Distinguishing Concept Categories from Single-Trial Electrophysiological Activity

Brian Murphy (brian.murphy@unitn.it)
Centre for Mind/Brain Sciences, Univ. of Trento, Italy

Michele Dalponte (michele.dalponte@unitn.it)
Dept. of Inf. Science and Engineering, Univ. Trento, Italy

Massimo Poesio (massimo.poesio@unitn.it)
Centre for Mind/Brain Sciences, Univ. of Trento, Italy /
Dept. Of Computing and Electronic Systems, Univ. Essex, United Kingdom

Lorenzo Bruzzone (lorenzo.bruzzone@unitn.it)
Dept. of Inf. Science and Engineering, Univ. Trento, Italy

Abstract

Differences in the neural representation of conceptual categories (such as buildings, tools, animals, plants, vehicles) are suggested by studies on brain-injury patients and by those using imaging techniques. Evidence that such conceptual distinctions are encoded spatially has been highlighted by fMRI techniques (Haxby et al, 2001), whereas ERP studies have identified differences in the temporal domain (Kiefer, 2001; Paz-Caballero et al., 2006). The study described here uses a machine learning technique (Dalponte et al., 2007), previously applied to Brain-Computer Interaction (BCI) tasks, to show that conceptual categories can be identified from single-trial spectral EEG responses to visually and auditorily presented stimuli in single participants. We found that using features extracted from frequency spectra, the categorial membership of a single stimulus presentation in one of three classes (animals, plants, tools) can be predicted with an average accuracy of 80%.

Keywords: Conceptual Categories, Semantic Spaces, Machine Learning, EEG.

Introduction

Data about brain activity are providing exciting new insights on conceptual knowledge. Studies such as Martin et al (1996) provided evidence of topographical encoding of conceptual information through fMRI; machine learning techniques have been successfully used to classify brain activity recorded with such techniques into conceptual categories (Haxby et al 2001, Hanson et al 2004, Shinkareva et al 2008).

All such studies rely on spatial information about neural patterns collected through fMRI. However, theories of conceptual memory based on the hypothesis that conceptual knowledge is distributed, taking the form of a 'conceptual map' (Spitzer, 1999) or 'word web' (Pulvermüller 2002), would predict a temporal or frequency encoding for conceptual information as well, involving some form of synchronization between the distinct brain areas activated in response to a concept. And indeed, several ERP studies reported by Pulvermüller (2002) found evidence for a temporal encoding of conceptual knowledge—e.g., for distinctions in the temporal domain between action verbs and nouns. To our knowledge, however, no previous study has attempted to identify regularities in conceptual knowledge activation on the basis of frequency information gathered through EEG. Yet if this is the case, and if similar categorical distinctions can be made with these techniques, our investigations of conceptual knowledge will greatly benefit, as the much lower overhead of EEG studies would make larger-scale investigations possible, particularly explorations of conceptual knowledge not directly concerned with issues of topography. The temporal resolution of EEG, relative to fMRI, also makes it possible to disentangle the effects of different stages of the process of perception and categorization.

In this paper we report the results of just such a study. Spectral EEG data collected from healthy participants presented with visual and audio representations of concepts belonging to three categories—animals, tools, and plants—were classified using supervised machine learning methods developed for BCI applications (which are typically applied to lower level cognitive states such as imagined movements). To examine the "steady-state" conceptual representations that we assume are the end result of perceptual processes, only the period *after* stimulus offset was considered. We found that these methods are able to classify neural patterns with excellent accuracy—i.e., that enough information is encoded in frequency spectra to allow discrimination between conceptual categories.

This paper is organized into six sections. After this introduction a background section presents the state-of-the-art and theoretical motivations. In the third section the experimental methodology is presented, and after that the main concepts of the data analysis are explained. Experimental results are presented in the fifth section. The last section is devoted to the discussion of the results.

Background

The earliest evidence for category-specific brain activation was provided by studies on patients with brain injuries (Warrington & Shallice, 1984; Caramazza & Shelton, 1998). Subsequent evidence was provided by functional brain imaging (Martin et al., 1996). For example, Haxby et al (2001) demonstrated with an fMRI study that the activation associated with semantic categories was distributed in the brain, and that areas outside of the principal foci of activity that are examined in neuropsychological studies could be used to successfully predict category membership in single participants.

Such studies also provided evidence that the location of these activations is consistent across individuals and across tasks, and that multiple areas of the brain are activated for each conceptual category (Martin & Chao, 2001; Pulvermüller, 2002). These results led to theoretical proposals such as Pulvermüller's 'word web' idea (2002), according to which concepts are represented in 'neuron webs' with distinct cortical topographies. This suggests the need for some form of synchronization between the firing of these neurons, either in the form of time synchronization or of frequency synchronizations. ERP studies such as Kiefer (2001), and later Paz-Caballero and colleagues (2006) found category specific differences in event-related potentials across groups of stimuli and participants.[1]

Experimental Design

Participants

The data were collected at the University of Essex BCI Lab. Four monolingual native speakers of English participated in the study. All four were male, with a college education, and reported that they did not suffer from any psychological or neurological condition. Their mean age was 26 years. One participant was left-handed. Participants were paid compensation of £6 per hour.

Procedure

Participants were presented with both a word visualization task (**auditory stimulus task**) and a silent image naming task (**visual stimulus task**). The auditory and visual stimuli were presented in two separate blocks, the order of which was alternated across participants. In each task the same full set of concepts was used, and their order was randomized on every run.

In the visual stimulus task participants sat in a relaxed upright position 1m from a computer monitor. Images were presented on a medium grey background and fell within a 9° viewing angle. Each image was preceded by 0.5s fixation cross, and followed by 2s of fixation and 2s of a blank screen. Participants were instructed to silently name the object represented, and to press the keyboard space-bar with the left-hand to indicate they had found an appropriate word. The image disappeared from the screen on this response, or on a time-out of four seconds.

In the audio stimulus task the same pattern of fixations and blanks was used while participants listened to words through a pair of earphones. In this case they were instructed to visualize an image that represented the word heard. Again, a keyboard response was used by participants to indicate that an image had been found.

After both tasks had been completed, each set of stimuli were again presented to verify that they had been correctly interpreted by participants.

Figure 1: Examples of visual stimuli.

Stimuli

The same set of 127 concepts were used as stimuli in both the visual and the auditory task. The stimuli for the visual task (Figure 1) were colourised line drawings from a replication of the Snodgrass object image set (Rossion & Pourtois, 2004). The audio stimuli were spoken words, recorded in-house by the experimenters. Stimulus concepts came from three categories: **animals** (50), small manipulable functional artefacts ('**tools**'; 50) and **plants** (27). Concepts were chosen to range from typical, familiar and frequent members of their category (e.g. *dog, hammer, flower*) to more obscure exemplars (e.g. *sea horse, thimble, artichoke*). The set of stimulus concepts was not manipulated to result in equal group averages for typicality, familiarity and similar norms, since group analyses of the categories were not planned (such group analyses are vulnerable to the effects of such confounding factors).

Recording

Variations in scalp voltages were measured at 64 electrode positions on a standard 10-20 montage, using a BioSemi ActiveTwo active electrode system.[2] An additional six electrodes measured voltages at the ear lobes, and around the eyes, for signal referencing and artifact identification purposes. Electrode activity was recorded on a dedicated PC at 512Hz.

[1] The only electrophysiological study known to the authors to find single participant category effects is Tanji et al. (2005), which used intracranial electrodes.

[2] http://www.biosemi.com/products.htm

Data analysis

Data Preprocessing

Before analysis all data was referenced to the average of the ear lobe electrodes, was resampled to 150Hz and was filtered with pass band of 1-60Hz. The EEGLAB suite[3] was used to perform an independent component analysis of the data, and components due to eye movements and to electrical mains-noise were manually identified and removed. No removal of epochs due to muscle or head-movement artefacts was necessary.

Automatic Categorization Procedure

The scheme of the automatic procedure adopted for categorizing the EEG signals is shown in Figure 2. It is made up of different blocks: i) an automatic system for the selection of optimal time and frequency intervals; ii) a feature extraction module; and iii) a categorization module. In the following some details of each block are given.

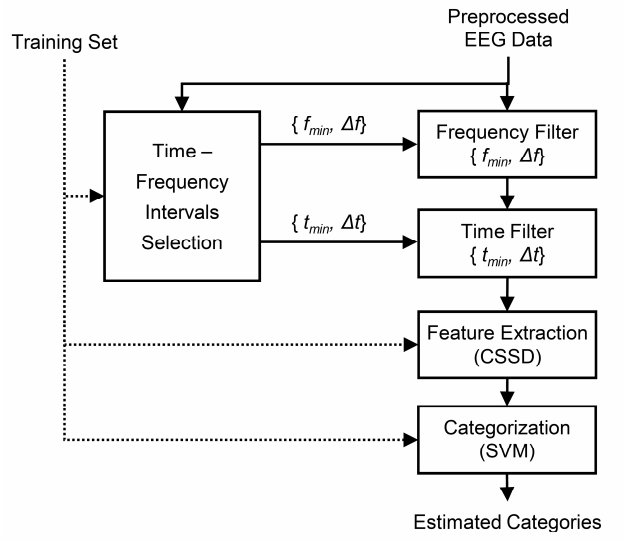

Figure 2: General architecture of the system used for categorization.

Time and Frequency Interval Selection

The preprocessed EEG signal is given as input to a system for the automatic identification of the most informative time and frequency intervals, which was developed for BCI applications (Dalponte et al., 2007), achieving state-of-the-art performance on a competition task involving imagined hand and foot movements.

The goal of this technique is to define the best combination of time and frequency intervals for the separation of the analyzed categories. This method can be divided into two parts: i) a search strategy, and ii) a separability measure computation.

[3] http://www.sccn.ucsd.edu/eeglab/

The search strategy adopted is based on a hierarchical search approach. Several combinations of time and frequency filter parameters are iteratively tested, changing the width (Δf and Δt) and position (t_{min} and f_{min}) of the time and frequency intervals considered.

The range of frequencies tested in this work is 1 to 50 Hz, and the time ranges from the stimulus offset to 5 seconds after that point.

For each combination of filter parameters the input data were filtered in frequency and in time. From filtered data features were extracted and separability among categories computed according to the Jeffries-Matusita (*JM*) distance (Bruzzone et al., 1995). This is a widely used measure in pattern recognition problems to estimate separability between classes, given a set of features. The *JM* distance between the categories ω_i and ω_j is defined as follows:

$$JM_{ij}(\boldsymbol{P}) = \left\{ \int_x \left[\sqrt{p(\boldsymbol{x}/\omega_i)} - \sqrt{p(\boldsymbol{x}/\omega_j)} \right]^2 d\boldsymbol{x} \right\}^{1/2} \quad (1)$$

where $p(\boldsymbol{x}/\omega_i)$ and $p(\boldsymbol{x}/\omega_j)$ are the conditional probability density functions for the feature vector $\boldsymbol{x}$ given the categories ω_i and ω_j. For a complete description of the time and frequency intervals selection method the reader is referred to Dalponte et al. (2007).

Once the most informative time and frequency intervals have been identified, these optimal parameters are used for all further analysis. In the predictive analysis the parameters are derived from a combination of parameters from a 3-fold cross-validation.

Feature Extraction

The feature extraction phase is performed on temporal and frequency filtered data. The algorithm adopted is based on the Common Spatial Subspace Decomposition (CSSD) algorithm (Wang et al., 1999). CSSD is a supervised transformation that decomposes the original EEG channels into a new time series which shows optimal variances for the discrimination of two populations of EEG signals. In particular a spatial filter is designed, and applied to the original data:

$$\boldsymbol{X} = \boldsymbol{SF} \cdot \boldsymbol{E} \quad (2)$$

where $\boldsymbol{E}$ is the matrix NxT of the original EEG data (where N is the number of channels and T the number of samples per channel), $\boldsymbol{SF}$ is the spatial filter and $\boldsymbol{X}$ is the set of derived signal components. The spatial filter $\boldsymbol{SF}$ is obtained by the simultaneous diagonalization of two covariance matrices, derived from the training data of the two categories considered.

From the new time series we extract as features for the categorization module the normalized variance of the first and last component of the transformed matrix $\boldsymbol{E}$, which are the more representative components of the two considered categories. Thus the two features for each epoch are the level of signal activity in each component. This can be viewed as comparing the event-related spectral activity (i.e. the relative event-related desynchronisation) of two synchronous neural structures which have been found to

have an optimal differential response to the stimulus categories of interest.

Categorization

The categorization step is based on a Support Vector Machines (SVM) classifier (Vapnik, 1998). This classifier is an advanced pattern recognition technique that has been adopted in many different fields in recent years. Three reasons for its success are: i) its high classification accuracy and very good generalization capability with respect to other classifiers; ii) the limited effort required in architecture design (i.e., it involves few control parameters) and low processing time overhead in the learning phase compared to other algorithms; and iii) its effectiveness in 'ill-posed' classification problems (those which have a low ratio between number of training samples and number of features).

The rationale of this classifier is to transform the original feature space into a space with a higher dimensionality, where a separation between the two categories by means of an optimal hyperplane is searched for, defined as:

$$f(\mathbf{x}) = w^* \cdot \Phi(\mathbf{x}) + b^* \qquad (3)$$

where $\boldsymbol{w}^*$ is a vector orthogonal to the separating hyperplane, b^* is a scalar value such that the ratio $b^*/||\boldsymbol{w}^*||$ represents the distance of the hyperplane from the origin, and the function Φ represents a non-linear transformation, called a 'kernel function'. The optimal hyperplane is the one that minimizes a cost function which combines two criteria; maximization of the inter-class margin, and minimization of classification errors.

The model selection of the parameters of the classifier was carried out according to the following strategy: i) randomly subdivide available data into two folds, containing respectively 20% and 80% samples; ii) train the classifier using the 20% fold, and test it with the 80% fold; iii) iterate steps i) and ii) 100 times; and iv) compute final accuracy as the mean accuracy over the 100 trials.

Results

Exploratory Analysis

In this first analysis we trained of the feature extraction algorithm with the complete set of patterns available. This means that the results obtained in this experiment are the optimal ones obtainable on this data set. We perform such an analysis in order to define an upper bound in terms of accuracies on these data sets, and to verify that category specific patterns are present in the recorded EEG activity.

The analysis was carried out separately for each combination of participant and task modality. In each of these, three pairwise categorizations were carried out: animals vs. tools; tools vs. plants; and plants vs. animals. The optimization procedure discovered optimal time-frequency windows that varied, but they lay predominantly in the 15-35Hz bands, and at 1.5 to 3 seconds after stimulus offset.

In all 24 analyses (4 participants x 2 modalities x 3 category pairs) the optimal components yielded described a semantic space in which conceptual categories formed largely separable clusters (see Figure 3 for an example). The power of categorization of the support vector machine was correspondingly high, achieving between 97% and 100% accuracy across all tests (see Table 1 for the complete results).

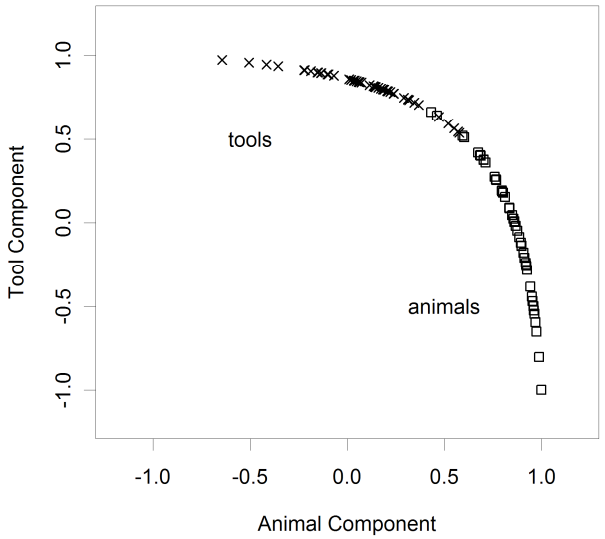

Figure 3: Semantic Space from Exploratory Analysis (Participant D, Auditory Stimulus, Animals vs. Tools). Animal concepts are represented by squares and Tools by crosses.

Table 1: Performance of Exploratory Analysis.

	Participant			
	A	B	C	D
Animals vs. Tools Auditory Task	98.6%	99.4%	97.6%	97.2%
Animals vs. Tools Visual Task	99.1%	96.9%	98.8%	98.8%
Animals vs. Plants Auditory Task	100%	100%	100%	99.8%
Animals vs. Plants Visual Task	100%	99.2%	98.9%	100%
Tools vs. Plants Auditory Task	100%	100%	100%	100%
Tools vs. Plants Visual Task	100%	100%	99.8%	100%

Predictive Analysis

The second analysis aims to understand the predictive capability of the system. In this context we carried out a 3-fold cross validation in the categorization step. Samples were divided into 3 arbitrary partitions or 'folds', two of which are iteratively used to train the feature extraction algorithm, while the remaining fold is used to compute class separability. Thus, we obtain three distances for each combination, that are averaged in order to have a single distance value. This procedure allows a predictive analysis, as the epochs used to train the feature extraction algorithm do not overlap with those used to compute the distance measure.

This analysis was carried out for all participants and both modalities on the category pair of Animals vs. Plants. In this task the optimization procedure discovered somewhat broader time-frequency windows, ranging from 10-45Hz, and from 1 to 4 seconds after stimulus offset.

The resulting semantic spaces show clear category clusters, with a more extensive overlap than in the previous analysis (see e.g. Figure 4). Categorization accuracy remained high, averaging 80% (see Table 2).

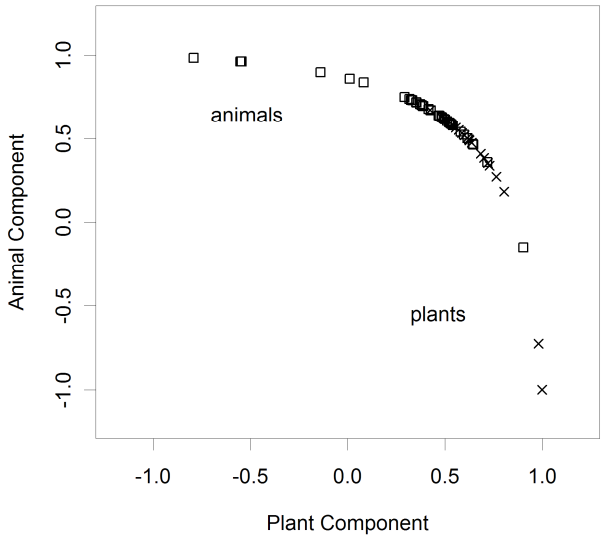

Figure 4: Semantic Space from Predictive Experiment (Participant C, Visual Stimulus, Animals vs. Plants). Animal concepts are represented by squares and Plants by crosses.

Table 2: Predictive Accuracy of Animals vs. Plants

	Visual Task	Auditory Task
Participant A	74.6%	88.2%
Participant B	72.4%	65.4%
Participant C	82.6%	92.7%
Participant D	81.9%	77.7%

Discussion

To our knowledge, this is the first study reporting that conceptual categories are distinguishable on the basis of neural activation data collected using EEG techniques. Our results suggest that it may be possible to investigate conceptual representations in the brain using techniques with much lower overhead than fMRI, which would be a great advantage for researchers—e.g., psycholinguists, computational linguists—who are primarily concerned with the organization of conceptual knowledge, rather than with its neural correlates. Of course, further work is needed to verify that information collected with such methods is also consistent across individuals as demonstrated for fMRI data e.g., by Shinkareva et al (2008). While differences were found, it is not yet clear if this is due to differences across participants in the timing or frequency encoding of cognitive states, or rather that these parameters identify parts of the time course and frequency spectrum that are less subject to task-related noise. Our results suggest that although optimal windows vary, good classification results can also be achieved with uniform windows. Further, while parts of frequency spectra found have been linked to object perception and representation (see e.g. Tallon-Baudry & Bertrand, 1999), activity in these bands has been found to be modulated by a wide variety of cognitive tasks in verbal, visual and spatial processing (see Kahana 2006 for a review).

Results concerning the scalp localization of the category-specific components provide further evidence for a distributed representation of conceptual knowledge, consistent with theories such as that of Barsalou (2003), as parietal, temporal and ventral areas of both hemispheres were seen to contribute to concept identification. However, it would be premature to interpret our results as providing evidence concerning localization of conceptual knowledge.

While BCI methodologies have been successfully applied to the decoding of task related cognitive states (such as imagined motor movements), this work is the first to study conceptual organization; and our results demonstrate the considerable power of the techniques described in Dalponte et al (2007).

Further work will also be required to investigate the correlation between the evidence of conceptual spaces obtained with our methods and the evidence provided by fMRI studies; spaces derived from corpus data (e.g. Baroni & Lenci, 2008); and other informant supplied data such as typicality judgements, semantic similarity judgements, and feature norms (McRae et al, 2005). We also continue to investigate the extent to which the predictive power of the system can generalize across participants and across modalities.

Acknowledgments

We wish to thank our colleague Heba Lakany for much of the experimental design, and Marco Baroni for insightful comments. We also wish to thank the University of Essex's BCI group, in whose facilities the experiments were

conducted. This research was supported in part by a CIMEC postdoctoral fellowship, and in part by a Research Promotion Fund grant from the University of Essex.

References

Baroni, M. & Lenci, A. (2008). Concepts and properties in word spaces. In A. Lenci (Ed.), *From Context to Meaning: Distributional Models of the Lexicon in Linguistics and Cognitive Science* (special issue of the Italian Journal of Linguistics).

Barsalou, L. (2003). Situated simulation in the human conceptual system. *Language and Cognitive Processes, 18*, 513-562.

Bruzzone, L., Roli, F., & Serpico, S. B. (1995). An extension to multiclass cases of the Jeffries-Matusita distance. *IEEE Transactions on. Geoscience and Remote Sensing, 33*, 1318-1321.

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

Dalponte, M., Bovolo, F., & Bruzzone, L. (2007). Automatic selection of frequency and time intervals for classification of EEG signals. *Electronics Letters, 43*, 1406-1408.

Hanson S. J., Matsuka T., Haxby J. (2004). Combinatorial codes in ventral temporal lobe for object recognition. *Neuroimage 23*, 156-166.

Haxby, J., Gobbini, I., Furey, M., Ishai, A., Schouten, J. & Pietrini, P. (2001). Distributed and Overlapping Representations of Faces and Objects in Ventral Temporal Cortex. *Science, 293*, 2425-2430.

Kahana, M. J. (2006). The cognitive correlates of human brain oscillations. *The Journal of Neuroscience, 26*, 1669-1672.

Kiefer, M. (2001). Perceptual and semantic sources of category-specific effects in object categorization: Event-related potentials during picture and word categorization. *Memory & Cognition, 29*, 100-116.

Martin, A. & Chao, L. (2001). Semantic memory and the brain: structure and processes. *Current Opinions 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.

McRae, K., Cree, G. S., Seidenberg, M. S., & McNorgan, C. (2005). Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods, 37*, 547-559.

Paz-Caballero, D., Cuetos, F., & Dobarro, A. (2006). Electrophysiological evidence for a natural/artifactual dissociation. *Brain Research, 1067*, 189-200.

Pulvermüller, F. (2002). *The neuroscience of language*. Cambridge: Cambridge University Press.

Rossion, B. & Pourtois, G. (2004). Revisiting Snodgrass and Vanderwart's object databank: the role of surface detail in basic level object recognition. *Perception, 33*, 217-236.

Shinkareva, S. V., Mason, R. A., Malave, V. L., Wang, W., Mitchell, T. M., et al. (2008). Using fMRI Brain Activation to Identify Cognitive States Associated with Perception of Tools and Dwellings. *PLoS ONE 3*.

Spitzer, M. (1999). *The mind within the net: models of learning, thinking and acting*. Cambridge: MIT Press.

Tallon-Baudry, C. & Bertrand, O. (1999). Oscillatory gamma activity in humans and its role in object representation. *Trends in Cognitive Science, 3*, 151-162.

Tanji, K., Suzuki, K., Delorme, A., Shamoto, H., & Nakasato, N. (2005). High-frequency gamma-band activity in the basal temporal cortex during picture-naming and lexical-decision tasks. *Journal of Neuroscience, 25*, 3287-3293.

Vapnik, V. N. (1998). *Statistical learning theory*. New York: Wiley.

Wang, Y., Berg, P., & Scherg, M. (1999). Common spatial subspace decomposition applied to analysis of brain responses under multiple task conditions: a simulation study. *Clinical Neurophysiology, 110*, 604-614.

Warrington, E. K., & Shallice, T. (1984). Category specific semantic impairments. *Brain, 107*, 829-853.