# Predicting the Semantic Category of Internally Generated Words from Neuromagnetic Recordings

Irina Simanova[1], Marcel A. J. van Gerven[1], Robert Oostenveld[1], and Peter Hagoort[1,2]

## Abstract

■ In this study, we explore the possibility to predict the semantic category of words from brain signals in a free word generation task. Participants produced single words from different semantic categories in a modified semantic fluency task. A Bayesian logistic regression classifier was trained to predict the semantic category of words from single-trial MEG data. Significant classification accuracies were achieved using sensor-level MEG time series at the time interval of conceptual preparation. Semantic category prediction was also possible using source-reconstructed time series, based on minimum norm estimates of cortical activity. Brain regions that contributed most to classification on the source level were identified. These were the left inferior frontal gyrus, left middle frontal gyrus, and left posterior middle temporal gyrus. Additionally, the temporal dynamics of brain activity underlying the semantic preparation during word generation was explored. These results provide important insights about central aspects of language production. ■

## INTRODUCTION

One of the basic properties of the conceptual system is its categorical organization. The ability of the brain to build categories and to generalize across objects with similar sensory and functional properties enables us to recognize and memorize new objects efficiently. Developmental studies show that perceptual categorization develops during the first months of life (Mandler, 2004), and multiple neuroimaging studies have reported differential activation for stimuli from different semantic domains (Mahon & Caramazza, 2009; Gerlach, 2007; Martin & Chao, 2001; Chao & Martin, 2000; Chao, Haxby, & Martin, 1999; Caramazza & Shelton, 1998; Martin, Wiggs, Ungerleider, & Haxby, 1996; Perani et al., 1995). More recent research has demonstrated that the semantic category or entity of an object can be successfully predicted from neural activity patterns when the object is presented visually (Murphy et al., 2011; Reddy & Kanwisher, 2007; Haynes & Rees, 2006; Kamitani & Tong, 2005; Cox & Savoy, 2003; Haxby et al., 2001) or orthographically (Simanova, Hagoort, Oostenveld, & van Gerven, 2014; Chan, Halgren, Marinkovic, & Cash, 2011; Murphy et al., 2011; Shinkareva, Malave, Mason, Mitchell, & Just, 2011; Simanova, van Gerven, Oostenveld, & Hagoort, 2010). It can be argued, however, that the use of visual or orthographical stimuli in these tasks introduce confounding visual or phonological effects related to semantic retrieval (Hwang, Palmer, Basho, Zadra, & Müller, 2009). An additional concern, more specific to visual stimulus presentation, is that certain perceptual attributes that are different between categories can bias the decoding outcome (Vindiola & Wolmetz, 2011; Simanova et al., 2010). It is possible to minimize these confounds by using an internally guided word generation task, without presenting any pictures or words to participants. Moreover, in this way central aspects of the process of language production can be investigated, instead of the common focus on language comprehension in this type of studies.

Few studies have used word generation to explore category-specific semantic representations. Vitali et al. (2005) examined fMRI activity and connectivity between cortical areas during silent production of tool and animal words. The authors report increased activation in the left frontal, left temporal, and parietal regions for tool compared with animal words. Another fMRI study (Hwang et al., 2009) reported that conceptual processing in speech preparation involves left lateral frontal cortex across different word categories. Category-specific activations were distributed across sensorimotor and perceptual cortices, according to semantic attributes of the word (Hwang et al., 2009). However, because both these studies used fMRI, it was not possible to explore the temporal dynamics of the reported effects and to separate processes underlying different stages of speech production.

Recently, van de Nieuwenhuijzen et al. (2013) demonstrated that categorical information can be decoded from temporally precise whole-head high-density MEG signals. In the experiment, participants viewed images of faces,

[1] Donders Institute for Brain, Cognition, and Behavior, Radboud University Nijmegen, [2] Max Planck Institute for Psycholinguistics, Nijmegen

© 2014 Massachusetts Institute of Technology

doi:10.1162/jocn_a_00690

objects, bodies and tools. Using source space signal time courses, the authors reconstructed the dynamics of visual category perception. In the present study, we set out to explore the categorical differences during internally guided word generation. Similar to the study by van de Nieuwenhuijzen et al. (2013), we use MEG and apply multivariate decoding to investigate category-specific semantic effects.

We used a constrained verbal fluency paradigm. Participants were asked to produce single words of different semantic categories (animals, nonliving objects, or countries). Participants were presented with a cue for the semantic category and a cue for the initial letter. They were asked to report with a button press if they had a word in mind that matched the cue requirements. With this, we achieved control over timing of responses and reduced the amount of speech-related artifacts. Following a short maintenance period, participants produced the word overtly, providing us with a measure of task performance. We employed whole-brain multivariate analysis of magnetoencephalographic recordings before speech production to find the neurophysiological markers of semantic processing. Three different approaches were used: classification was performed on the signal time series at the sensor level as well as on temporally and spatially decomposed time series. The source-level time series analysis made it possible to identify brain regions that contributed most to category prediction. We discuss the results in the light of the literature on the neurobiology of semantic memory and single-word production.

## METHODS

### Participants

Eighteen healthy native Dutch-speaking participants took part in the study (five men, age = 18–25 years, mean age = 21 ± 2 years). Data from two more participants were not included in the current study because of excessive head movements. One more participant was excluded because the experimental session was interrupted. All participants were right-handed and reported that they did not suffer from any psychological or neurological disorders. The study was approved by the local ethics committee (Commissie Mensgebonden Onderzoek Regio Arnhem-Nijmegen). All the participants gave written informed consent before the experiment. Participants received either monetary compensation or course credits for their participation.

### Experimental Design

Participants were asked to produce single words of different semantic categories: animals, nonliving objects, and countries (the latter was used as a filler category and is not included in the presented MEG analysis). Participants were presented with the semantic category cue and then the initial letter cue. Subsequently, they were asked to report with a button press if they had generated a word that fulfills the requirements indicated by the cues and to produce the word overtly after a short maintenance period.

The experiment was organized in short blocks, with the target semantic category alternating between blocks. Each block included 10 trials. At the beginning of each block, the category title was displayed on the screen for 4 sec. The experiment consisted of nine blocks per task category (90 trials per category). The order of presented letters and the order of blocks were alternated across participants. The letters were selected according to their frequency in initial position in Dutch nouns based on CELEX (Max Planck Institute for Psycholinguistics, the Netherlands, 2001). For example, letter S appeared more often than D, whereas letters X, Y, and Q were not presented, because words beginning with these letters are very infrequent in Dutch language.

The time structure of a single trial is shown in Figure 1. Each experimental trial started with a 1000-msec baseline interval. Subsequently, a letter was displayed for 200 msec. Participants were instructed to press a button to report that they had a suitable word in mind. If no button press was registered within the 3500 msec response interval, the trial was terminated (no-response trial). If the participant pressed the button within this interval, the trial was extended for an additional 1500 msec. Subsequently, an icon depicting a microphone appeared on the screen.

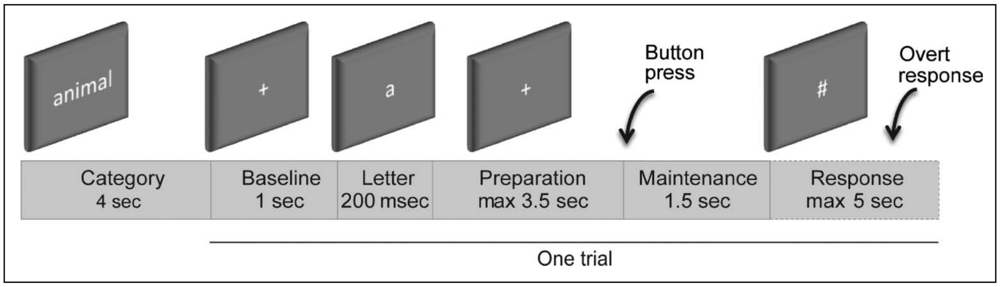


**Figure 1.** The structure of an experimental trial. The experiment was organized in short blocks, with the target semantic category alternating between blocks. Each block included 10 trials. In the beginning of the block, the category title was presented, followed by the letter cue. Participants were instructed to press a button if they had a word in mind that fulfilled both cues. If the participant pressed the button within the responsive interval of 3500 msec, the trial was extended for an additional 1500 msec. Subsequently, the participant was invited to vocalize the word. The next trial started in 2000 msec with a new letter cue (see Experimental Design for more details).

Now the participant was invited to vocalize the word. The next trial started in 2000 msec. A fixation cross was presented on the screen before presentation of the cues. All overt responses were registered with a microphone and recorded to a hard drive.

The experiment lasted approximately 60 min and included two breaks. Participants remained seated during the breaks. After the experiment participants were asked to fill in an evaluation form. They had to indicate task difficulty, separately for each of the three categories on a scale from 1 to 10 (1 = *very easy*; 10 = *very difficult*). They also had to describe in writing the strategy they used when performing the task.

### Analysis of Task Performance

All recorded overt responses were annotated by a native Dutch speaker. For each word the semantic category was identified to check participants' compliance with the task. The most commonly produced words from all participants are summarized in Table 1. For each trial, the RT (time from the letter presentation to the button press) was extracted. A two-way ANOVA with participant as a random factor and category as a fixed factor was used to test for the differences in RTs in different categories across participants (IBM SPSS Statistics 19). A two-sample paired *t* test was used to test for the difference in the number of responses for two categories. The behavioral measures were later correlated with MEG results (see Analysis of Confounding Factors).

### MEG Acquisition

MEG data were acquired with a 275-sensor axial gradiometer system (CTF systems Inc., Port Coquitlam, Canada) in a magnetically shielded room. Participants were seated comfortably in a chair with their head inside the sensor helmet. They were told not to move during the experiment and to fixate their eyes on a back-projection screen placed in front of them.

The participant's head position was determined using coils positioned at the participant's nasion and in the left and right ear canal. Continuous registration of head movements relative to the original position was performed during the experiment. Additionally, horizontal and vertical EOGs as well as EMGs from two electrodes placed above and below the participant's mouth were recorded. The ongoing EOG, EMG, and MEG signals were digitized at 1200 Hz and stored for offline analysis. Overt responses were recorded by a microphone mounted on the wall of the experimental room.

In addition, a whole-brain high-resolution structural T1-weighted MP-RAGE sequence was used to image each participant's anatomy (repetition time = 2300 msec, 192 slices with voxel size of 1 mm$^3$, field of view = 256°), accelerated with GRAPPA parallel imaging.

**Table 1.** Twenty Most Commonly Produced Words from Each Task Category

| *Animals* | | | *Nonliving Objects* | | |
|---|---|---|---|---|---|
| hond | dog | 16 | jas | coat | 9 |
| aap | monkey | 15 | mes | knife | 9 |
| vis | fish | 15 | tafel | table | 9 |
| uil | owl | 14 | vork | fork | 7 |
| egel | hedgehog | 12 | bal | ball | 7 |
| muis | mouse | 12 | boek | book | 6 |
| beer | bear | 11 | pen | pen | 6 |
| olifant | elephant | 11 | sleutel | key | 5 |
| paard | horse | 10 | fiets | bicycle | 5 |
| rat | rat | 10 | nietmachine | stapler | 5 |
| slang | snake | 10 | plank | shelf | 5 |
| vogel | bird | 10 | zaag | saw | 5 |
| kat | cat | 10 | kast | cupboard | 5 |
| neushoorn | rhino | 9 | lamp | lamp | 5 |
| giraffe | giraffe | 8 | beker | cup | 4 |
| arend | eagle | 8 | lepel | spoon | 4 |
| tijger | tiger | 7 | afstandsbediening | remote control | 4 |
| kangeroe | kangaroo | 7 | bank | sofa | 4 |
| koe | cow | 7 | deur | door | 4 |
| ooievaar | stork | 7 | kapstok | coat-peg | 4 |

Dutch words are given with English translation. Numbers indicate the number of participants (out of 18) who mentioned the word upon encountering the initial letter.

### MEG Preprocessing

All preprocessing and analysis steps were performed using MATLAB R2011a (The MathWorks, Inc., Natick, MA) and FieldTrip, an open source Matlab toolbox for the analysis of neuroimaging data (Oostenveld, Fries, Maris, & Schoffelen, 2011).

The data segments from −200 msec before letter presentation up to 1500 msec after the button press were extracted. Segments containing system-related artifacts or muscular activity were identified based on signal variance. Identified segments were inspected visually and rejected if contamination with artifacts was confirmed. The number of rejected trials varied across participants from 9 to 76 (on average 15 ± 7% of the total number of trials). In the remaining data, line noise (50 Hz and harmonics) was removed using a discrete Fourier transform. The data were subsequently resampled at 300 Hz and baseline corrected to 200 msec of the baseline interval. Subsequently, independent component analysis (ICA) was performed (Makeig, Bell, Jung, & Sejnowski, 1996). Components explaining horizontal and vertical eye movements, eye blinks, and ECG were discarded based on visual inspection. Sensor-level time series were reconstructed from the remaining components. At the same time, the remaining components were stored for subsequent classification.

After preprocessing, all data segments were redefined such that the zero time point corresponded to the button press. The time interval from −500 to 0 msec before the button press was used for analysis. After exclusion of contaminated and no-response trials, the number of remaining trials varied from 85 to 221 per participant (on average 150 ± 38).

### Minimum Norm Estimate Source Reconstruction

We used dynamic statistical parametric mapping to reconstruct the sources of neuronal activity (Dale et al., 2000). This minimum norm estimate method is favorable when there are no a priori assumptions on the location and/or number of current sources (Hämäläinen & Ilmoniemi, 1994). For each participant, a volume conductor model of the head was created using a single-shell approximation based on an individual segmented structural image (Nolte, 2003). The source space was defined as a triangulated cortical mesh, consisting of ~8000 approximately equally sized triangles. Cortical mesh reconstruction was performed with the Freesurfer image analysis suite (surfer.nmr.mgh.harvard.edu/). The volume conductor model and the cortical mesh, as well as the gradiometer positions for individual participants, were used to create the forward model. Finally, the inverse solution was computed, using the minimum norm estimate of the cortical activity at the selected time interval. The noise covariance matrix was estimated at the time window from −100 to 0 msec preceding the button press. The noise covariance-scaling factor was set to $10^{-8}$. The source time series were then computed at the time interval from −500 to 0 msec before the button press.

### MEG Single-trial Analysis

The choice of the time window for MEG analysis was based on the duration of the shortest trial (i.e., the minimal amount of time required to perform the task). In all participants, the minimal duration of trials was around 550 msec. Therefore, we chose the interval of 500 msec before the button press for the analysis, assuming that the process of conceptual preparation takes place within this interval. We also took into account previous literature on word production. On the basis of a meta-analysis of a large number of picture naming studies, Indefrey and Levelt concluded that conceptual preparation and lexical selection occur approximately 600–350 msec before the response in overt naming (Indefrey, 2011; Indefrey & Levelt, 2004; see also Levelt et al., 1998). Although the preparation stages in picture naming could be different from the current task, we assume that the speed of conceptual preparation and lexical selection lies approximately within the range of 500 msec.

In the first analysis, single-trial classification of the MEG data was performed at this time interval. A Bayesian logistic regression classifier (Simanova et al., 2010; van Gerven, Cseke, de Lange, & Heskes, 2010) was trained to identify the semantic category (animals vs. nonliving objects) of the to-be-produced word in each trial. The method is explained in detail in the Supplementary Text 1. A fivefold cross-validation was performed in which the data set was partitioned into five random subsets. Thus, after training on 80% of the data, the classification algorithm was tested separately on each trial within the remaining 20% of the data. This process was repeated five times. Classification accuracy (proportion of correctly classified trials) was computed in each fold and then averaged to produce a single accuracy estimate. The number of trials belonging to each of the two categories was balanced before classification: The number of trials in the minority class was estimated, and the same number of trials was randomly selected from the majority class. In all tests, before classification, the signal over all trials was standardized to have zero mean and a standard deviation of one.

All classification analyses were repeated with three different projections of the data: (i) the sensor-level time series, (ii) the reconstructed source-level time series, and (iii) the component time series produced by ICA (see MEG Preprocessing). In all three types of classification, single-trial time series, without averaging or collapsing over time, were used as input to the classifier. Classification accuracies obtained with these three data projections were compared using one-way ANOVA.

The category "countries" served as a filler category in the experiment and was not included in the current analysis. We focused on the contrast between animals and nonliving objects because differences in semantic processing of these categories are well studied (Martin, 2007; Gerlach, Law, & Paulson, 2002; Caramazza & Shelton, 1998), which allows us to link our results to the existing literature.

### Statistical Analysis of Classification Accuracies

Chance-level accuracy was 0.5 in all tests. Within participants, the significance of each classification outcome was computed using a binomial test, which compares the performance of the trained classifier with that of a baseline classifier that assigns all trials to same class (Salzberg, 1997). Between participants, the resulting classification accuracies from the group (18 participants) were compared against the chance level of 0.5 using a right-handed *t* test.

### Analysis of Confounding Factors

As is evident from the analysis of RTs, in some participants the response in nonliving object trials took longer than in animal trials. RT differences could bias the classification in the interval before the button press, because short trials may contain distinctly different cognitive processes than long trials. To explore the possible impact of this confounding measure on classification performance, we analyzed if the classifier's predictions depended on the trials' length. We computed the Pearson correlation between the RTs in individual trials and classification outcome—the probability that each trial belongs to the category "animals." *p* values were computed against the hypothesis that there was a negative correlation between these measures (i.e., the classifier tends to predict the class "animals" for shorter trials). This analysis was performed for each participant based on the sensor-level classification results.

As described in the Experimental Design section, the experiment consisted of short blocks, and the task category alternated between blocks. It is therefore possible that the block structure biased the classifier's performance. One possibility is that the classifier predicted that trials belong to the same block (based on correlations or drifts in the signal), rather than same category. Another possibility is that the mental state induced by the target category at the beginning of each block could be decoded during the entire duration of the block, irrespectively of the word production task. In both cases the classifier would be able to decode the semantic category from the MEG signal during the baseline period, before the letter presentation. To address this possibility, we conducted classification analysis at the baseline interval. For each participant, the classification (same as the main procedure) was applied at the sensor-level MEG data in an interval of 200 preceding the presentation of the letter in each trial.

### Feature Localization

To identify which data features were used in the Bayesian logistic regression for predicting the semantic category, the maps indicating the importance of data features were extracted (Simanova et al., 2010; van Gerven et al., 2010). The relative importance of data features is expressed in terms of the variance of auxiliary variables in the Bayesian logistic regression classifier (see Simanova et al., 2010, for details). The importance weights associated with each data feature were visualized as feature maps (Chan et al., 2011; Simanova et al., 2010; van Gerven et al., 2009). We focus on feature maps obtained for source-level classification. In each participant, the importance weights for each location were averaged across all time points of the analyzed interval. Resulting values were mapped onto the cortical surface of individual participants and then interpolated to a three-dimensional grid. This grid was defined as an individualized warp of a template grid, and therefore, the resulting three-dimensional maps were directly interpretable in Montreal Neurological Institute (MNI) space. The maps were then averaged across participants to produce a group estimate of the spatial distribution of important features. Labels identifying each brain region were extracted using the Anatomical Automatic Labeling toolbox (Tzourio-Mazoyer et al., 2002).

For the group-level analysis of the feature maps, we thresholded individual maps at the 0.95 quantile of the distribution of importance weights. Subsequently, a group-level binomial test was applied, under the .05 probability of success at each location, to retain only those grid points that survived statistical testing.

### Analysis of Temporal Dynamics

We additionally repeated the classification analysis in the source level on 100 msec time bins, covering the interval from −500 msec to the button press. For each time bin, the classification analysis (see "MEG Single-trial Analysis") as well as feature localization (see "Feature Localization") were performed. We report the averaged classification accuracy across 18 participants and group-level feature localization maps at each 100-msec time bin.

## RESULTS

### Task Performance

On average participants gave a response within the 3500-msec response interval in two-thirds of all trials (58 ± 15 words for animals and 56 ± 12 words for objects). There was no significant difference in the number of responses between categories. There was, however, a difference in the RTs between two categories. Participants tended to respond faster for animals (mean RT = 1630 ± 780 msec) than nonliving objects (1840 ± 780 msec), $F = 11.3(1)$, $p < .01$. In the postexperimental evaluation, some participants indicated that the category nonliving objects was the most difficult. On the basis of the ratings from all participants, however, the difference between categories in the task difficulty was not significant: Average rating scores were 5.3 ± 2.5 for animals and 6.3 ± 1.8 for nonliving objects.

Inspection of the most commonly produced words (Table 1) suggests that there was less agreement between participants in nonliving objects than in animals. Altogether, the behavioral results indicate that the task was not of equal difficulty among the categories.

After the experiment, participants were asked to report what strategy they used to name objects. Most of the participants indicated that they used visualization of certain subcategories when performing the task. For example, to name an animal, they would think of a particular animal group, for instance, farm animals or zoo animals. For naming nonliving objects, the most common strategy was to list common household objects by visualizing a kitchen or a living room. For the category "countries," all participants reported that they imagined a map and "read" names, following the geographical proximity of countries. On the basis of these strategy reports, we decided not to include the data from the latter category in the final analysis, as mentioned above. Because of the differences

**Table 2.** Mean Accuracies When Performing Classification Based on Sensor Space Data, Source Space Data and ICA-decomposed Data

| | *Mean Accuracy* | *SD* | *p* | *No. of Significant Participants* |
|---|---|---|---|---|
| Sensor level | 0.56 | 0.02 | $p < .01$ | 7 |
| Source level | 0.58 | 0.01 | $p < 10e{-4}$ | 10 |
| ICA | 0.61 | 0.01 | $p < 10e{-6}$ | 11 |

The *p* values resulting from the group-level statistical analysis of classification accuracies are shown in the third column. The last column shows the number of participants for whom the classification accuracies were significantly higher than chance according to the within-subject statistical analysis.

in strategy, the outcome of countries–animals and countries–objects comparisons would be very difficult to interpret.

## MEG Single-trial Analysis

Table 2 summarizes results of the single-trial classification in the conceptual preparation interval. There were differences in the classification performance across participants; whereas for some participants classification was at chance, for others it was significantly above chance. The maximal accuracy was 0.69 for the sensor space classification ($p < 10^{-6}$ in the within participant's binomial test), 0.67 for the source space classification ($p < 10^{-5}$), and 0.70 for the ICA-based classification ($p < 10^{-3}$). We observed that ICA decomposition increased classification performance slightly. One-way ANOVA showed a marginal difference in the classification accuracies between ICA, sensor, and source-based classification across participants ($F = 2.85$, $df = (2, 34)$, $p = .07$). None of the separate post hoc comparisons was significant after Bonferroni correction.

Both sensor-based and ICA-based analyses did not allow for comparing features across participants. The position of MEG sensors relative to the brain is different in all participants, preventing the comparison of individual sensor-level feature maps. It is also difficult to match ICA components from different participants, because the spatial filters are determined individually. To circumvent these difficulties, we performed the classification analysis in source space. The classifier performed slightly better at the source than at the sensor level. By further warping individual feature maps to MNI space, we were able to compare them across participants and localize brain regions that contributed to category discrimination at the group level.

## Feature Localization Results

Single-trial analysis on the source level allowed for mapping classifier parameters to native brain space. Feature maps from two participants with highest classification accuracies in source space are shown in Figure 2, whereas

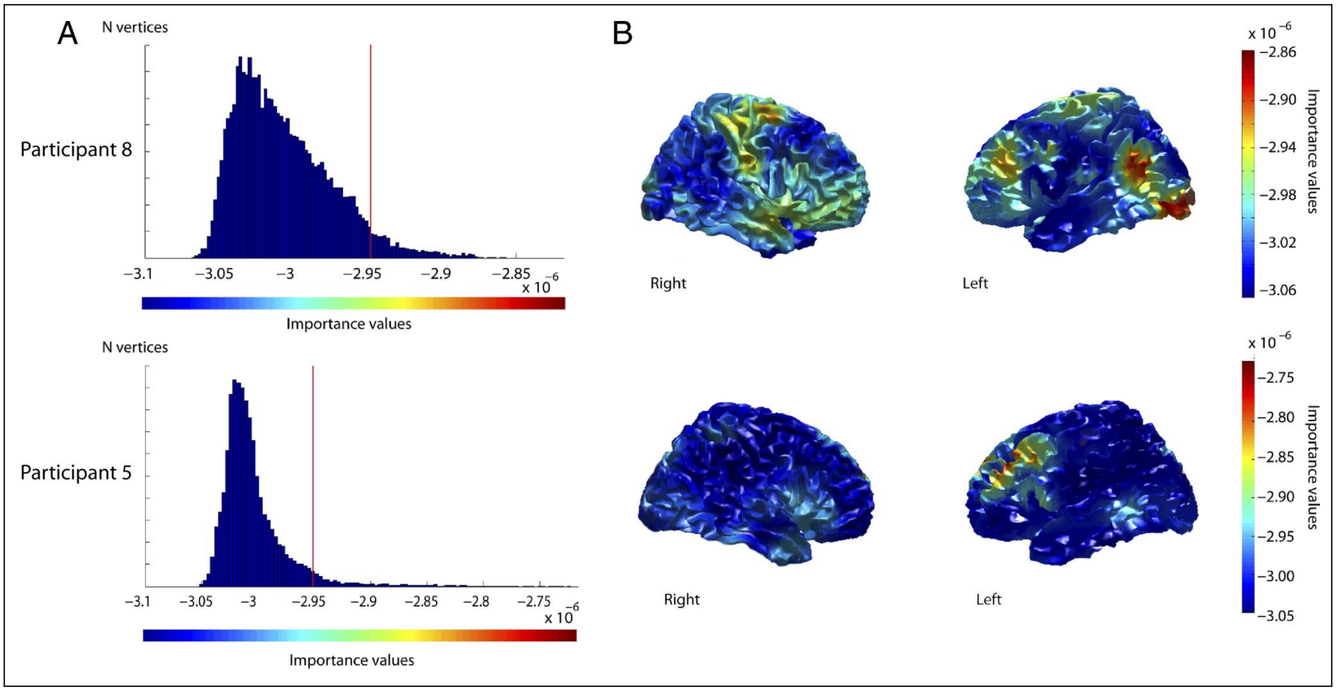


**Figure 2.** Histograms (A) and surface maps (B) of the feature importance in two participants (animals–objects classification based on the source time series). Feature importance was quantified in terms of the variance of auxiliary variables in the Bayesian logistic regression classifier. The importance weights for each location were averaged across all time points of the analyzed interval. Red line on A marks the 0.95 quantile of the distribution, the threshold used for the group-level statistical analysis. Color bar under the histogram on A shows how the weights are scaled on B. Blue color represents low importance, and red represents high importance. For this figure, we chose two participants with highest classification accuracies in the source space. Feature maps from all 18 experimental participants are shown in Supplementary Figure 1.

**Figure 3.** Averaged importance map masked with the results of the group-level statistical analysis ($p < .05$, uncorr., $n = 18$).

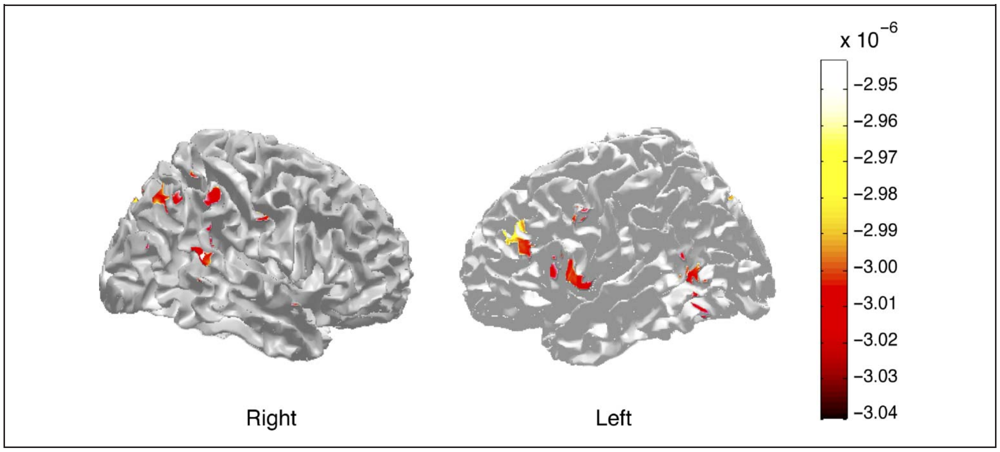


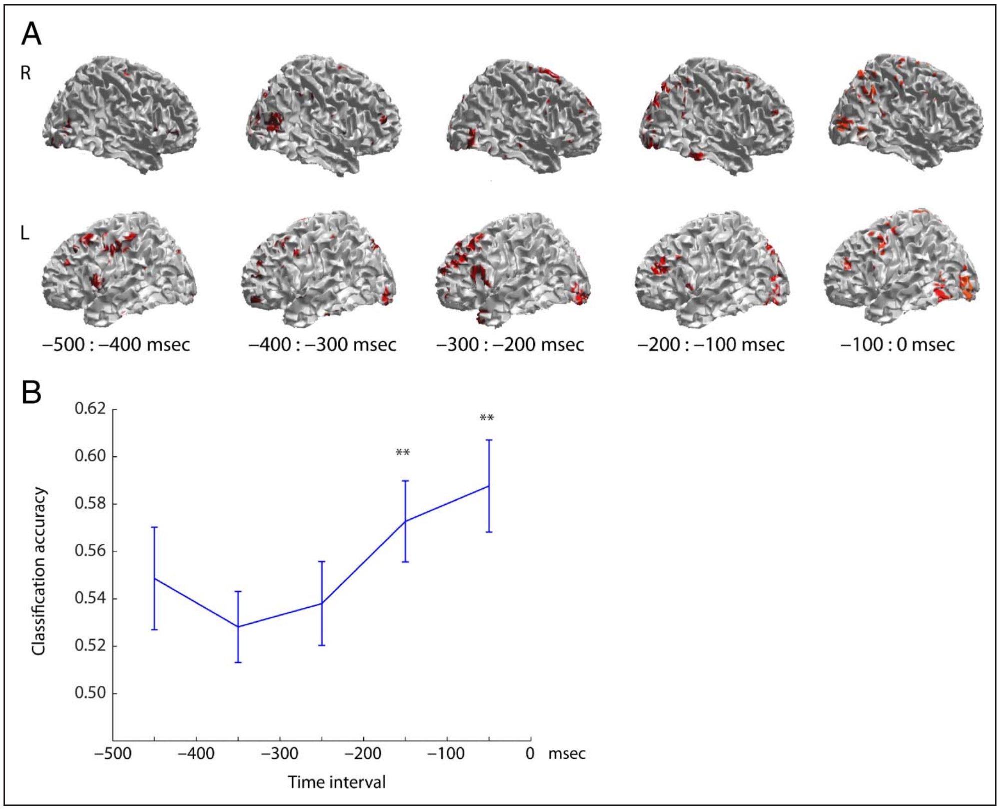
Figure 3 shows the outcome of the group-level statistical analysis. That is, Figure 3 depicts the group-averaged map masked at $p < .05$ (uncorrected). Several salient brain regions can be identified in the averaged map. These are the left middle frontal gyrus (peak location at [−40 44 32]), left posterior middle temporal gyrus ([−52 60 12]), left inferior frontal gyrus ([−56 12 12]), regions in the right parietal lobe ([40 −64 48] and [56 −36 48]), and right precuneus ([20 −80 48]).

For the clusters located in left middle frontal, inferior frontal, and posterior middle temporal gyrus, the group level $p$ values were lower than $10^{-2}$. However, none of the clusters survived a more stringent multiple comparison correction. Therefore, the maps in Figure 3, as well as in Figure 4, are included for descriptive, not inferential purposes.

## Analysis of Confounding Factors

We tested if the classifier's prediction in single trials correlated with the RTs. We found a weak significant positive correlation ($p < .05$, not corrected for multiple tests) in 5 of 18 participants (Participants 1, 3, 4, 14, and 16). This result indicates that the classification in these participants could be driven in part by timing differences between conditions. Note, however, that only in three of these participants the classification performance based

**Figure 4.** Results of the analysis of temporal dynamics within the conceptual preparation interval. (A) The temporal development of the feature maps ($p < .05$, uncorr., $n = 18$). The maps were constructed in the same way as for Figure 3, based on the source-level signal in 100-msec bins. (B) The averaged classification accuracy across 18 participants at each time bin. The group-level accuracies in the last two bins (from −200 to −100 msec and from −100 msec to the button press) were significantly above chance level ($p < .001$) after Bonferoni correction for the number of conducted tests.



on the sensor-level signal time series was significantly higher than chance level.

The classification analysis at the baseline interval showed a significant outcome only in 1 of 18 participants (Participant 18). For other participants, the accuracies were close to chance level, and the average accuracy across the group was 0.51 ± 0.01. Overall, based on this result, we conclude that the block structure is not a major confounding factor in the employed experimental design.

### Analysis of Temporal Dynamics

Results of the analysis of temporal dynamics are presented in Figure 4. Panel B shows the averaged classification accuracy across 18 participants at each time bin. The averaged accuracies in the last two bins (from −200 to −100 msec and from −100 msec to the button press) were significantly above chance level, after Bonferoni correction of the statistical threshold. Panel A shows the temporal development of the feature maps. In the left hemisphere, the location of the most important features for the classifier changed from the pFC (left inferior frontal gyrus, left middle frontal gyrus, left precentral gyrus) toward posterior areas (left fusiform, left inferior temporal gyrus, left inferior, and middle occipital gyri). In the right hemisphere, important features shift from parietal cortex to the inferior temporal and occipital locations by the end of the measured interval.

## DISCUSSION

### Single-trial Classification Accuracies

In this study, we explored the possibility of predicting the semantic category of internally generated words from observed brain activity before word onset. The results indicate that such prediction is feasible. Classification accuracies in the time interval before speech production were above chance in a majority of the participants. To our knowledge, this is the first study demonstrating successful semantic decoding in a word generation task. Notably, the range of decoding accuracies obtained in this study is similar to results reported before for decoding semantic information when people are reading or listening to verbal stimuli. For instance, in the study of Simanova et al. (2010), participants were presented with animal and tool words, and the Bayesian logistic regression classifier was trained to predict the word's category from EEG data. Classification in the auditory modality showed a mean accuracy of 0.61 and classification in the orthographic modality showed a mean accuracy of 0.56 (across 20 participants). These numbers are very close to the present results, indicating that the present classification results are close to what may be expected based on previous studies.

Previous studies have shown that the spatial filtering methods, such as ICA, improve classification performance on electrophysiological data (see, e.g., Farquhar & Hill, 2012; Lemm, Blankertz, Curio, & Müller, 2005). A recent MEG study showed an improvement of the classification accuracies in source space compared with sensor space (van de Nieuwenhuijzen et al., 2013). Differences in results between projections are because of changes in signal-to-noise ratio; spatial or temporal filtering suppresses noise from unrelated sources, which would lead to more robust input to the classifier. Here, we observed a slight increase in classification accuracy when applying the ICA decomposition or source-level decomposition to the signal, relative to sensor-level results.

### Feature Localization

Previous studies have shown that Bayesian logistic regression is not only an effective classification technique, but also a useful tool for studying cognitive processes (Simanova et al., 2010; van Gerven et al., 2010). The feature maps produced by the classifier can be related to the discriminability between experimental conditions. Here we conducted a group-level statistical analysis of the source-level feature maps to identify brain locations commonly involved in semantic preparation in the word generation task across all the experimental participants. The effect was statistically significant only when uncorrected for multiple comparisons; the group-level maps should therefore be interpreted as descriptive, rather than inferential. However, because the common structure in the importance maps revealed by the group-level test agreed with previous studies on semantic memory and categorical contrasts (Simanova et al., 2014; Martin, 2007; Gerlach et al., 2002; Caramazza & Shelton, 1998), here we briefly discuss these localization results.

The task used in this study is very similar to verbal fluency, which is widely used in neuropsychology to evaluate language and executive control in patients. Neurological observations and neuroimaging studies indicate that the left frontal lobe region is particularly important for verbal fluency (e.g., Birn et al., 2010; Meinzer et al., 2009; Costafreda et al., 2006; Hodges et al., 1999; Thompson-Schill et al., 1998; Milner, 1964). At the same time, it has been suggested that the network of active brain regions is not identical for two commonly used types of fluency tasks. A number of neuroimaging studies reported that letter fluency (generation of words for a letter cue) yields activation in the left inferior frontal gyrus, whereas semantic fluency (generation of words for a semantic category cue) yields greater activation in the left middle frontal gyrus (Birn et al., 2010; Meinzer et al., 2009; Perani et al., 2003). Both fluency tasks were combined in the current study, and both frontal regions emerged at the group-level analysis. The biggest cluster with the high averaged importance values was found in the left middle frontal gyrus and thus coincides with the reported activations for semantic fluency.

A second prominent contribution of importance values in posterior middle temporal gyrus during the conceptual preparation also agrees with previous studies (see, e.g., Simanova et al., 2014; Chao, Weisberg, & Martin, 2002; Maess, Friederici, Damian, Meyer, & Levelt, 2002; Chao et al., 1999; see also reviews by Martin, 2007; Martin & Chao, 2001). Posterior middle temporal region plays a central role in retrieval of semantic knowledge (Binder & Desai, 2011; Binder, Desai, Graves, & Conant, 2009). Previous studies using the same semantic contrast showed that this region is involved in discrimination between semantic categories independently of the sensory modality used for stimulus presentation (Simanova et al., 2014; Fairhall & Caramazza, 2013).

We also studied the temporal dynamics of the classification accuracies and feature importance maps in the interval of conceptual preparation (Figure 4). Results suggest a gradual transition of the importance weights in the left hemisphere from prefrontal to posterior cortex over the course of 500 msec. This may indicate that semantic retrieval in speech production is initiated in pFC and posterior areas come into play later. Classification accuracies increased over the 500-msec interval, which suggests that category-related information is accumulating in the signal toward the end of conceptual preparation.

The reported results are very similar to our previous findings on decoding of semantic information; yet, there is an important difference. The current study investigates word production in the absence of any perceptual cues. With this, this study provides evidence that conceptual processing in speech production and perception relies on a common functional substrate. The source space localization results replicate previous findings, suggesting an important role of left inferior and middle frontal gyri in conceptual preparation and word retrieval processes (see review by Price, 2010). Extraction of conceptual information is underlined by dynamic interplay between these frontal regions and posterior temporal cortex (see Hagoort, 2013; Binder & Desai, 2011; Price, 2010). Present results suggest that, in case of internally guided word production, the initial category-specific activity in frontal cortex is followed by a build-up of categorical information in temporal areas.

### Limitations of the Present Study

An important downside of the volitional word generation task compared with standard stimulus-driven experimental paradigms is the impossibility to control beforehand the characteristics of responses, such as word frequency, imageability, length, syllable structure, et cetera. It is therefore important to collect behavioral measures and take into consideration possible unwanted differences between experimental conditions. In this study, differences in RTs were revealed. The response for nonliving objects took longer than for animals in many participants. The difference in timing could have had an impact on classification accuracy. The conducted analysis of confounding factors shows, however, that the classifier's predictions for single trials did not necessarily correlate with RTs. This result provides further support for the validity of our findings.

Differences in employed recall strategy and task difficulty between the categories may also act as confounding factors. The current results on localization of discriminating features should therefore be interpreted with caution.

We believe that more research is needed on semantic effects in internally generated word production, and possibly a more optimal experimental paradigm could be established in the future. For instance, various studies have recorded EEG during overt, not delayed, speech production, and the methods of correction for speech-related artifacts have improved in the last years (Aristei, Melinger, & Abdel Rahman, 2011). Hence, taking into account the present results, it might be possible to apply single-trial classification techniques for semantic decoding from EEG/MEG signal in overt word production.

To summarize, this study shows that semantic information can be decoded from neuromagnetic recordings during internally generated word production, in the absence of any perceptual cues. The results give an important insight into the temporal dynamics of brain activity underlying volitional word generation and as such about central aspects of language production.

### Acknowledgments

We would like to thank Jan Mathijs Schoffelen for helpful discussions on data analysis, and Dimitra Makri for assistance in data acquisition. This work was supported by the BrainGain Smart Mix Programme of the Netherlands Ministry of Economic Affairs and the Netherlands Ministry of Education, Culture and Science.

Reprint requests should be sent to Irina Simanova, Donders Institute for Brain, Cognition and Behaviour, Radboud University Nijmegen, Montessorrilaan 3, 6525 HR, Nijmegen, Netherlands, or via e-mail: i.simanova@donders.ru.nl.

## REFERENCES

Aristei, S., Melinger, A., & Abdel Rahman, R. (2011). Electrophysiological chronometry of semantic context effects in language production. *Journal of Cognitive Neuroscience, 23,* 1567–1586.

Binder, J. R., & Desai, R. H. (2011). The neurobiology of semantic memory. *Trends in Cognitive Sciences, 15,* 527–536.

Binder, J. R., Desai, R. H., Graves, W. W., & Conant, L. L. (2009). Where is the semantic system? A critical review and meta-analysis of 120 functional neuroimaging studies. *Cerebral Cortex, 19,* 2767–2796.

Birn, R. M., Kenworthy, L., Case, L., Caravella, R., Jones, T. B., Bandettini, P. A., et al. (2010). Neural systems supporting lexical search guided by letter and semantic category cues: A self-paced overt response fMRI study of verbal fluency. *Neuroimage, 49,* 1099–1107.

Caramazza, A., & Shelton, J. R. (1998). Domain-specific knowledge systems in the brain: The animate-inanimate distinction. *Journal of Cognitive Neuroscience, 10,* 1–34.

Chan, A. M., Halgren, E., Marinkovic, K., & Cash, S. S. (2011). Decoding word and category-specific spatiotemporal representations from MEG and EEG. *Neuroimage, 54,* 3028–3039.

Chao, L. L., & Martin, A. (2000). Representation of manipulable man-made objects in the dorsal stream. *Neuroimage, 12,* 478–484.

Chao, L. L., Haxby, J. V., & Martin, A. (1999). Attribute-based neural substrates in temporal cortex for perceiving and knowing about objects. *Nature Neuroscience, 2,* 913–919.

Chao, L. L., Weisberg, J., & Martin, A. (2002). Experience-dependent modulation of category-related cortical activity. *Cerebral Cortex, 12,* 545–551.

Costafreda, S. G., Fu, C. H., Lee, L., Everitt, B., Brammer, M. J., & David, A. S. (2006). A systematic review and quantitative appraisal of fMRI studies of verbal fluency: Role of the left inferior frontal gyrus. *Human Brain Mapping, 27,* 799–810.

Cox, D. D., & Savoy, R. L. (2003). Functional magnetic resonance imaging (fMRI) "brain reading": Detecting and classifying distributed patterns of fMRI activity in human visual cortex. *Neuroimage, 19,* 261–270.

Dale, A. M., Liu, A. K., Fischl, B. R., Buckner, R. L., Belliveau, J. W., Lewine, J. D., et al. (2000). Dynamic statistical parametric mapping: Combining fMRI and MEG for high-resolution imaging of cortical activity. *Neuron, 26,* 55–67.

Fairhall, S. L., & Caramazza, A. (2013). Brain regions that represent amodal conceptual knowledge. *Journal of Neuroscience, 33,* 10552–10558.

Farquhar, J., & Hill, N. J. (2012). Interactions between pre-processing and classification methods for event-related-potential classification: Best-practice guidelines for brain-computer interfacing. *Neuroinformatics, 11,* 175–192.

Gerlach, C. (2007). A review of functional imaging studies on category specificity. *Journal of Cognitive Neuroscience, 19,* 296–314.

Gerlach, C., Law, I., & Paulson, O. B. (2002). When action turns into words. Activation of motor-based knowledge during categorization of manipulable objects. *Journal of Cognitive Neuroscience, 14,* 1230–1239.

Hagoort, P. (2013). MUC (memory, unification, control) and beyond. *Frontiers in Psychology, 4,* 416.

Hämäläinen, M. S., & Ilmoniemi, R. J. (1994). Interpreting magnetic fields of the brain: Minimum norm estimates. *Medical & Biological Engineering & Computing, 32,* 35–42.

Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., & Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science, 293,* 2425–2430.

Haynes, J. D., & Rees, G. (2006). Decoding mental states from brain activity in humans. *Nature Reviews Neuroscience, 7,* 523–534.

Hodges, J. R., Patterson, K., Ward, R., Garrard, P., Bak, T., Perry, R., et al. (1999). The differentiation of semantic dementia and frontal lobe dementia (temporal and frontal variants of frontotemporal dementia) from early Alzheimer's disease: A comparative neuropsychological study. *Neuropsychology, 13,* 31–40.

Hwang, K., Palmer, E. D., Basho, S., Zadra, J. R., & Müller, R. A. (2009). Category-specific activations during word generation reflect experiential sensorimotor modalities. *Neuroimage, 48,* 717–725.

Indefrey, P. (2011). The spatial and temporal signatures of word production components: A critical update. *Frontiers in Psychology, 2,* 255.

Indefrey, P., & Levelt, W. J. (2004). The spatial and temporal signatures of word production components. *Cognition, 92,* 101–144.

Kamitani, Y., & Tong, F. (2005). Decoding the visual and subjective contents of the human brain. *Nature Neuroscience, 8,* 679–685.

Lemm, S., Blankertz, B., Curio, G., & Müller, K. R. (2005). Spatio-spectral filters for improving the classification of single trial EEG. *IEEE Transactions on Biomedical Engineering, 52,* 1541–1548.

Levelt, W. J., Praamstra, P., Meyer, A. S., Helenius, P., & Salmelin, R. (1998). An MEG study of picture naming. *Journal of Cognitive Neuroscience, 10,* 553–567.

Maess, B., Friederici, A. D., Damian, M., Meyer, A. S., & Levelt, W. J. (2002). Semantic category interference in overt picture naming: Sharpening current density localization by PCA. *Journal of Cognitive Neuroscience, 14,* 455–462.

Mahon, B. Z., & Caramazza, A. (2009). Concepts and categories: A cognitive neuropsychological perspective. *Annual Review of Psychology, 60,* 27–51.

Makeig, S., Bell, A. J., Jung, T. P., & Sejnowski, T. J. (1996). Independent component analysis of electroencephalographic data. In D. S. Touretzky, M. C. Mozer, & M. E. Hasselmo (Eds.), *Advances in neural information processing systems* (pp. 145–151). MIT Press.

Mandler, J. M. (2004). Thought before language. *Trends in Cognitive Sciences, 8,* 508–513.

Martin, A. (2007). The representation of object concepts in the brain. *Annual Review of Psychology, 58,* 25–45.

Martin, A., & Chao, L. L. (2001). Semantic memory and the brain: Structure and processes. *Current Opinion in Neurobiology, 11,* 194–201.

Martin, A., Wiggs, C. L., Ungerleider, L. G., & Haxby, J. V. (1996). Neural correlates of category-specific knowledge. *Nature, 379,* 649–652.

Meinzer, M., Flaisch, T., Wilser, L., Eulitz, C., Rockstroh, B., Conway, T., et al. (2009). Neural signatures of semantic and phonemic fluency in young and old adults. *Journal of Cognitive Neuroscience, 21,* 2007–2018.

Milner, B. (1964). Some effects of frontal lobectomy in man. In J. Warren & K. Akert (Eds.), *The frontal granular cortex and behavior* (pp. 313–331). New-York: McGraw-Hill.

Murphy, B., Poesio, M., Bovolo, F., Bruzzone, L., Dalponte, M., & Lakany, H. (2011). EEG decoding of semantic category reveals distributed representations for single concepts. *Brain and Language, 117,* 12–22.

Nolte, G. (2003). The magnetic lead field theorem in the quasi-static approximation and its use for magnetoencephalography forward calculation in realistic volume conductors. *Physics in Medicine & Biology, 48,* 3637–3652.

Oostenveld, R., Fries, P., Maris, E., & Schoffelen, J. M. (2011). FieldTrip: Open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Computational Intelligence and Neuroscience, 2011,* 156869.

Perani, D., Abutalebi, J., Paulesu, E., Brambati, S., Scifo, P., Cappa, S. F., et al. (2003). The role of age of acquisition and language usage in early, high-proficient bilinguals: An fMRI study during verbal fluency. *Human Brain Mapping, 19,* 170–182.

Perani, D., Cappa, S. F., Bettinardi, V., Bressi, S., Gorno-Tempini, M., Matarrese, M., et al. (1995). Different neural systems for the recognition of animals and man-made tools. *NeuroReport, 6,* 1637–1641.

Price, C. J. (2010). The anatomy of language: A revirew of 100 fMRI studies published in 2009. *Annals of the New York Academy of Sciences , 1191,* 62–88.

Reddy, L., & Kanwisher, N. (2007). Category selectivity in the ventral visual pathway confers robustness to clutter and diverted attention. *Current Biology, 17,* 2067–2072.

Salzberg, S. L. (1997). On comparing classifiers: Pitfalls to avoid and a recommended approach. *Data Mining and Knowledge Discovery, 1,* 317–328.

Shinkareva, S. V., Malave, V. L., Mason, R. A., Mitchell, T. M., & Just, M. A. (2011). Commonality of neural representations of words and pictures. *Neuroimage, 54,* 2418–2425.

Simanova, I., Hagoort, P., Oostenveld, R., & van Gerven, M. A. (2014). Modality-independent decoding of semantic information from the human brain. *Cerebral Cortex, 24,* 426–434.

Simanova, I., van Gerven, M., Oostenveld, R., & Hagoort, P. (2010). Identifying object categories from event-related EEG: Toward decoding of conceptual representations. *PLoS One, 5,* e14465.

Thompson-Schill, S. L., Swick, D., Farah, M. J., D'Esposito, M., Kan, I. P., & Knight, R. T. (1998). Verb generation in patients with focal frontal lesions: A neuropsychological test of neuroimaging findings. *Proceedings of the National Academy of Sciences, U.S.A., 95,* 15855–15860.

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., et al. (2002). Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *Neuroimage, 15,* 273–289.

van de Nieuwenhuijzen, M. E., Backus, A. R., Bahramisharif, A., Doeller, C. F., Jensen, O., & van Gerven, M. A. (2013). MEG-based decoding of the spatiotemporal dynamics of visual category perception. *Neuroimage, 83,* 1063–1073.

van Gerven, M. A., Cseke, B., de Lange, F. P., & Heskes, T. (2010). Efficient Bayesian multivariate fMRI analysis using a sparsifying spatio-temporal prior. *Neuroimage, 50,* 150–161.

van Gerven, M., Farquhar, J., Schaefer, R., Vlek, R., Geuze, J., Nijholt, A., et al. (2009). The brain–computer interface cycle. *Journal of Neural Engineering, 6,* 041001.

Vindiola, M., & Wolmetz, M. (2011). Mental encoding and neural decoding of abstract cognitive categories: A commentary and simulation. *Neuroimage, 54,* 2822–2827.

Vitali, P., Abutalebi, J., Tettamanti, M., Rowe, J., Scifo, P., Fazio, F., et al. (2005). Generating animal and tool names: An fMRI study of effective connectivity. *Brain and Language, 93,* 32–45.