The Semantics of Adjective Noun Phrases in the Human Brain

Alona Fyshe[a,*], Gustavo Sudre[b], Leila Wehbe[c], Nicole Rafidi[d], Tom M. Mitchell[d]

[a]*Department of Computer Science, University of Victoria, 3800 Finnerty Road, Victoria, BC, Canada, V8P 5C2*
[b]*National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, USA 20814*
[c]*Helen Wills Neuroscience Institute, University of California, Berkeley, CA, USA. 94720*
[d]*Machine Learning Department, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA, USA, 15213*

Abstract

As a person reads, the brain performs complex operations to create higher order semantic representations from individual words. While these steps are effortless for competent readers, we are only beginning to understand how the brain performs these actions. Here, we explore semantic composition using magnetoencephalography (MEG) recordings of people reading adjective-noun phrases presented one word at a time. We track the neural representation of semantic information over time, through different brain regions. Our results reveal several novel findings: 1) the neural representation of adjective semantics observed during adjective reading is reactivated after phrase reading, with remarkable consistency, 2) a neural representation of the adjective is also present during noun presentation, but this neural representation is the reverse of that observed during adjective presentation 3) the neural representation of adjective semantics are oscillatory and entrained to alpha band frequencies. We also introduce a new method for analyzing brain image time series called Time Generalized Averaging. Taken together, these results paint a picture of information flow in the brain as phrases are read and understood.

Keywords: Language comprehension, Semantic representations, Semantic composition, Magnetoencephalography, Oscillatory brain dynamics

1. Introduction

Semantic composition, the process of combining small linguistic units to build more complex meaning, is fundamental to language comprehension. It is a skill that children acquire with amazing speed, and that adults perform with little effort. Still, we are just beginning to understand the neural processes involved in semantic composition, and the neural representation of composed meaning. Multiple studies have identified cortical regions exhibiting increased activity during increased semantic composition load ([1, 2, 3, 4, 5], to name only a few). Here we considered a different question: Where and how are semantic representations stored in preparation for, and used during semantic composition?

[*] Corresponding author. Mailing Address: Department of Computer Science, University of Victoria, ECS Room 504, PO Box 1700 STN CSC, Victoria, BC, Canada, V8W 2Y2

Email address: afyshe@uvic.ca (Alona Fyshe)

Semantic composition in the brain has been studied using semantically anomalous sentences [1, 2, 6], as well as in simple phrases [4, 5], typically by comparing the *magnitude* of brain activity between conditions (e.g. composing words into phrases vs. reading word lists). Several such studies have implicated right and left anterior temporal lobes (RATL and LATL) as well as ventro-medial prefrontal cortex (vmPFC) and left inferior frontal gyrus (IFG) in compositional processing [4, 7, 3]. Magnetoencephalography (MEG) studies have shown effects in these areas as early as 180ms post stimulus onset, until around 480 ms post stimulus onset [4, 7], which aligns well with the N400 effect observed in electroencephalography [1]. Semantic composition effects have also been seen around 600ms post stimulus onset (P600)[8]. Though the P600 is more commonly associated with *syntactic* violations, it can be evoked by syntactically sound stimuli that violate a semantic constraint. In particular, semantic constraints related to the violation of animacy constraints have been shown to elicit a P600 (E.g. “Every morning at breakfast the eggs would eat toast.” is a syntactically sound sentence, but “egg” violates the semantic constraint imposed by “eat”, which requires an animate subject) [2].

The computational (rather than neuroscientific) study of language semantics has been greatly influenced by the idea that word meaning can be inferred by the context surrounding a given word, averaged over many examples of the word’s usage [9, 10, 11, 12, 13]. For example, we might see the word *ball* with verbs like *kick*, *throw*, *catch*, with adjectives like *bouncy* or with nouns like *save* and *goal*. Context cues suggest meaning, so we can use large collections of text to compile statistics about word usage (e.g. the frequency of pairs of words), which form models of word meaning. These statistics are typically compressed using dimensionality reduction algorithms like singular value decomposition (SVD), as in latent semantic analysis (LSA) [14]. LSA and similar models represent each word with a vector. Together, the vectors of many words form a Vector Space Model (VSM).

The brain’s semantic representations can be studied by quantifying the information present in neural activity at particular cortical locations and times. For example, one can train machine learning algorithms to predict which word a person is reading based on their neural activity, by training the algorithm to predict the vector associated with that word in a particular VSM [15, 16]. Such algorithms do not require large differences in brain activity between conditions, but rather leverage differences in the *spatio-temporal patterns* of neural activity, which may involve differences in signal in both the positive and negative direction in different areas of the brain at different times. Such techniques have been used in a variety of settings to predict words from brain activity [15, 16, 17, 18, 19].

To our knowledge, the study presented here represents the first effort to study semantic representations of adjective-noun phrases using the fine time resolution offered by Magnetoencephalography (MEG). To study adjective-noun phrases in the brain, we traced the flow of information through time and brain space. The ability of these algorithms to predict the stimuli from MEG recordings is indicative of the information present in the underlying brain activity, and thus is indicative of the brain’s neural representation of that stimuli. Note that the implication flows only one way, we cannot say that poor prediction accuracy at a particular time/space point is proof that semantic information is not represented there. Poor prediction accuracy could be due to multiple factors, including noise and the limitations of brain imaging technology, and is not necessarily due to a property of the brain. In addition, the neural representation of a stimulus property could actually be the representation of some correlated property. Even with these caveats, prediction accuracy is a useful measure for exploring the time-space trajectories of information processing in the brain.

The following section outlines the methods used to collect and process MEG data, and to measure predictive accuracy. Section 3 outlines the results of our experiments, and Section 4 reflects on the results to build a hypothesis of how adjective-noun phrases are read and composed in the human brain.

2. Materials and Methods

To examine the neural representations of adjective and noun semantics, we presented phrases consisting of an (adjective or determiner) and a noun. To maximize our probability of detecting the individual words of the phrase, we chose 8 nouns that have been shown to be easily predicted from MEG recordings [16]. Six adjectives were then chosen to modulate the most predictable semantic qualities of the words (e.g. edibility, manipulability, size). We also paired nouns with "the" to isolate noun meaning. In total, there are 30 word pairs (phrases). For a full list, see the Appendix. Though some of the phrases start with a determiner (the), for simplicity, throughout this paper we will refer to all phrases as adjective-noun phrases. MEG data was recorded for 9 subjects (4 female), all neurologically healthy, right-handed, native English speakers with normal or corrected to normal vision.

Phrases are shown in rapid serial visual presentation (RSVP) format (See Figure 1). During each trial, the first word of the phrase appears in text on the display at 0 seconds, and is removed from the display at 500 ms. The noun appears at 800 ms and is removed at 1300 ms. To ensure subjects were engaged during the experiment, 10% of the stimuli were adjective-adjective pairs (oddballs), for which the subjects were instructed to press a button with their left hand. Neither the adjective-adjective trials, nor the adjective-noun trial immediately following the oddball were used for analysis. Excluding these omitted trials, each phrase was presented 20 times, and analysis was performed on the mean MEG time series over all 20 trials. The experiment was carried out in 7 blocks of approximately equal length, with the opportunity to rest between blocks.

Because of the experimental setup, there is a strong correlation between the adjectives and nouns in our data. For example, the word "rotten" only appears with food words "tomato" and "carrot". For this reason, we were careful to avoid analyses in which we seek to predict a property of the adjective, but might instead predict a correlated property of the noun (and vice versa). This is explained further in Section 2.6.

2.1. Data Acquisition and Preprocessing

All 9 subjects gave their written informed consent approved by the University of Pittsburgh (protocol PRO09030355) and Carnegie Mellon (protocol HS09-343) Institutional Review Boards. MEG data were recorded using an Elekta Neuromag device (Elekta Oy). As much as possible, we adhered to the best practices of MEG data collection[20]. The

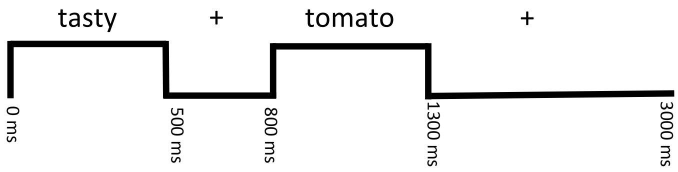


Figure 1: Stimulus presentation protocol. Adjective-noun phrases were presented one word at a time with a fixation cross both between words and between phrases. The first word of the phrase appears at 0 ms, and disappears at 500ms, the second word is visible 800 ms – 1300 ms. The first words in successive phrases are 3000 ms apart.

data were acquired at 1 kHz, high-pass filtered at 0.1 Hz and low-pass filtered at 330 Hz. Eye movements (horizontal and vertical eye movements as well as blinks) were monitored by recording the differential activity of muscles above, below, and beside the eyes. At the beginning of each session, the position of the subject's head was recorded with four head position indicator (HPI) coils placed on the subject's scalp. The HPI coils, along with three cardinal points (nasion, left and right pre-auricular), were digitized into the system to allow for head translation to register data collected in different blocks.

The data was preprocessed using the temporal extension of SSS (tSSS) [21] to remove artifacts and noise unrelated to brain activity. In addition, we used tSSS to realign the head position measured at the beginning of each block to a common location. The MEG signal was then low-pass filtered to 50 Hz to remove the contributions of line noise and down-sampled to 200 Hz. The Signal Space Projection method (SSP) [22] was used to remove signal contamination by eye blinks or movements, as well as MEG sensor malfunctions or other artifacts. The MEG data was parsed into trials, one for each phrase presentation. Each trial begins at the onset of the first word of the phrase, and ends 3000 ms later, for a total of 600 time points of data per sample (See Figure 1). MEG sensor amplitudes are known to drift with time, so we corrected each trial by subtracting from every sensor the mean signal amplitude during the 200ms before stimulus onset. During behavioral tests it was found that phrases containing the noun "thing" were inconsistently judged by human subjects, and so the 8 phrases containing the noun "thing" were omitted from further analysis, leaving a total of 30 phrases for analysis.

After processing, the MEG data for each subject consisted of 20 repetitions for each of the 30 phrases. Each repetition has a 600 dimensional time series for each of the 306 sensors. For each subject, we averaged all 20 repetitions of a given phrase to create one data instance per phrase, 30 instances in all. The dimensions of the final data matrix for each subject were $30 \times 306 \times 600$.

2.2. Source Localization

In order to transform MEG sensor recordings into estimates of neural activity localized to areas of the brain, we used a multi-step process. First, Freesurfer (`http://surfer.nmr.mgh.harvard.edu`) was used to construct a 3D model of each subject's brain, based on a structural MRI. Freesurfer was used to segment the brain into ROIs based on the Desikan-Killiany Atlas. Then the Minimum Norm Estimate method [23] was used to generate estimates of sources on the cortical sheet, spaced 5mm apart. The noise covariance matrix was estimated using approximately 2 minutes of MEG recordings collected without a subject in the room (empty room recordings) either directly before or after the subject's session. Source localization resulted in approximately 12000 sources per subject derived from the 306 MEG sensor signals.

2.3. Prediction Tasks

We used two prediction tasks to track the words representations while reading adjective-noun phrases. In each case, the task is to predict the stimulus from the MEG recording. Differences in the time course of prediction accuracy for each task will show us how the brain processes information during adjective-noun phrase comprehension. The tasks are:

1. **Predicting Adjective Semantics:** Predict the identity of the first word in the phrase (any of the 6 adjectives or the word "the")
2. **Predicting Noun Semantics:** Predict the identity of the noun (one of 8).

For both tasks, we trained models to predict the dimensions of a VSM vector for a given word. We then predict word identity based on the similarity of the predicted vectors to the corresponding true VSM vectors. A more detailed description follows in Section 2.4. For each of the words in our study, we used word vectors that are based on the sentence dependency relationships for each word of interest, averaged over a very large number of sentences [24]. We use the first $m = 500$ SVD dimensions to summarize the dependency statistics, which represents a reasonable trade-off between computation time and accuracy [24].

2.4. Prediction Framework

To study adjective-noun composition in the brain we devised a simple prediction framework. Cross-validation was performed independently for each subject, wherein 2 of the 30 phrases are withheld during training, and subsequently used to test the framework's predictions. This hold out and test procedure was repeated multiple times. For each of the prediction tasks described in Section 2.3, every stimulus phrase is represented by two vectors from a VSM, which represent the semantics of either the adjective/determiner or the noun. The elements of this vector are the targets in the prediction framework ($s^{(k)}$ in Equation 1) .

We created a data matrix $X \in \mathbb{R}^{N \times P}$ where N is the total number of phrases, and $P = s \times t$, for $s = 306$ sensors and t time points. Each element of the data matrix, $x_{i,j}$, represents the value for training instance i at a point j in MEG sensor/time space. To predict each of the dimensions of the semantic vector, we trained an L_2 regularized (ridge) regression model, $\hat{\boldsymbol{\beta}}$:

$$\begin{aligned} \hat{\boldsymbol{\beta}^{(k)}} &= \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\{ \sum_{i=1}^{N} (s_i^{(k)} - \sum_{j=1}^{P} \beta_j x_{i,j})^2 + \lambda \sum_{j=1}^{P} \beta_j^2 \right\} \\ &= \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\{ ||\boldsymbol{s}^{(k)} - X\boldsymbol{\beta}||_F^2 + \lambda ||\boldsymbol{\beta}||_F^2 \right\} \end{aligned} \qquad (1)$$

where $s_i^{(k)}$ is the kth element of the VSM word vector for training instance i, and λ is a regularization parameter that controls overfitting in $\boldsymbol{\beta}$. We append a column of 1s to X to incorporate a bias term. Notice that each semantic dimension k has its own $\hat{\boldsymbol{\beta}^{(k)}}$. Initially, λ was tuned to maximize accuracy for each semantic dimension using leave-one-out cross-validation (LOOCV) over the training data, but it was found to be very stable, so a set value of $\lambda = 10^{-6}$ was used to save computation time. Training data was normalized during cross-validation so that each time-sensor feature has mean 0 and standard deviation 1, and the same correction was applied to test data. Time windows used to train each regression model are 100ms wide and overlap by 80ms with adjacent windows.

2.5. *The 2 vs. 2 Test*

For each stimulus word w in this study, we have an m-dimensional VSM vector $\mathbf{s_w} = \{s_1 \dots s_m\}$, created from corpus data. We refer to each feature s_i in this vector as one of the semantic features of the stimulus word w. Note the semantic vector may correspond to the adjective or the noun, depending on the analysis being performed.

Using Equation 1, we trained m independent functions $f^{(1)}(\boldsymbol{x}) \rightarrow \hat{s}^{(1)}, \dots, f^{(m)}(\boldsymbol{x}) \rightarrow \hat{s}^{(m)}$, where $\hat{s}^{(i)}$ represents the predicted value of the ith semantic feature, and $f^{(k)}(\boldsymbol{x}) = (\hat{\boldsymbol{\beta}}^{(k)}\boldsymbol{x})$ using $\hat{\boldsymbol{\beta}}^{(k)}$ from Equation 1. We combined the output of $f^{(1)} \dots f^{(m)}$ to create the final predicted semantic vector $\hat{\mathbf{s}} = \{\hat{s}^{(1)} \dots \hat{s}^{(m)}\}$. We used a distance function to quantify the dissimilarity between two semantic vectors. Many distance metrics could be used, we chose cosine distance due to its popularity for VSM-related tasks in computational linguistics.

To test performance we used the forced choice **2 vs. 2 test** [15]. For each test we withheld the MEG recording for two of the 30 available adjective-noun phrases, and trained $\hat{\boldsymbol{\beta}}$ on the remaining 28. We then used the MEG data of the held out phrases ($\boldsymbol{x}$) to predict the semantic vectors for both of the held out phrases ($\hat{\mathbf{s}}$). The task is to choose the correct assignment of predicted vectors $\hat{\mathbf{s}}_\mathbf{i}$ and $\hat{\mathbf{s}}_\mathbf{j}$ to the true VSM semantic vectors $\mathbf{s_i}$ and $\mathbf{s_j}$. We will make this choice by comparing the sum of distances for the two assignments:

$$d(\mathbf{s_i}, \hat{\mathbf{s}}_\mathbf{i}) + d(\mathbf{s_j}, \hat{\mathbf{s}}_\mathbf{j}) \overset{?}{<} d(\mathbf{s_i}, \hat{\mathbf{s}}_\mathbf{j}) + d(\mathbf{s_j}, \hat{\mathbf{s}}_\mathbf{i}) \tag{2}$$

If the left hand side of the above equation is indeed smaller than the left, we mark the 2 vs. 2 test correct. **2 vs. 2 accuracy** is the percentage of correct 2 vs. 2 tests. The 2 vs. 2 test is advantageous because it allows us to use two predictions per test, resulting in a higher signal-to-noise ratio; two weak predictions can still result in a correct assignment of true to predicted phrase vectors. Under the null hypothesis that MEG data and semantics are unrelated, the expected chance accuracy for the 2 vs. 2 test is 50%.

2.6. *Correlations between the words of phrases*

There is a correlation between the adjectives and nouns in our experimental paradigm. For example, the word “rotten” is always followed by a food word. The word “carrot” never appears with the adjective “gentle”. If we were to not correct for this correlation, we could build a model that was actually detecting the correlated semantics of the noun, when we had intended to build a model that leverages only the semantics of the adjective. To avoid reporting results that rely on this confound, we only consider predicting the adjective or noun when the other word (noun or adjective, respectively) is shared in the 2 vs. 2 pair. That is, when we encounter a 2 vs. 2 pair that contrasts adjectives "rotten" and "big", we will include it in our analysis only when the noun is the same for both phrases (e.g. "rotten tomato" and "big tomato"). Thus, if our prediction framework leveraged the correlated semantics of the noun to predict the adjective, it would be of no use for differentiating between these test phrases. The same is true for the noun; when we encounter a 2 vs. 2 pair that contrasts nouns "dog" and "bear", we include it only when the adjective is the same for both phrases (e.g. "gentle dog" and "gentle bear"). Thus we can be sure that we are not relying on any correlations between adjectives and nouns for our analyses. There are (30 choose 2) $= 435$ distinct 2 vs. 2 tests. Amongst those 435 tests, 51 share the same adjective and 60 share the same noun. These 111 pairs are the only ones included in our analyses.

2.7. Significance Testing

We used permutation tests to determine the probability of obtaining our prediction results by chance. Permutation tests require shuffling the data labels (words) and running the identical prediction framework (cross validation, training $\boldsymbol{\beta}$, predicting $\hat{s}$, computing 2 vs. 2 accuracy) on the permuted data. When we do this many times, we approximate the null distribution under which the data and labels have no relationship. From this null distribution we calculate a p-value for the performance we observe when training on the correct (un-permuted) assignment of words to MEG data. In the experiments that follow, we will train and test multiple predictors on multiple time windows. To account for the multiple comparisons performed over the time windows, we used the Benjamini-Hochberg-Yekutieli (BHY) procedure, with correction for arbitrary correlation amongst the tests [25].

2.8. Time Generalization Matrices

To test the consistency of the neural code in time, we use Temporal Generalization Matrices (TGMs) [26]. TGMs mix train and test data from different time windows to test the stability of the neural representation over time. For our TGMs, we used the prediction framework described in Section 2.4, but mix train and test data selected from different time windows for each of the 2 vs. 2 tests. The entry at (i, j) of a TGM ($T_{(i,j)}$) contains the 2 vs. 2 accuracy when we train using MEG data from a time window centered at time i, and test using MEG data from a time window centered at time j. Thus, depending on the value of i and j we may be mixing train and test data from different time periods, possibly comparing times when the subject is viewing a different word type, or no visual stimuli at all. If the neural representation of a concept is stable, then models can be trained and tested with data from different time windows with little or no impact on accuracy. TGM analysis has been performed in the past, and has shown that information from previous stimuli can be detected during a memory task, even after stimulus presentation[27, 28].

Examples of hypothetical TGMs appear in Figure 2. Along the y axis is the time used to train a model, and along the x axis is the time used to test the model (generalization time), and the color corresponds to the accuracy level. The diagonal of the TGMs (i.e. $T_{i,i}$ for all i) corresponds to training and testing with the same time window (diagonal

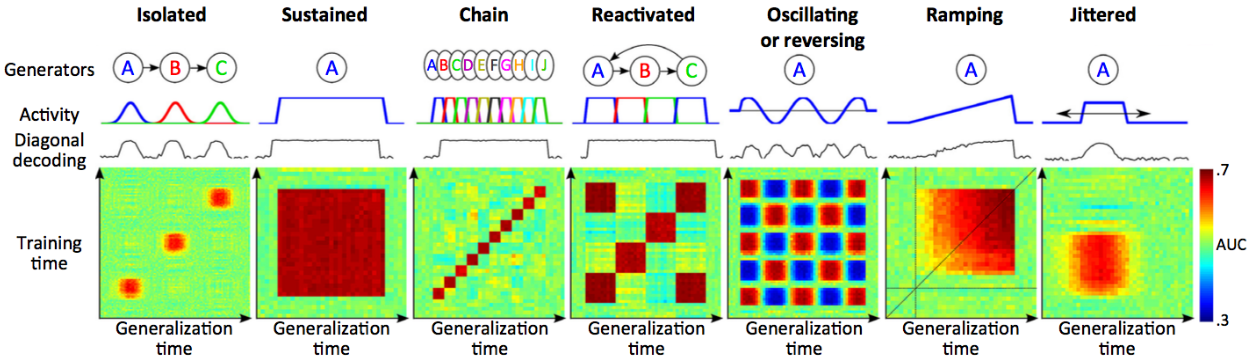


Figure 2: Hypothetical TGMs based on different neural activation patterns. Depending on the generators and resulting neural activity, the pattern of the TGMs differs. From King and Dehane [26].

decoding in Figure 2). Comparing the TGM to the diagonal decoding in Figure 2 shows that TGMs can pick up on patterns in data that might otherwise go undetected. All of the examples in Figure 2 (except ramping) assume a constant neural representation within a generator block. Figure 3A shows an example where the generator (A) doesn't produce a constant neural pattern, but rather a pattern that evolves over time, enters a negated phase (A^{-1}), and then repeats itself. Thus, instead of the block patterns we see in Figure 2, we see diagonal lines.

We extended the TGM concept to create a new type of analysis: Time Generalization Averaging (TGA). TGA is based on TGMs, but combines results from many columns of a TGM to maximize statistical power. To combine, we first align the columns of the TGM so that the coordinates corresponding to training and testing on the same time window are aligned. This requires shifting the second column of the TGM down by one row, the third column down by two, and so on. A hypothetical TGM appears in Figure 3A, and the column-shifted version of the example appears in Figure 3B. After this shifting, the strong diagonal lines in Figure 3A become horizontal lines (Figure 3B). After column-shifting the matrix, the x axis still corresponds to the test time, but now the y-axis represents the *difference* between the midpoints of the training window and testing windows. We select a subset of the columns of the column-shifted matrix, average them, and this becomes the final TGA (Figure 3C). Note that, for any given row of the column-shifted matrix, a subset of the columns may not have an entry (e.g. the top row has only one entry in column one, row two has entries in columns 1 and 2, etc.). TGA allows us to test the significance of any diagonally-aligned oscillations in accuracy, and wether the pattern for a generator is indeed negating and repeating itself. We will use a permutation test and the same TGA analysis to ensure that our results are not due to features of the data unrelated to the stimuli.

3. Results

We will begin with a typical time and space analysis of the neural representation of adjective an nouns, and then explore TGM and TGA analyses of our data.

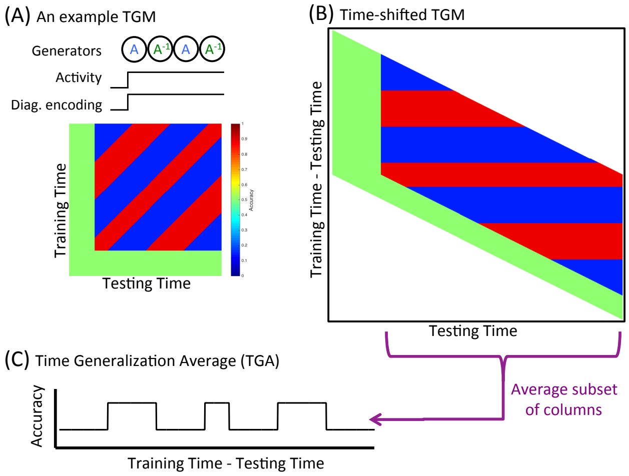


Figure 3: Time Generalized Averaging (TGA). **(A)** A hypothetical TGM that shows oscillations in accuracy aligned with the diagonal. This TGM is caused by a generator that emits a pattern that repeats, and is not constant within a time block. Note that the "Generators" and "Diagonal Decoding" are like that of "Reactivated" in Figure 2, but the TGM is very different. **(B)** We shift the columns of the TGM in (A) so that rows that corresponds to training and testing on the same time window are aligned for each column. Now the y axis corresponds to the *difference* between the train and test windows, and the x axis is still testing time. **(C)** When we average over the aligned time columns in (B), the result is a Time Generalized Average (TGA). A TGA allows us to condense the information in a TGM into a simple 1D plot.

3.1. Predicting Adjective and Noun Semantics

2 vs. 2 accuracy for predicting the adjective and noun from whole brain MEG sensor data observed within a sliding 100 millisecond interval appears in Figure 4. Here we can see that the adjective remains detectable until well after the onset of the noun, dipping below statistical significance for the first time at about 1500 ms after the onset of the adjective (700 ms after the onset of the noun). Adjective prediction accuracy again rises above the statistical significance threshold at several subsequent time points, and is significantly above chance for the last time around 2000 ms after the onset of the adjective. Thus, our model can predict the adjective during adjective *and* noun presentation, and also for a prolonged period *after* the noun stimuli has disappeared. This implies that there is a neural representation associated with the adjective that persists during the entire phrase presentation, as well as after the phrase has been presented (we call this the phrase wrap-up period). From these results we cannot tell if the neural representation of the adjective changes over time, we can only infer that there exists a reliable representation during each significantly above-chance time window.

Figure 4 also shows noun prediction accuracy as a function of time. After its initial peak around 950 ms, the accuracy for predicting noun semantics dips below the significance threshold for the first time at $\sim$ 1700 ms. Though prediction accuracy is sustained after the offset of the noun, there is no resurgence later in time. The accuracy plot for the noun is more peaked than the adjective, and it is significant for less time than the adjective. There are also two above chance points very early in time (windows centered at 770 and790 ms, corresponding to time windows 720-820 ms and 740-840 ms respectively). Though these windows do overlap with the first few milliseconds of noun presentation, they may be too early to be a true effect. We will see that these points are no longer significant when we employ the more powerful TGA analysis in Section 3.3.

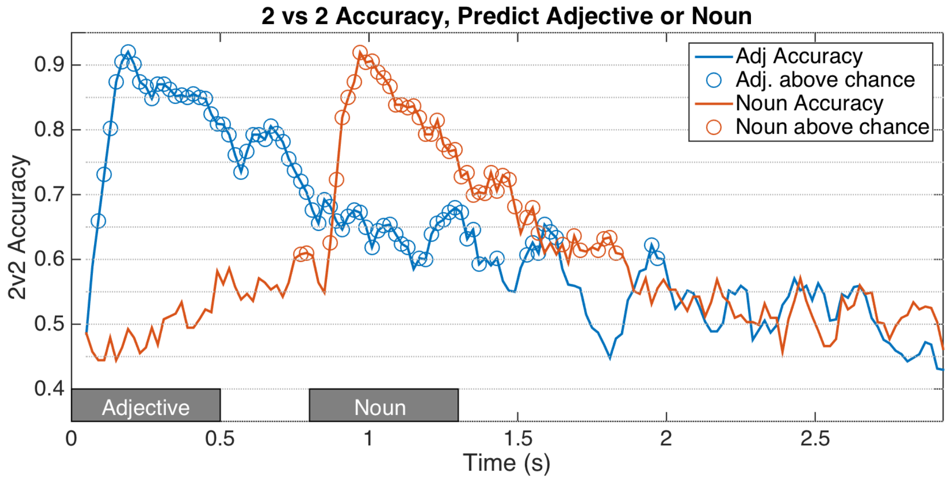


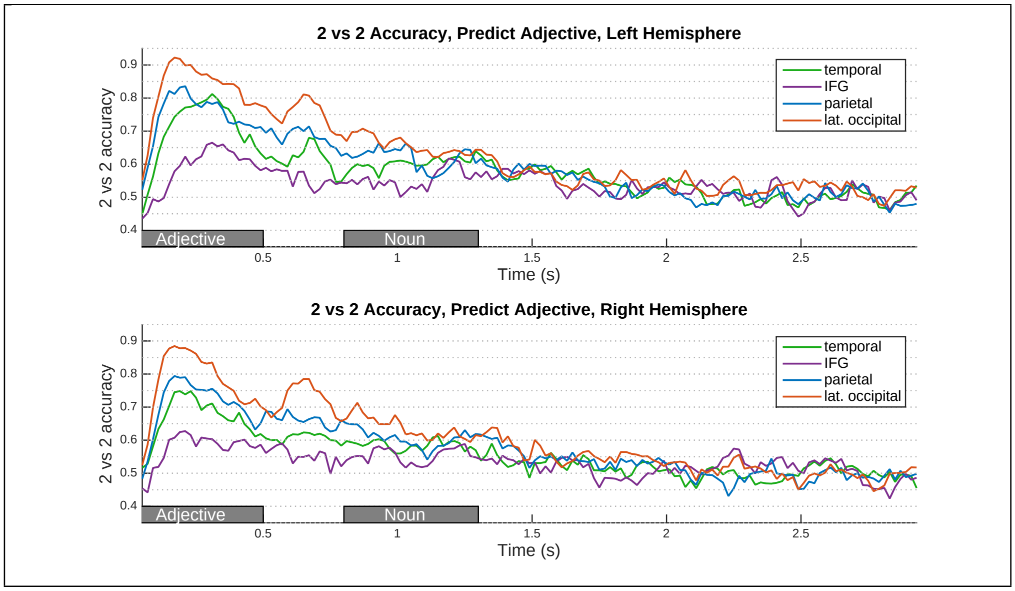
Figure 4: 2 vs. 2 accuracy as a function of time for predicting the words of the phrase, averaged over 9 subjects. Time windows for the presentation of the adjective and noun are indicated with grey rectangles. Vertical axis indicates the prediction accuracy for the adjective or noun, based on predicting its semantic feature vector from a 100 millisecond interval of MEG data in sensor space. Significantly above chance accuracies are highlighted with circular markers. Here a separate prediction model was trained for each time point on the horizontal time axis.

3.2. Localizing Adjective and Noun Semantics

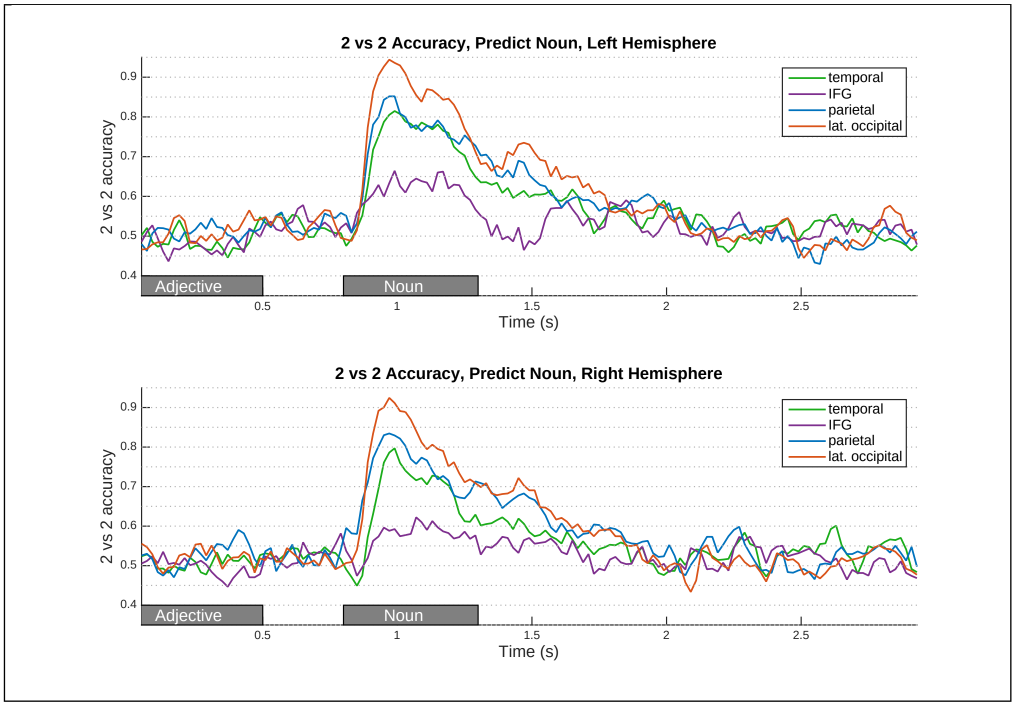
Where in the brain do the neural representations for the adjective and noun lie? To answer this question, we recreated the plot in Figure 4, but used only a subset of the source-localized ROIs to build and test our models. Here, each subject's source localized MEG data was parcellated based on the Desikan-Killiany Atlas [29]. We analyzed groups of ROIs that reside in temporal, occipital, IFG and parietal areas. For the sub-ROIs included in each of the groups, see the Appendix.

Figure 5a shows the 2 vs 2 accuracy for predicting the adjective, averaged over each ROI group in each hemisphere, as a function of time. For completeness we include right hemisphere (RH) temporal and IFG, though previous research has associated most compositional processing with these areas in the left hemisphere (LH) [4, 30, 3]. In general, the 2 vs. 2 accuracy of RH ROIs is dominated by LH ROIs, but the difference is not large. This reinforces previous studies which found that semantic representations are highly distributed [16, 31]. Accuracy in left IFG and parietal resurges about 400ms after the onset of the noun (1200ms after the onset of the adjective). In general, accuracy in occipital and parietal ROIs exceeds that of the temporal and IFG ROI groups, until about 1400ms. Based on these plots, it is not clear which area drives the resurgence of the adjective representation at 2 seconds.

Figure 5b shows the accuracy of predicting the noun, averaged over each ROI group. As with the whole brain analysis, these plots are much more peaked at the onset of the noun, and there is less sustained accuracy when compared



(a) 2 vs. 2 accuracy for predicting the semantics of the adjective.



(b) 2 vs. 2 accuracy for predicting the semantics of the noun.

Figure 5: 2 vs. 2 accuracy for predicting (a) the adjective or (b) the noun, as a function of time, averaged over several ROI groups. Within each subfigure, top: left hemisphere, bottom: right hemisphere. Stimuli timing appears in grey rectangles.

to the adjective plots. During the peak, the ordering of the accuracy for the ROIs is the same as for the adjective (highest to lowest: lateral occipital, parietal, temporal, IFG). Interestingly, though the whole brain analysis (Figure 4) showed nearly equal peak prediction accuracy for both adjective and noun, the ROI analysis shows that the noun is slightly better predicted by the individual ROIs than the adjective.

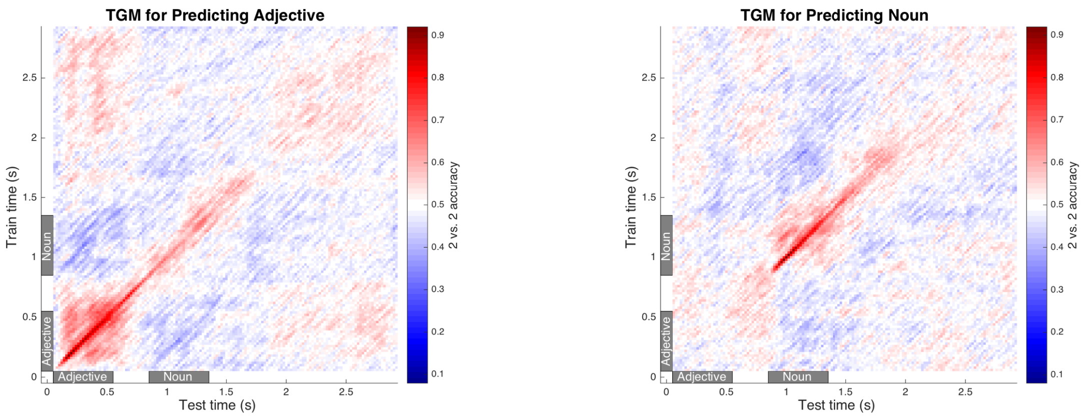
3.3. Consistency of the Neural Code in Time

How consistent in time is the neural code of the adjective? That is, does the neural code for the adjective during adjective presentation resemble the neural code used for the adjective during noun presentation, or during the phrase wrap-up period? To test this, we created Temporal Generalization Matrices (Figure 6), as described in Section 2.8, using whole brain MEG sensor data. Recall that each point along the *diagonal* of the TGMs corresponds to training and testing the prediction model using MEG data taken from the same time window, and so represents exactly the results plotted in Figure 4. In contrast, the off-diagonal elements correspond to training the model using MEG data from one time interval, but testing its prediction accuracy on data taken from a second time interval, allowing us to test whether the neural representation of this predicted information varies over time.

The TGMs in Figure 6 show two key features. Firstly, there are off-diagonal patches that correspond to both high and low prediction accuracies. Only the adjective TGM shows off-diagonal accuracy patches above 60% (Figure 6a), and they are strongest when training on a time point after 2 s, and testing on a period during the presentation of the adjective. Low accuracy (below the chance accuracy of 50%) patches appear for both the adjective and noun, and are strongest when training on a window just after the offset of that word and testing during the time when a word is visible. Secondly, the TGMs show a highly oscillatory pattern, which manifests as many diagonal lines parallel to the main diagonal. This pattern implies that the neural representation of the adjective is repeating itself periodically. We found this time constant to be very close to 100ms for most subjects, meaning that the pattern is entrained to alpha band frequencies (10 Hz).

In order to test the significance of the oscillations and above/below chance accuracy patches, we perform Time Generalization Averaging (TGA) as described in Section 2.8. Each of the columns of the TGM represents a time series of prediction accuracy for models tested on a particular time period (See Figure 7A). Recall that the first step of TGA is shifting the columns of the TGM so that the coordinates corresponding to training and testing on the same time window are aligned (See Figure 7B). We average across a subset of the aligned columns to create the time series in Figure 7C.

To study the representation of the adjective, we selected the columns corresponding to peak predictive accuracy for the adjective in Figure 4 (windows centered at 190 - 450 ms), and the rows for which 50% of columns have an entry after shifting. We average this selection of the aligned data to create our TGA graph (Figure 7C). This allowed us to test if the above and below chance regions in Figure 6 are significant. Note that, in TGAs, each point on the plot is an average over several models trained on several time windows. To account for this, the stimuli window annotations in



(a) TGM for predicting the semantics of the adjective.

(b) TGM for predicting the semantics of the noun.

Figure 6: 2 vs. 2 accuracy for predicting the words of the phrase, presented in TGMs .] Within each TGM, the color at point i, j indicates the prediction accuracy when the model is trained using data from an interval centered at time point i, then tested by its ability to predict the noun or verb based on MEG data centered at time point j. Time windows are 100ms wide and overlap by 80ms with adjacent windows. Time 0 is the onset of the adjective, 0.8 is onset of the noun, as annotated with grey rectangles.

Figure 7C represent the times during which *any* of the models' training data overlaps with the stimulus presentation period (thus the windows are wider).

Figure 7C shows several points that are significantly above/below chance. The permutation test is wonderfully flexible; we can perform a TGA analysis on permuted data and compare to the distribution of those results, knowing that our statistical test remains sound, even though our analysis has become more complex. After the onset of the noun, several of the peaks (8/10) aligned with the 10 Hz oscillations we noted in Figure 6 (for convenience we have annotated all above and below chance peaks outside of the adjective presentation). Our TGA analysis shows that the long above/below chance time periods are indeed significant. The timing of the below chance segment of the TGA analysis coincides with windows we would expect to contain N400 and P600 effects, and so could be a byproduct of the processing required for composition.

We performed a TGA analysis for predicting the noun (Figure 8), and found that there are no above chance points outside of the noun stimuli presentation window, but there are below chance points after the offset of the noun. Thus, for both words of the phrase, a TGA analysis shows that the regions of low 2 vs. 2 accuracy just after the offset of the word are statistically significant.

We performed a TGA analysis on the ROIs from Figure 5 to determine the origin of the oscillations and above/below chance patterns for the adjective's representation. TGA analysis increased our statistical power, and showed that the significantly above-chance accuracy we saw *after* 2s in Figure 7 is likely originating from the left occipital and parietal cortices. Interestingly, the significantly below chance accuracy around 800 ms appears to come from *right* parietal and temporal cortex. We performed the same TGA analysis for the noun (not shown), but the accuracy lines were very flat.

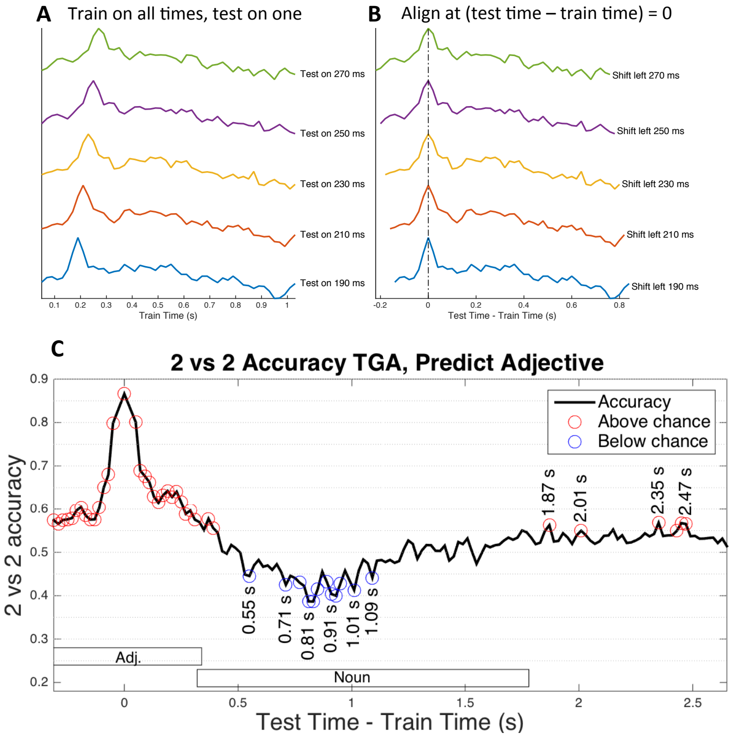


Figure 7: Time Generalized Averaging (TGA). A) Each of the columns of a Time Generalization Matrix (TGM) is a time series. B) We shift the time series from A to align the windows corresponding to training and testing on the same time. Now the x axis corresponds to the *difference* between the train and test windows. C) We average over aligned time series to create a grand average. This graph shows the TGA analysis for the semantics of the adjective, averaged over test times [0.19 - 0.45s]. White rectangles denote windows during which at least one of the trained models used data overlapping with word stimuli presentation. Our TGA analysis shows that there are points both above and below the significance threshold, even after the adjective stimuli has left the screen.

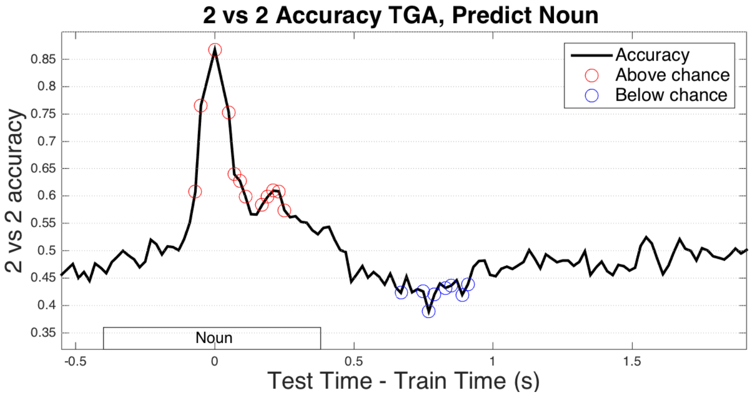


Figure 8: TGA analysis for predicting the semantics of the noun over time windows [0.95-1.17s]. The white rectangle denotes windows during which at least one of the trained models used data overlapping with word stimulus presentation. Note that there are significant points both above and below the significance threshold, and the significantly below chance points occur after the presentation of the noun stimulus.

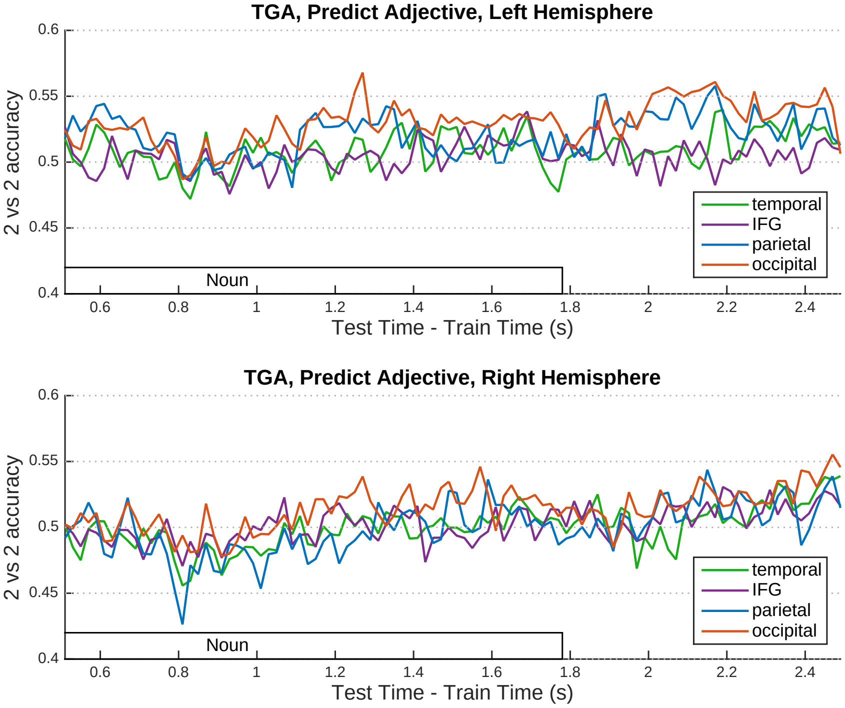


Figure 9: TGA analysis for predicting the adjective using several ROI groups. Top: left hemisphere, bottom: right hemisphere. The white rectangle denotes windows during which at least one of the trained models used data overlapping with word stimulus presentation.

3.4. *Results Summary*

To summarize, the main findings from our data analyses are: 1) we can predict the adjective well after noun stimuli presentation (as late as 2s after adjective presentation), 2) the neural representation of adjective semantics observed during adjective reading is reactivated after the entire phrase is read, with remarkable consistency, 3) there is a period of below chance prediction performance just after the presentation of each word, and 4) the neural representation of adjective semantics are oscillatory and entrained to alpha band frequencies.

4. Discussion

Here we compare and contrast the results from Section 3 to build a hypothesis of how the brain represents and processes adjective-noun phrases. To foreshadow, our interpretation of the above analyses for adjective noun processing is as follows:

1. During the time the adjective is read, the brain maintains a neural representation for the adjective;
2. During the time the noun is read, the brain holds both the representation for the noun, but also a representation of the adjective that appears to be the negation of its representation during adjective reading

3. After the noun stimuli ends, the noun enters a negated representation state, followed by a resurgence of the adjective representation in its original (non-negated) form.

Both the TGM and TGA analyses show highly oscillatory activity, which implies that semantics is represented in the brain using a repeating pattern entrained to alpha band frequencies. Our ROI analysis points to a distributed representation of composed semantics.

4.1. Adjective Semantics in Early and Late Time Windows

Figure 7C (adjective semantics TGA), shows that the pattern for the adjective is fairly consistent within the time that the adjective is being presented. In addition, there are significantly above chance points very late in time, as late as 2.45s relative distance between the training and testing windows. This result implies that the early and late representations of adjective semantics are highly similar. This late above chance accuracy could be due to the intersective quality of most of the adjectives chosen for this study; the meaning of our selected adjectives is largely unaffected by the semantics of the nouns in this experiment. For example, a rotten has very similar meaning when paired with either tomato or carrot. Though the two foods may spoil in slightly different ways, the end result is inedible. Thus, it is reasonable that, for phrases that contain intersective adjective, the neural representation of the adjective alone should be very similar the phrase's neural representation.

At first glance, one might think the late above chance prediction accuracy for the adjective conflicts with previous work showing that semantic composition begins as early as 140ms after the onset of the noun, and as late as 600ms [30, 2]. In our experiments, this would correspond to 940 ms and 1400 ms after the onset of the adjective. The early effects of semantic composition are typically studied using contrasting stimuli that either does or does not require composition. Thus, the timings reported in previous work are the "switching on" of the machinery required to perform semantic composition, but not necessarily the time when we would expect to see the final product of semantic composition, or cessation of thinking about the phrase semantics.

There is support for semantic effects as late as the effects we see here. Previous work has shown effects during joke comprehension as late as 1100 ms after the onset of the final word [32]. Semantic violation effects have been reported as late as 2.5s after the onset of the critical word in the sentence [6]. When the semantic plausibility of sentence critical words is varied, differences in EEG recordings for anomalous vs expected words extend to the edge of the 1.2s analysis window (and possibly beyond) [33]. Many analyses have restricted themselves to the time period ending 1s after the onset of the critical word, possibly because the windows of analysis in seminal papers extended only that far [1, 2]. A review of several other accounts of this late activation appears in VanPetten and Luka [34]. The results of the present study show that analyzing MEG data beyond 1s post stimulus onset can give new insight into semantic processing.

4.2. Noun Semantics

The neural representation of the noun is detectable until about 1.7s after the onset of the adjective (0.9s after the onset of the noun). This duration is shorter than that of the adjective, which is detectable in the brain even 2s after the onset of the adjective stimulus. After the noun stimulus has left the screen, a negated representation of the noun appears in the brain, again for a much shorter time than the adjective.

Noun semantics are not predictable during the phrase wrap up period (1.3-3s). It is somewhat counter-intuitive that the semantics of the adjective should be more salient than the semantics of the noun during the contemplation of the phrase. This could be the result of our choice of adjectives, which manipulate the most predictable features of the noun to their extreme ends.

4.3. The Localization of Semantic Representations

Occipital cortex is the ROI with the highest 2 vs. 2 accuracy, while temporal and IFG regions contribute less (Figure 5, Figure 9). Though temporal and IFG ROIs have been implicated in multiple previous studies of composition [4, 30], much of this previous work compared brain activation for composing vs not composing words, rather than searching for the final output of semantic composition.

One might question the results for the occipital ROI (Figure 5a), as it is a region known to handle visual perception. However, previous work has found that many semantic properties are represented in the occipital region of the brain, including properties related to the size, manipulability, animacy and threatening nature of the stimuli (see supplementary material for [16]). These semantic features are highly related to the attributes we manipulated with our adjective choices.

In general, visual stimulus features (e.g. word length) are most predictable in occipital cortex before 200ms post stimulus onset, and not predictable later in time. Note that in the TGM in Figure 6a, high off-diagonal accuracy appears for *test* times at 200ms and later, so is likely not attributable to a visual feature of the stimulus being recalled. As a sanity check, we ran our prediction framework with the task of predicting the number of letters in the adjective and found that after 700ms, we could not predict the word length of the adjective with significant accuracy. In addition, we performed a TGA analysis using time windows centered at 70 to 130 ms post adjective onset (peak time for predicting adjective word length) and found no significant points, implying that the resurgence of adjective semantics is not simply the image of the word being visually recalled, or some ghosting of the visual stimuli.

Recall, also, that we are using a corpus-derived semantic representation of the adjective for the prediction tasks throughout this paper. Though there are some correlates to the perceptual features of word strings in these corpus-derived features (e.g. frequent words are, on average, shorter than infrequent words) we are, by and large, predicting the semantic features of the words when we use these vectors.

4.4. Significantly Below Chance Prediction Accuracy in Temporal Generalization Averaging

For both words of the phrase, a period of below chance prediction accuracy appears after the offset of the word stimuli. Significantly below chance accuracy may seem counter-intuitive; how can the framework's predictions be systematically *worse* than random guessing? If the prediction is systematically inverted, perhaps the MEG signal itself is negated. To test this, we negated the MEG signal, and found that for TGM coordinates with below chance accuracy, the 2 vs. 2 accuracy on negated data was not only above chance, it was exactly $1 - a$ where a is the 2 vs. 2 accuracy on non-negated MEG data. This is a byproduct of our prediction framework. Negated MEG signal leads to negated predictions ($\hat{s}_i^{(k)}$ from Equation 1), which causes the predicted vector to point in exactly opposite of the prediction for non-negated data. The opposing direction results in a negation of the cosine of the angle, and thus flips the decision in Equation 2. This negated representation of the word could be how the brain to maintains context while processing a new word. Figure 9 provides some evidence that the right hemisphere may coordinate this negated representation, but more tests are needed to confirm this hypothesis.

What does it mean for the MEG signal to be negated? MEG measures the post-synaptic potential of many parallel dendrites. Thus, the negation of the MEG signal equates to the reversal of the underlying magnetic field, i.e. neurons with opposite direction are firing. This could be caused by several phenomena, perhaps related to the neural loops currently thought to be related to neural oscillations [35]. It is interesting that the negated representation for the adjective appears during the time we would expect to see N400 and P600 effects during noun reading (approximately 500 - 1000 ms in Figure 7C).

We hypothesize that for both the adjective and the noun, the brief negated representation is a holding pattern that allows the brain to store a word's meaning in memory while performing another action. In the case of the adjective, the next action is reading the noun. In the case of the noun, the next action is recalling the adjective for composition. The beginning of the negated noun representation aligns well with the end of the negated adjective representation. After the negated noun representation completes, we see the return of the adjective in its original, non-negated, form.

4.5. The Oscillatory Nature of Prediction Accuracy

One of the most striking patterns in the TGMs is the oscillatory nature of prediction accuracy (Figure 6). In Figure 7C we can clearly see 10 Hz (alpha) oscillations in the prediction accuracy for the adjective. Originally, alpha band oscillations were thought to be the brain's natural idling state, as they are observed in visual cortex when subjects close their eyes [36]. Note that we are not implying that the neural representation resides in the alpha band. Indeed, in studying the weight maps for adjective semantics, the oscillations are are closer to 30 Hz and are *entrained* to alpha, meaning they *repeat* every 100 ms. It has been suggested that gamma band activity may synchronize to alpha band phase in order to share information between brain areas [37]. In previous work, memory tasks have shown oscillatory representations, though at a lower frequency than reported here [27]. There is also evidence for alpha-entrained neural signals for visual tasks [38], and oscillations of above and below chance accuracy in TGMs [39].

To test if the neural representation resides in the alpha band, we bandstop filtered the MEG data to eliminate frequencies between 7 and 14 Hz. Even after this aggressive filtering, the oscillatory dynamics remain, as do the above and below chance patterns of performance. In fact, after bandstop filtering, the TGMs show clearer oscillations, but they appear to be closer to 5 Hz (theta band frequency). Power in the theta band has been shown to be related to semantics in a number of studies [40, 41, 42, 6]. Interestingly, when we highpass filtered our data to remove all frequencies below 14 Hz the TGA analysis became extremely oscillatory and showed no significant prediction accuracy (high or low) outside of the stimuli windows. Certainly, more exploration of these phenomena is required.

Is it possible that the alpha-aligned prediction oscillations are somehow related to a noise artifact? Each MEG session had 7 blocks (segments of continuous MEG recording with short breaks in between), and each phrase used for train/test is actually the mean of 20 repetitions, so the chance of any noise artifact being aligned over all 7 blocks is very unlikely. As an additional test, we also collected an additional subject's data using the same paradigm with a MEG located at Aalto University in Finland, and produced similar results.

We were concerned that the paradigm itself may have been causing the oscillation in accuracy. Because all stimulus onsets were locked to multiples of 100ms with no jitter, we thought anticipation of stimuli onset could create oscillations. To rule out this possibility, we collected an additional subject with jittered noun and adjective onsets (between 1-100ms added to onset of each word stimuli). We found the same oscillatory pattern when we corrected the train/time windows to account for the jitter. This result assures us that the oscillatory patterns are not a function of the timing of our paradigm.

We also thought that the alpha entrained pattern could be a byproduct of the alpha ringing associated with viewing a stimuli [43]. However, that even after bandstop filtering 7-14 Hz, the oscillatory dynamics remain, as do the above and below chance patterns of performance. We also wanted to ensure that our preprocessing had not introduced a ringing artifact. To test this, we reprocessed our data, eliminating all pre-processing steps except for tSSS. We then conducted the same TGA analyses and saw the performance drop by less than 1% on average, and the oscillatory patterns remained.

Together, these sanity checks provide strong evidence that the oscillatory nature of word prediction is truly a brain-related signal, and not some noise superimposed on the signal itself. But what is the purpose of this oscillatory representation? There is strong evidence that oscillations are involved in working memory [44], and working memory must play a role in language understanding. We hypothesize that, during semantic composition, oscillatory activity is used to coordinate brain areas, to retain and to recall information as it is needed. Here we explored a very controlled composition task; it will be interesting to see if the same oscillatory patterns appear in tasks requiring more complex compositional processing.

5. Conclusion

This paper conveyed several new findings regarding adjective-noun composition in the human brain as we tracked the flow of information during semantic composition. Our analysis showed that adjective semantics are predictable for an extended period of time, continuously until 1.6s after the onset of the adjective, and are reactivated during late processing, 2-3s after the onset of the adjective (1.2-2.2s after the onset of the noun). The reactivated neural representation matches the representation seen during the initial reading of the adjective. After the offset of each word, a negated representation of the word appears, possibly mediated by the right hemisphere. Semantic representations are oscillatory and alpha-entrained, as evidenced by the strong diagonal peaks and troughs of accuracy in the TGMs.

The resurgence of adjective semantics is much later than the activation of the machinery responsible for combinatorics, as documented in previous research [4, 7]. Perhaps the combinatorial machinery of LATL and LIFG is the hub that coordinates areas, readying them for the compositional processing. This would require them to activate sooner than areas of the brain that *store* the composed semantic meaning. Our results imply that future research interested in the composed representation should look beyond the typical 1s time window after the onset of a word.

With respect to semantic composition in the brain, several new research questions have emerged. For example, does adjective resurgence appear even for non-intersective adjectives? We would also like to explore more complex composition tasks like sentences, paragraphs, stories, and beyond. We are interested in the underlying mechanisms that give rise to significantly below chance accuracy, and would like to explore their role in compositional processing. Our work also raises several new questions regarding the role of oscillations in the neural processing of language. By exploring simple composition in a controlled setting, our study establishes an analysis framework for such future research directions.

Acknowledgements

This material is based upon work supported by the National Institutes of Health grant number 1r01hd075328-0, IARPA afrl award #31081.1 and the CNBC Multimodal Neuroimaging Training Program Fellowship.

References

[1] M. Kutas and S. SA Hillyard. Reading senseless sentences: brain potentials reflect semantic incongruity. *Science*, 207(4427):203–5, 1980.

[2] Gina R Kuperberg. Neural mechanisms of language comprehension: challenges to syntax. *Brain research*, 1146:23–49, may 2007.

[3] Peter Hagoort. On Broca, brain, and binding: a new framework. *Trends in cognitive sciences*, 9(9):416–23, sep 2005.

[4] Douglas K Bemis and Liina Pylkkänen. Simple composition: a magnetoencephalography investigation into the comprehension of minimal linguistic phrases. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, 31(8):2801–14, feb 2011.

[5] Masha Westerlund and Liina Pylkkänen. The role of the left anterior temporal lobe in semantic composition vs. semantic memory. *Neuropsychologia*, 57:59–70, may 2014.

[6] Marcel Bastiaansen, Lilla Magyari, and Peter Hagoort. Syntactic unification operations are reflected in oscillatory dynamics during on-line sentence comprehension. *Journal of cognitive neuroscience*, 22(7):1333–47, jul 2010.

[7] D K Bemis and L Pylkkänen. Basic linguistic composition recruits the left anterior temporal lobe and left angular gyrus during both listening and reading. *Cerebral cortex (New York, N.Y. : 1991)*, 23(8):1859–73, aug 2013.

[8] L Osterhout and P J Holcomb. Event-related brain potentials elicted by syntactic anomaly. *Journal of Memory and Language*, 31:785–806, 1992.

[9] Peter D Turney and Patrick Pantel. From Frequency to Meaning : Vector Space Models of Semantics. *Journal of Artificial Intelligence Research*, 37:141–188, 2010.

[10] William Blacoe and Mirella Lapata. A Comparison of Vector-based Representations for Semantic Composition. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 546–556, Jeju Island, Korea, 2012.

[11] Elia Bruni and Marco Baroni. Multimodal Distributional Semantics. *Journal of Artificial Intelligence Research*, 48, 2013.

[12] Jeffrey Pennington, Richard Socher, and Christopher D Manning. GloVe : Global Vectors for Word Representation. In *Conference on Empirical Methods in Natural Language Processing*, Doha, Qatar, 2014.

[13] Marco Baroni, Georgiana Dinu, and German Kruszewski. Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *52nd Annual Meeting of the Association for Computational Linguistics*, 2014.

[14] T Landauer and S Dumais. A solution to Plato's problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104(2):211–240, 1997.

[15] Tom M Mitchell, Svetlana V Shinkareva, Andrew Carlson, Kai-Min Chang, Vicente L Malave, Robert A Mason, and Marcel Adam Just. Predicting human brain activity associated with the meanings of nouns. *Science (New York, N.Y.)*, 320(5880):1191–5, may 2008.

[16] Gustavo Sudre, Dean Pomerleau, Mark Palatucci, Leila Wehbe, Alona Fyshe, Riitta Salmelin, and Tom Mitchell. Tracking Neural Coding of Perceptual and Semantic Features of Concrete Nouns. *NeuroImage*, 62(1):463–451, may 2012.

[17] Brian Murphy, Partha Talukdar, and Tom Mitchell. Selecting Corpus-Semantic Models for Neurolinguistic Decoding. In *First Joint Conference on Lexical and Computational Semantics (*SEM)*, pages 114–123, Montreal, Quebec, Canada, 2012.

[18] Leila Wehbe, Ashish Vaswani, Kevin Knight, and Tom Mitchell. Aligning context-based statistical models of language with brain activity during reading. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014.

[19] Leila Wehbe, Brian Murphy, Partha Talukdar, Alona Fyshe, Aaditya Ramdas, and Tom Mitchell. Simultaneously Uncovering the Patterns of Brain Regions Involved in Different Story Reading Subprocesses. pages 1–19, 2014.

[20] Joachim Gross, Sylvain Baillet, Gareth R. Barnes, Richard N. Henson, Arjan Hillebrand, Ole Jensen, Karim Jerbi, Vladimir Litvak, Burkhard Maess, Robert Oostenveld, Lauri Parkkonen, Jason R. Taylor, Virginie van Wassenhove, Michael Wibral, and Jan-Mathijs Schoffelen. Good-practice for conducting and reporting MEG research. *NeuroImage*, oct 2012.

[21] Samu Taulu and Riitta Hari. Removal of magnetoencephalographic artifacts with temporal signal-space separation: demonstration with single-trial auditory-evoked responses. *Human brain mapping*, 30(5):1524–34, may 2009.

[22] M A Uusitalo and R J Ilmoniemi. Signal-space projection method for separating MEG or EEG into components. *Medical & biological engineering & computing*, 35(2):135–40, mar 1997.

[23] M S Hämäläinen and R J Ilmoniemi. Interpreting magnetic fields of the brain: minimum norm estimates. *Medical & biological engineering & computing*, 32:35–42, 1994.

[24] Alona Fyshe, Partha Talukdar, Brian Murphy, and Tom Mitchell. Documents and Dependencies: an Exploration of Vector Space Models for Semantic Composition. In *Computational Natural Language Learning*, Sofia, Bulgaria, 2013.

[25] Yoav Benjamini and Daniel Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, 29(4):1165–1188, 2001.

[26] J-R. King and S. Dehaene. Characterizing the dynamics of mental representations: the temporal generalization method. *Trends in Cognitive Sciences*, 18(4):203–210, 2014.

[27] Lluís Fuentemilla, Will D. Penny, Nathan Cashdollar, Nico Bunzeck, and Emrah Düzel. Theta-Coupled Periodic Replay in Working Memory. *Current Biology*, 20(7):606–612, 2010.

[28] Ethan M Meyers, David J Freedman, Gabriel Kreiman, Earl K Miller, and Tomaso Poggio. Dynamic population coding of category information in inferior temporal and prefrontal cortex. *Journal of Neurophysiology*, 100(June 2008):1407–1419, 2008.

[29] Rahul S. Desikan, Florent Segonne, Bruce Fischl, Brian T. Quinn, Bradford C. Dickerson, Deborah Blacker, Randy L. Buckner, Anders M. Dale, R. Paul Maguire, Bradley T. Hyman, Marilyn S. Albert, and Ronald J. Killiany. An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *NeuroImage*, 31(3):968–980, 2006.

[30] Douglas K Bemis and Liina Pylkkänen. Flexible composition: MEG evidence for the deployment of basic combinatorial linguistic mechanisms in response to task demands. *PloS one*, 8(9):e73949, jan 2013.

[31] Sean G Baron and Daniel Osherson. Evidence for conceptual combination in the left anterior temporal lobe. *NeuroImage*, 55(4):1847–52, apr 2011.

[32] Ksenija Marinkovic, Sharelle Baldwin, Maureen G Courtney, Thomas Witzel, Anders M Dale, and Eric Halgren. Right hemisphere has the last laugh: neural dynamics of joke appreciation. *Cognitive, affective & behavioral neuroscience*, 11(1):113–30, mar 2011.

[33] Katherine a DeLong, Laura Quante, and Marta Kutas. Predictability, plausibility, and two late ERP positivities during written sentence comprehension. *Neuropsychologia*, 61:150–62, aug 2014.

[34] Cyma Van Petten and Barbara J. Luka. Prediction during language comprehension: Benefits, costs, and ERP components. *International Journal of Psychophysiology*, 83(2):176–190, 2012.

[35] Ole Jensen, Bart Gips, Til Ole Bergmann, and Mathilde Bonnefond. Temporal coding organized by coupled alpha and gamma oscillations prioritize visual processing. *Trends in neurosciences*, 37(7):357–369, jul 2014.

[36] G. Pfurtscheller, a. Stancák, and Ch. Neuper. Event-related synchronization (ERS) in the alpha band an electrophysiological correlate of cortical idling: A review. *International Journal of Psychophysiology*, 24(1-2):39–46, nov 1996.

[37] Mathilde Bonnefond and Ole Jensen. Gamma Activity Coupled to Alpha Phase as a Mechanism for Top-Down Controlled Gating. *Plos One*, 10(6):e0128667, 2015.

[38] Rufin Vanrullen and James S P MacDonald. Perceptual echoes at 10 Hz in the human brain. *Current Biology*, 22(11):995–999, 2012.

[39] Hinze Hogendoorn, Frans A.J. Verstraten, and Patrick Cavanagh. Strikingly rapid neural basis of motion-induced position shifts revealed by high temporal-resolution EEG pattern classification. *Vision Research*, 113:1–10, 2015.

[40] Peter Hagoort, Lea Hald, Marcel Bastiaansen, and Karl Magnus Petersson. Integration of Word Meaning and World Knowledge in Language. *Science*, 304(April):438–442, 2004.

[41] Marcel C M Bastiaansen, Marieke van der Linden, Mariken Ter Keurs, Ton Dijkstra, and Peter Hagoort. Theta responses are involved in lexical-semantic retrieval during language processing. *Journal of cognitive neuroscience*, 17(3):530–41, 2005.

[42] Marcel C M Bastiaansen, Robert Oostenveld, Ole Jensen, and Peter Hagoort. I see what you mean: Theta power increases are involved in the retrieval of lexical semantic information. *Brain and Language*, 106(1):15–28, 2008.

[43] H Bhatt, E M Hedgecock, S Kim, E Fox, and E M Hedgecock. Dynamic Brain Sources of Visual Evoked Responses. *Science*, 295(January):690–694, 2002.

[44] Frederic Roux and Peter J. Uhlhaas. Working memory and neural oscillations: Alpha-gamma versus theta-gamma codes for distinct WM information? *Trends in Cognitive Sciences*, 18(1):16–25, 2014.

Appendix

Stimuli

The phrases for the adjective noun brain imaging experiment are made from 6 nouns (“dog”, “bear”, “tomato”, “carrot”, “hammer”, “shovel”) and 8 adjectives (“big”, “small”, “ferocious”, “gentle”, “light”, “heavy”, “rotten”, ’‘tasty”), as well as two null words: “the” and “thing”. The phrases are:

- the dog
- the bear
- the tomato
- the carrot
- the hammer
- the shovel
- big dog
- big bear
- big tomato
- big carrot
- big hammer
- big shovel
- small dog
- small bear
- small tomato
- small carrot

- small hammer
- small shovel
- ferocious dog
- ferocious bear
- gentle dog
- gentle bear
- light hammer
- light shovel
- heavy hammer
- heavy shovel
- rotten carrot
- rotten tomato
- tasty carrot
- tasty tomato

Due to multiple word senses, the word “light” was not used in the adjective-adjective oddballs.

ROI Membership

For all ROI analyses in the paper, temporal ROIs include: superior temporal gyrus, middle temporal gyrus, inferior temporal gyrus, banks of the superior temporal sulcus, and fusiform gyrus; inferior frontal gyrus (IFG) includes: pars opercularis, pars orbitalis, and pars triangularis; occipital includes: lingual, lateral occipital, and cuneus; parietal includes superior parietal, inferior parietal and the precuneus. Post central and supramarginal gyrus were omitted from the parietal ROIs because their performance was much lower than the ROIs we included here.