal activity increased as the comprehension of a story increased (Maguire, Frith, & Morris, 1999). Similar results have been obtained for the visual modality in an MEG study on cortico-cortical connectivity (Kujala et al., 2007). In this study, the ventromedial orbitofrontal cortex was identified as one of nine strongly connected network nodes during story reading. The synchrony between the orbital region and two other network nodes (the left inferior occipitotemporal cortex and the left superior temporal cortex) was stronger for connected stories than for isolated words. Further, this synchrony was modulated by the presentation rate of the story, such that faster presentation led to stronger synchrony. The authors interpreted this as reflecting increased processing demands for comprehending the story under faster presentation conditions. Medial orbitofrontal activity has also been reported to increase when subjects are asked to complete sentences with a word that fits the context as opposed to completing a sentence with a word that does not fit (Nathaniel-James & Frith, 2002). Clearly, the former task

is more parallel to natural language processing than the latter, and thus this result is straightforwardly explained by the hypothesis that the vmPFC participates in the basic construction of complex semantic representations. Finally, the involvement of the vmPFC in comprehension is also supported by recent studies on referential ambiguity, which has been shown to elicit more vmPFC activity that referentially failing expressions (Nieuwland, Petersson & Van Berkum, 2007). In ERPs, similar manipulations have elicited increased anterior negativities, qualitatively different from the N400 response associated with semantic anomalies (Van Berkum et al., 2003, 2007; Nieuwland & Van Berkum, 2008). Thus it is possible that these anterior negativities may be related to the MEG AMF response.

The vmPFC and higher cognition in general

The results described above suggest that the the vmPFC contributes to language processing in a way that is both modality-independent as well as shared between production and comprehension. However, rather than language, the vmPFC has been much more prominently associated with various types of non-linguistic higher cognition, such emotion (Damasio, 1994; Bechara, Damasio, & Damasio, 2000), decision making (Bechara, Tranel, & Damasio, 2000; Fellows & Farah, 2007; see Wallis, 2007 for a recent review), representation of reward value (Schoenbaum, Roesch, & Stalnaker, 2006), and social cognition, including theory-of-mind (Amodio & Frith, 2006; Baron-Cohen et al., 1994; Baron-Cohen & Ring, 1994; Gallagher & Frith, 2003; Krueger, Barbey, & Grafman, 2009; Rowe, Bullock, Polkey, & Morris, 2001). It is thus unsurprising that ventromedial effects of linguistic manipulations are usually thought to relate to one of these non-linguistic functions. For example, successful comprehension and production have both been

speculated to be rewarding experiences, thus engaging the vmPFC (Maguire et al., 1999; Nathaniel-James & Frith, 2002). However, this hypothesis would not explain our current results, as we observe larger vmPFC activation for those sentences for which a well-formed semantic representation cannot be built. Kujala et al. (2006), on the other hand, related their orbitofrontal findings to visual recognition, but this explanation is clearly inapplicable to Maguire et al's auditory findings (Maguire et al., 1999). Finally, given that sentence processing tasks usually involve some type of judgments on the stimuli, vmPFC effects of linguistic manipulations might plausibly relate to decision-making. However, Pylkkänen et al. (2009) ruled out this interpretation in a recent study where resolvable semantic mismatch elicited increased vmPFC activation even in the absence of any decision task.

Of the various functions that have been associated with the vmPFC, social cognition relates to language processing perhaps the most naturally. In some ways, language comprehension is a type of theory-of-mind task: on the basis of sensory stimulation, the interlocutor must reconstruct the mental model of the message that the speaker had in mind. Thus it is possible that some aspects of language processing share mechanisms with social cognition, or are evolutionarily evolved from computations involved in social cognition. Future research directly contrasting semantic manipulations and nonlinguistic social tasks is obviously needed to elucidate this possibility. One domain where semantic processing and social cognition strongly intersect is the interpretation of sarcasm, which requires overriding the literal interpretation of an expression on the basis of recognizing that the speaker's intention is somehow opposed to it. Thus if semantic interpretation and theory-of-mind engage some of the same mechanisms, one might expect sarcasm to be a robust recruiter of ventromedial prefrontal activity. Deficit-lesion studies have indeed shown that ventromedial damage leads to profound deficits in comprehending sarcasm

(Shamay-Tsoory et al., 2005). One the basis of this, one might expect the vmPFC to also be sensitive to other more classically "pragmatic" phenomena, such as implicatures or presupposition failure.

Given that the vmPFC is sensitive to such a broad range of manipulations, developing theoretical models of this region, and the orbitofrontal cortex more generally, has been particularly challenging (Zald, 2007). There is a general consensus that the orbitofrontal cortex is necessary for flexible behaviour and for navigating a complex social environment (Barbas, 2007; Rempel-Clower, 2007), but a more mechanistic understanding of this region is lacking. So far language has not generally been considered a core vmPFC function, but as reviewed above, there is strong evidence that the vmPFC is sensitive to many linguistic manipulations. Given that linguistics offers extremely detailed representational theories for language, language may provide a fruitful window to vmPFC function. Our present findings show that the vmPFC is sensitive to the semantic well-formedness but not the plausibility of an expression. This suggests that the vmPFC is involved in building semantic representations, but not in assessing how likely the situation described by the expression is, given world knowledge. Thus one general hypothesis of vmPFC function might be the composition of complex representations, perhaps in multiple domains. An integrative role for the medial PFC has already been proposed in the domain of social cognition (Krueger et al., 2009), as well in reward-guided behavior, where it has been proposed that the orbitofrontal cortex, including medial PFC, is responsible for integrating different decision variables to derive an abstract value signal (Wallis, 2007). We propose that a domain general version of this type of hypothesis might have potential in relating the many reported functions of the vmPFC, which so far have largely been studied in separate lines of research.

ACKNOWLEDGEMENT

This research was supported by the National Science Foundation grant BCS-0545186.

REFERENCES

Amodio, David, and Chris Frith. 2006. Meeting of minds: the medial frontal cortex and social cognition. *Nature Reviews Neuroscience* 7, 268-277.

Andrews, E. (1986). Analysis of de- and un- in American English. *American Speech*, 61, 221-32.

Barbas, H. (2007). Specialized Elements of Orbitofrontal Cortex in Primates. *Ann N Y Acad Sci.* 1121:10-32.

Baron-Cohen, S., Ring, H., Moriarty, J., Schmitz, B., Costa, D. & Ell, P. (1994). The brain basis of theory of mind: the role of the orbito-frontal region. *Br. J. Psychiatry* 165, 640–649.

Baron-Cohen, S., & Ring, H. (1994). A model of the mindreading system: neuropsychological and neurobiological perspectives. In: P. Mitchell and C. Lewis, Editors, *Origins of an Understanding Mind*, Erlbaum (1994), pp. 183–207.

Bechara A, Damasio H, Damasio AR. (2000). Emotion, decision making and the orbitofrontal cortex. *Cerebral Cortex*, 10(3):295-307.

Bechara A, Tranel D, Damasio H.(2000). Characterization of the decision-making deficit of patients with ventromedial prefrontal cortex lesions. Brain. 2000 Nov;123 (Pt 11):2189-202.

Bowerman, M. (1982). Reorganizational processes in lexical and syntactic development. In E. Wanner & L. Gleitman, *Language acquisition: the state of the art,* 319-46). Cambridge: Cambridge University Press.

Brennan, J. & Pylkkänen, L. (2008). Processing Events: Behavioral and Neuromagnetic Correlates of Aspectual Coercion. *Brain and Language*. 106, 132-143.

Caponigro, I. & Heller, D. (2007). The Non-Concealed Nature of Free Relatives: Implications for Connectivity in Specificational Sentences. In Chris Barker and Pauline Jacobson (eds.), *Direct Compositionality*, 237–263. Oxford, UK: Oxford University Press.

Clark, E., Carpenter, K., & Deutsch, W. (1995). Reference states and reversals: undoing actions with verbs. *Journal of Child Language*, 22, 633-62.

Covington, M. (1981). Evidence for Lexicalism: A Critical Review. Bloomington: Indiana University Linguistics Club.

Damasio, A. R. (1994). *Descartes' Error: Emotion, Reason, and the Human Brain.* New York: Putnam.

Dowty, D. (1979).*Word Meaning and Montague Grammar*, Reidel, Dordrecht, Netherlands.

Fellows, L. K., Farah, M. J. (2007). The Role of Ventromedial Prefrontal Cortex in Decision Making: Judgment under Uncertainty or Judgment Per Se? *Cerebral Cortex*, 17, 2669-74.

Funk, W. P. (1988). On the semantic and morphological status of reversative verbs in English and German. *Papers and Studies in Contrastive Linguistics*, 26, 19-35.

Gallagher, H. L., Frith, C. D. (2003). Functional imaging of 'theory of mind'. *Trends Cognitive Science* 7(2):77-83.

Grodzinsky, Y., & Amunts, K. (eds.) (2006).*Broca's Region*. New York: Oxford University Press.

Hagoort, P., Hald, L., Bastiaansen, M., Petersson, K.M. (2004) Integration of word meaning and world knowledge in language comprehension. *Science*, 304, 438-41.

Heim, I. and Kratzer, A. (1998). *Semantics in Generative Grammar*. Oxford: Blackwell Publishers.

Horn, L. (2002). Uncovering the un-word: a study in lexical pragmatics. *Sophia Linguistica* 49, 1–64.

Kemmerer, D., & Wright, S. K. (2002). Selective impairment of knowledge underlying un-prefixation: Further evidence for the autonomy of grammatical semantics. *Journal of Neurolinguistics*, 15(3-5), 403-432.

Krueger F, Barbey AK, Grafman J. (2009). The medial prefrontal cortex mediates social event knowledge. *Trends in Cognitive Sciences* 13, 103-9.

Kujala, J., Pammer, K., Cornelissen, P., Roebroeck, A., Formisano, E., Salmelin, R. (2007). Phase coupling in a cerebro-cerebellar network at 8–13 Hz during reading. *Cerebral Cortex*, 17, 1476-85.

Kutas, M. & Hillyard, S.A. (1980). Reading senseless sentences: brain potentials reflect semantic incongruity. *Science, 207*, 203–205.

Li, P. (1992). Cryptotypes, meaning-form mappings, and overgeneralizations. In *Proceedings of the 24th Annual Child Research Forum*, 162-178.

Maguire, E. A., Frith, C. D., & Morris, R. G. (1999). The functional neuroanatomy of comprehension and memory: the importance of prior knowledge. *Brain*. 122 (Pt 10): 1839-50.

Marchand, H. (1969). The categories and types of present-day English word formation. 2nd ed. München: Beck.

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG and MEG data. *Journal of Neuroscience Methods,* 164, 177-190.

Mosher, J. C., & Leahy, R. M. (1998). Recursive MUSIC: A Framework for EEG and MEG Source Localization. *IEEE Transactions On Biomedical Engineering*, 45, 11, 1342-54.

Nathan, L. (2006). The Interpretation of Concealed Questions. PhD Thesis, MIT, MA.

Nathaniel-James, D. A. & Frith, C. D. (2002). The role of the dorsolateral prefrontal cortex: evidence from the effects of contextual constraint in a sentence completion task. *Neuroimage*, 16(4):1094-1102.

Nieuwland, M. S., Petersson, K. M., & Van Berkum, J. J. A. (2007). On sense and reference: examining the functional neuroanatomy of referential processing. *Neuroimage*. 1;37(3):993-1004.

Nieuwland, M. S. & Van Berkum, J. J. A. (2008). The interplay between semantic and referential aspects of anaphoric noun phrase resolution: Evidence from ERPs. *Brain and Language*, doi:10.1016/j.bandl.2008.05.001.

Pustejovsky, J. (1995*). The Generative Lexicon*. Cambridge, MA: MIT Press.

Pylkkänen, L. (2008). Mismatching Meanings in Brain and Behavior. *Language and Linguistics Compass*. 2/4, 712–738.

Pylkkänen, L., Martin, A. E., McElree, B., & Smart, A. (2009). The Anterior Midline Field: Coercion or Decision Making? *Brain and Language,* 108, 184-90.

Pylkkänen, L. & McElree, B. (2006). The syntax-semantic interface: On-line composition of sentence meaning. In M. Traxler & M.A. Gernsbacher (eds.), *Handbook of Psycholinguistics* (2nd Ed), pp. 537-577. NY: Elsevier.

Pylkkänen, L. & McElree, B. (2007). An MEG Study of Silent Meaning. *Journal of Cognitive Neuroscience*, 19, 1905-1921.

Schoenbaum, G., Roesch, M.R., Stalnaker, T.A. (2006). Orbitofrontal cortex, decision-making and drug addiction. *Trends Neurosci*. 29(2):116-24.

Rempel-Clower, N. L. (2007). Role of orbitofrontal cortex connections in emotion. *Ann N Y Acad Sci.* Dec;1121:72-86.

Rowe, A.D., Bullock, P.R., Polkey, C.E., Morris, R.G. (2001). "Theory of mind" impairments and their relationship to executive functioning following frontal lobe excisions. *Brain.* 124(Pt 3):600-16.

Sawada, S. (1995). On the verb-forming prefix un-. *English Linguistics*, 12, 222-47.

Shamay-Tsoory, S.G., Tomer, R., Aharon-Peretz, J., (2005). The neuroanatomical basis of understanding sarcasm and its relationship to social cognition. *Neuropsychology*. 19(3):288-300.

Van Berkum, J. J. A., Brown, C. M., Hagoort, P. & Zwitserlood, P. (2003). Event-related brain potentials reflect discourse-referential ambiguity in spoken language comprehension, *Psychophysiology* 40, 235–248.

Van Berkum, J. J. A., Koornneef, A.W., Otten, M. & Nieuwland, M.S. (2007). Establishing reference in language comprehension: An electrophysiological perspective, *Brain Research* 1146, 158–171.

Wallis, J. D. (2007). Neuronal mechanisms in prefrontal cortex underlying adaptive choice behavior. *Ann N Y Acad Sci*. 1121:447-60.

Zald, D. (2007). Orbital versus dorsolateral prefrontal cortex: anatomical insights into content versus process differentiation models of the prefrontal cortex. *Ann N Y Acad Sci*. 1121, 395-406.

TABLE AND FIGURE CAPTIONS

Table 1: Examples of stimuli. The critical manipulation is underlined, with MEG responses recorded at the un-prefixed word.

Table 2: Lengths and frequencies (HAL corpus) of the target items with un-prefixes, as well as mean well-formed judgments and LSA co-occurrence estimates for the three conditions.

Figure 1: Results of the vmPFC and LIPC ROI-analysis. The top panel shows the regions entered into the analysis, the middle panels the time-course of activity within the regions, with significant differences indicated in black, and the bottom panel shows the mean amplitudes per condition for the significant clusters.

Figure 2: Results of the whole brain analysis of distributed source activity. Each plotted region represents a spatiotemporal neighborhood where the conditions differed reliably. Red indicates increased activation for the violation condition, and blue decreased. The vmPFC effect of the semantic violations in boxed in yellow and the LIPC effect shared by the semantic and world knowledge violations in blue.

Figure 3: A direct contrast between semantic violations and world knowledge violations at 200-350ms using a whole brain analysis. An increase in vmPFC activity is observed for the semantic violations, although somewhat more anteriorly than in the comparison between semantic violations and the well-formed controls (Fig 2), suggesting a slightly different center of the effect.

Figure 4: Ventromedial prefrontal activity per condition for four individuals for whom MRIs were available. All four showed an increase in vmPFC activity for the semantic violations.

Table 1: Examples of stimuli

Item	Condition: Stimulus
1a	Well-formed: None of the waitresses noticed that the wine was being uncorked for the wedding reception
1b	Semantic violation: The experienced waitress firmly ensured that the wine was being unchilled for the next meal.
1c	World violation: All of the waitresses knew that the thirst was being uncorked for the main course.
2a	Well-formed: The maid informed her boss that the toilet was being unclogged in the upstairs bathroom.
2b	Semantic violation: One of the maids discovered that the toilet was being unflushed in the old house.
2c	World violation: Some of the maids thought that the towel was being unclogged in the restroom downstairs.
3a	Well-formed: The diligent handyman was sure that the lightbulb was being unscrewed in the upstairs hallway.
3b	Semantic violation: The friendly handyman was told that the lightbulb was being unswitched in the dining room.
3c	World violation: Some of the handymen feared that the shadow was being unscrewed in the desk lamp.

Table 2: Lexical statistics, well-formedness judgment data and LSA similarity scores

	Well-formed (*the shirt was being unbuttoned*)	Semantic violation (*the shirt was being unironed*)	World violation (*the zipper was being unbuttoned*)
Length (un-verb)	8.8	8.6	8.8
Surface frequency (un-verb)	155	16	155
Stem frequency (un-verb)	12983	23237	12983
Well-formedness judgment	5.3	2.8	2.4
Noun–un-verb co-occurrence (LSA cosines)	0.21	0.18	0.16
Noun–verbal stem co-occurrence (LSA cosines)	0.17	0.18	0.14

Figure 1

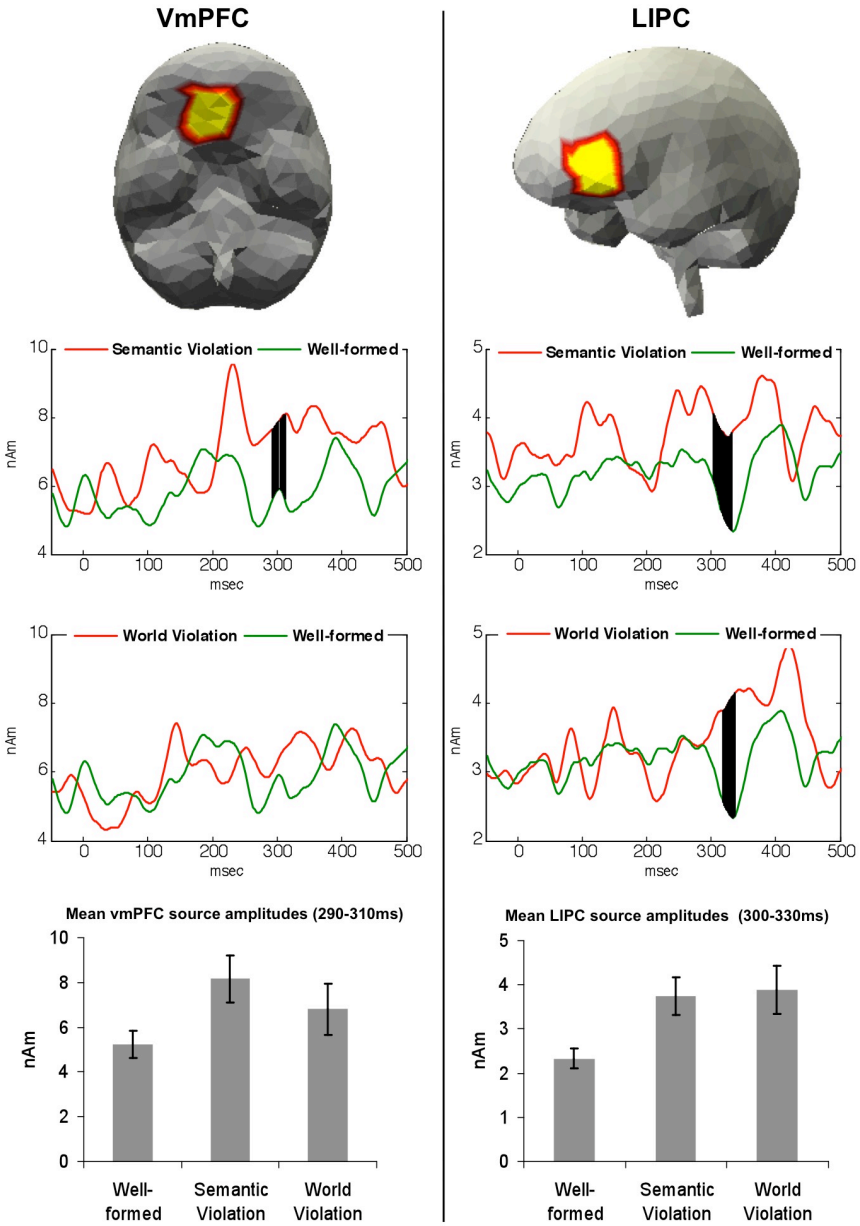

VmPFC
LIPC
Semantic Violation
Well-formed
World Violation
nAm
msec
Mean vmPFC source amplitudes (290-310ms)
Mean LIPC source amplitudes (300-330ms)
Well-formed
Semantic Violation
World Violation

Figure 2

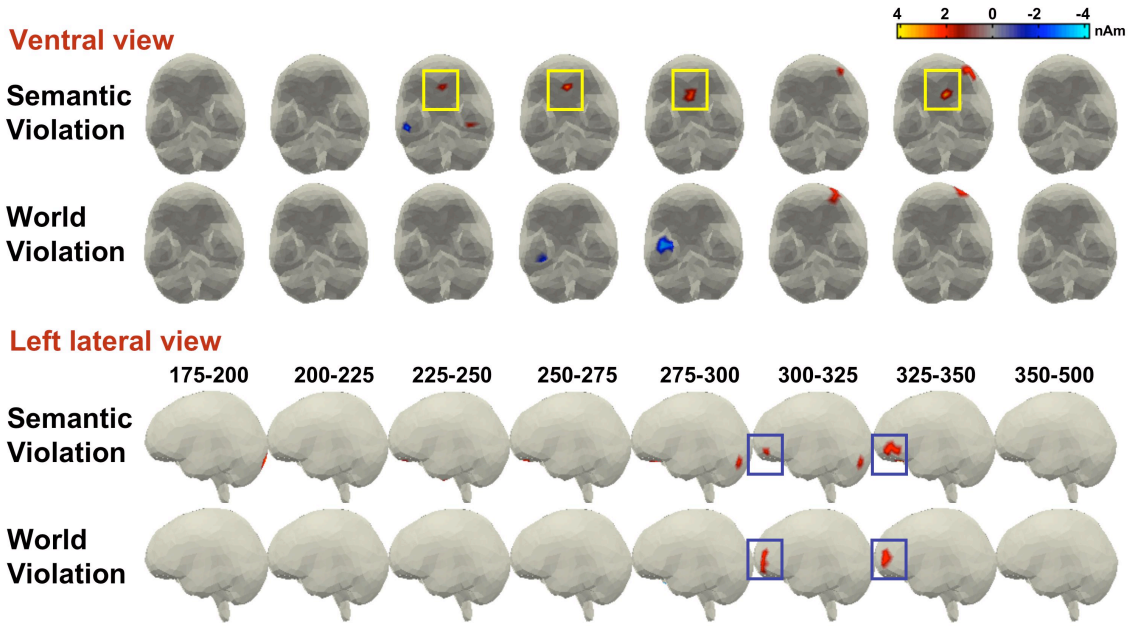

4
2
0
-2
-4
nAm
Ventral view
Semantic Violation
World Violation
Left lateral view
175-200
200-225
225-250
250-275
275-300
300-325
325-350
350-500
Semantic Violation
World Violation

Figure 3

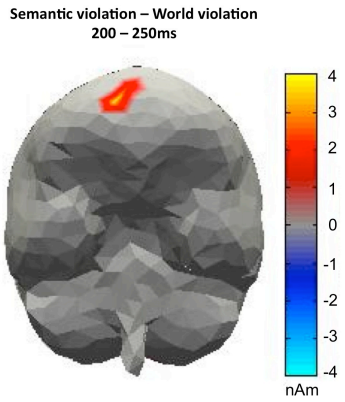
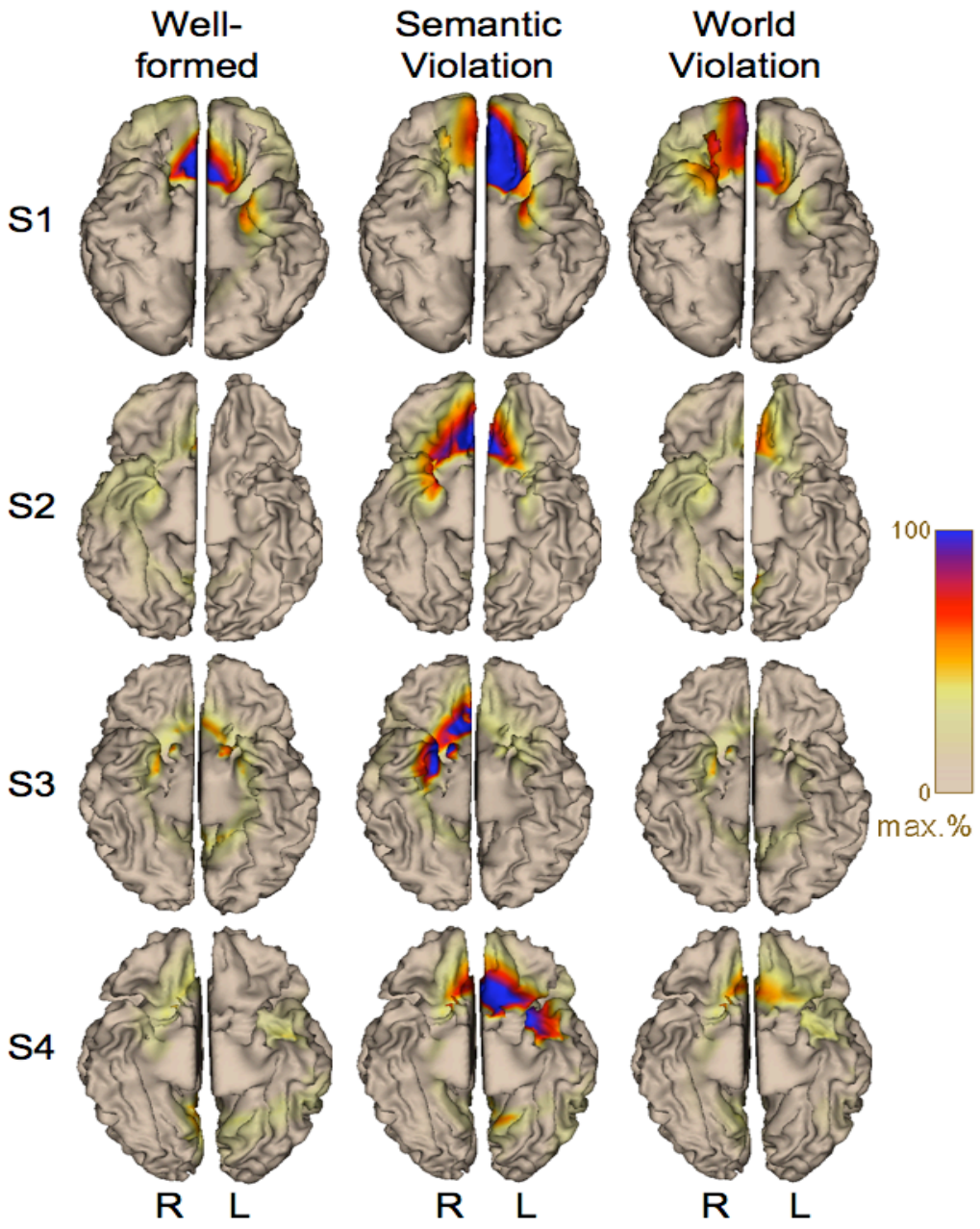

Semantic violation – World violation
200 – 250ms
4
3
2
1
0
-1
-2
-3
-4
nAm

Figure 4


Well-
formed
Semantic
Violation
World
Violation
S1
S2
S3
S4
100
0
max.%
R
L
R
L
R
L