EEG responds to conceptual stimuli and corpus semantics

Brian Murphy
CIMeC, University of Trento
Rovereto 38068, Italy
brian.murphy@unitn.it

Marco Baroni
CIMeC, University of Trento
Rovereto 38068, Italy
marco.baroni@unitn.it

Massimo Poesio
CIMeC, University of Trento
Rovereto 38068, Italy
massimo.poesio@unitn.it

Abstract

Mitchell et al. (2008) demonstrated that corpus-extracted models of semantic knowledge can predict neural activation patterns recorded using fMRI. This could be a very powerful technique for evaluating conceptual models extracted from corpora; however, fMRI is expensive and imposes strong constraints on data collection. Following on experiments that demonstrated that EEG activation patterns encode enough information to discriminate broad conceptual categories, we show that corpus-based semantic representations can predict EEG activation patterns with significant accuracy, and we evaluate the relative performance of different corpus-models on this task.

1 Introduction

Models of semantic relatedness induced from corpus data have proven effective in a number of empirical tasks (Sahlgren, 2006) and there is increasing interest in whether distributional information extracted from corpora correlates with aspects of speakers' semantic knowledge: see Lund and Burgess (1996), Landauer and Dumais (1997), Almuhareb (2006), Padó and Lapata (2007), Schulte im Walde (2008), among many others. For this purpose, corpus models have been tested on datasets that are based on semantic judgements (metalinguistic or meta-cognitive intuitions about synonymy, semantic distance, category-membership) or behavioural experiments (semantic priming, property generation, free association). While all these data are valuable, they are indirect reflections of semantic knowledge, and when the predictions they make diverge from those of corpora, interpretation is problematic: is the corpus model missing essential aspects of semantics, or are non-semantic factors biasing the data elicited from informants?

Reading semantic processes and representations directly from the brain would be an ideal way to get around these limitations. Until recently, analysis of linguistic quantities using neural data collected with EEG (measurement at the scalp of voltages induced by neuronal firing) or fMRI (measurement of changes of oxygen concentrations in the brain tied to cognitive processes) had neither the advantages of corpora (scale) nor of informants (finer grained judgements).

However, some clear patterns of differential activity have been found for broad semantic classes. Viewing images of natural (typically animals and plants) and non-natural (typically artefacts like tools or vehicles) objects elicits different loci of activity in fMRI (Martin and Chao, 2001) and EEG (Kiefer, 2001), that persist across participants. Differences have also been found in response to auditorily or visually presented words of different lexical classes, such as abstract/concrete, and verb/noun (Pulvermüller, 2002). But interpretation of such group results remains somewhat difficult, as they may be consistent with more than one distinction: the natural/artefactual division just mentioned, may rather be between living/nonliving entities, dynamic/static entities, or be based on embodied experience (e.g. manipulable or not).

More recently, however, machine learning and other numerical techniques have been successfully applied to extract semantic information from neural data in a more discriminative fashion, down to the level of individual concepts. The work presented here builds on two strands of previous work: Murphy et al. (2008) use EEG data to perform semantic categorisation on single stimuli; and Mitchell et al. (2008) introduce an fMRI-based method that detects word level distinctions by learning associations between features of neural activity and semantic features derived from a

Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, pages 619–627,
Singapore, 6-7 August 2009. ©2009 ACL and AFNLP

corpus. We combine these innovations by introducing a method that extracts featural representations from the EEG signal, and uses corpus-based models to predict word level distinctions in patterns of EEG activity. The proposed method achieves a performance level significantly above chance (also when distinguishing between concepts from the same semantic category, e.g., *dog* and *cat*), and approaching that achieved with fMRI.

The paper proceeds as follows. The next section describes a simple behavioural experiment where Italian-speaking participants had to name photographic images of mammals and tools while their EEG activity was being recorded, and continues to detail how the rich and multidimensional signals collected were reduced to a small set of optimally informative features using a new method. Section 3 describes a series of corpus-based semantic models derived from both a raw-text web corpus, and from various parsings of a conventional corpus. In Section 4 we describe the training of a series of linear models, that each learn the associations between a set of corpus semantic features and an individual EEG activity feature. By combining these models it is possible to predict the EEG activity pattern for a single unseen word, and compare this to the observed pattern for the corresponding concept. Results (Section 5) show that these predictions succeed at a level significantly above chance, both for coarser distinctions between words in different superordinate categories (e.g., differentiating between *drill* and *gorilla*), and, at least for the model based on the larger web corpus, for those within the same category (e.g., *drill* vs *spanner*, *koala* vs *gorilla*).

2 Neural Activation Data

2.1 Data collection

EEG data was gathered from native speakers of Italian during a simple behavioural experiment at the CIMeC/DiSCoF laboratories at Trento University. Seven participants (five male and two female; age range 25-33; all with college education) performed a silent naming task. Each of them was presented[1] on screen with a series of contrast-normalised greyscale photographs of tools (garden and work tools) and land mammals (excluding emotionally valent domesticated animals and predators), for which they had to think of the most appropriate name (see figure 1). They were not explicitly asked to group the entities into superordinate categories, or to concentrate on their semantic properties, but completing the task involved resolving each picture to its corresponding concept. Images remained on screen until a keyboard response was received from the participant to indicate a suitable label had been found, and presentations were interleaved with three second rest periods. Thirty stimuli in each of the two classes were each presented six times, in random order, to give a total of 360 image presentations in the session. Response rates were over 95%, and a post-session questionnaire determined that participants agreed on image labels in approximately 90% of cases. English terms for the concepts used are listed below.

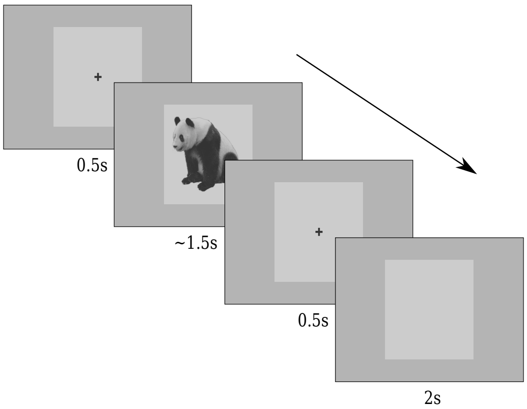

Figure 1: Presentation of image stimuli

Mammals anteater, armadillo, badger, beaver, bison, boar, camel, chamois, chimpanzee, deer, elephant, fox, giraffe, gorilla, hare, hedgehog, hippopotamus, ibex, kangaroo, koala, llama, mole, monkey, mouse, otter, panda, rhinoceros, skunk, squirrel, zebra

Tools Allen key, axe, chainsaw, craft knife, crowbar, file, garden fork, garden trowel, hacksaw, hammer, mallet, nail, paint brush, paint roller, pen knife, pick axe, plaster trowel, pliers, plunger, pneumatic drill, power drill, rake, saw, scissors, scraper, screw, screwdriver, sickle, spanner, tape measure

The EEG signals were recorded at 500Hz from 64 scalp locations based on the 10-20 standard

[1] Using the E-Prime software package: http://www.pstnet.com/e-prime/.

montage.[2] The EEG recording computer and stimulus presentation computer were synchronised by means of parallel port transmitted triggers. After the experiment, pre-processing of the recorded signals was carried out using the EEGLAB package (Delorme and Makeig, 2003): signals were band-pass filtered at 1-50Hz to remove slow drifts and high-frequency noise, and then down-sampled to 120Hz. An ICA decomposition was subsequently applied (Makeig et al., 1996), and signal components due to eye-movements were manually identified and removed.

As a preliminary test to verify that the recorded signals included category specific patterns, we applied a discriminative classification technique based on source-separation, similar to that described in Murphy et al. (2008). This found that the categories of mammals and tools could be distinguished with an accuracy ranging from 57% to 80% (mean of 72% over the seven participants).

2.2 Feature extraction

The features extracted are metrics of signal power at a particular scalp location, in a particular frequency band, and at a particular time latency relative to the presentation of each image stimulus. Termed Event Related Synchronisation (ERS) or Event Related Spectral Perturbation (ERSP), such frequency-specific changes in signal amplitude are known to correlate with a wide range of cognitive functions (Pfurtscheller and Lopes da Silva, 1999), and have specifically been shown to be sensitive to category distinctions during the processing of linguistic and visual stimuli (Murphy et al., 2008; Gilbert et al., 2009).

Feature extraction and selection is performed individually on a per-participant basis. As a first step all signal channels are z-score normalised to control for varying conductivity at each electrode site, and a Laplacian sharpening is applied to counteract the spatial blurring of signals caused by the skull, and so minimise redundancy of information between channels.

For each stimulus presentation, 14,400 signal power features are extracted: 64 electrode channels by 15 frequency bands (of width 3.3Hz, between 1 and 50Hz) by 15 time intervals (of length 67ms, in the first second after image presentation). A z-score normalisation is carried out across all stimulus presentations to equalise variance across frequencies and times: to control both for the low-pass filtering action of the skull, and for the reduced synchronicity of activity at increasing latencies. For each stimulus a mean is then taken over each of six presentations to arrive at a more reliable power estimate for each feature.[3]

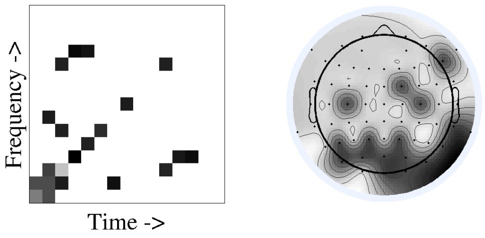

Figure 2: Mean rank of selected features in the time/frequency space (left panel) and on the scalp (right panel) for participant E

The feature ranking method used in Mitchell et al. (2008) evaluates the extent to which the relationship among stimuli is stable across across presentations, using a correlational measure,[4] but preliminary analyses with this selection method on EEG features proved disappointing. Here, two additional ranking criteria are used: each feature is evaluated for its noisiness (the amount of power variation seen across presentations of the same stimulus), and for its distinctiveness (the amount of variation in power estimates across different stimuli). A combination of these three strategies is used to rank the features by their informativeness, and the top 50 features are then selected for each participant.[5]

A qualitative evaluation of the feature selection strategy can be carried out by examining the distribution of features selected. Figure 2 shows the distribution of selected features over the time/frequency spectrum (left panel), and over the scalp (right panel - viewed from above, with the nose pointing upwards). The distribution seen is

[2] Using a Brain Vision BrainAmp system: http://www.brainproducts.com/.

[3] Stimulus power features are isolated by band-pass filtering for the required frequencies, cropping following the relevant time interval relative to each image presentation, and then taking the variance of the resulting signal, which is proportional to power.

[4] See the associated supplementary materials of Mitchell et al. (2008) for details: http://www.sciencemag.org/cgi/content/full/320/5880/1191/DC1.

[5] Several combinations of these parameters (selection thresholds of 5, 20, 50, 100, 200 features; ranking criteria in isolation and in combination) were investigated - the one chosen gave highest overall performance with the web-derived corpus model: 50 features, combined ranking criteria.

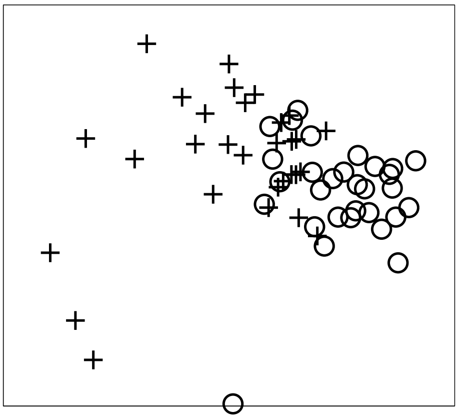

Figure 3: First two components of principal components analysis of selected features for participant E (crosses: mammals; circles: tools)

plausible in reference to previous work: lower frequencies (Pfurtscheller and Lopes da Silva, 1999), latencies principally in the first few hundred milliseconds (Kiefer, 2001), and activity in the visual centres at the rear of the head, as well as parietal areas (Pulvermüller, 2005). A principal components analysis can also be performed on the selected features to see if they reflect any plausible semantic space. As figure 3 shows, the feature selection stage has captured quite faithfully the mammal/tool distinction in a totally unsupervised fashion.

3 Corpus-based semantic models

Data from linguistics (Pustejovsky, 1995; Fillmore, 1982) and neuroscience (Barsalou, 1999; Barsalou, 2003; Pulvermüller, 2005) underline how certain verbs, by emphasising typical ways in which we interact with entities and how they behave, are pivotal in the representation of concrete nominal concepts. Following these traditions, Mitchell et al. (2008) use 25 manually picked verbs as their corpus-based features.

Here that approach is replicated by translating these verbs into Italian. Mitchell et al. (2008) selected verbs that denote our interaction with objects and living things, such as *smell* and *ride*. While the translations are not completely faithful (because frequent verbs of this sort tend to span different sets of senses in the two languages), the aim was to respect the same principle when building the Italian list. The full list, with our back translations into English, is presented in Table 1. We refer to this set as the "Mitchell" verbs.

alzare "raise"	annusare "smell/sniff"
aprire "open"	ascoltare "listen"
assaggiare "taste"	avvicinare "near"
cavalcare "ride"	correre "run/flow"
dire "say/tell"	entrare "enter"
guidare "drive"	indossare "wear"
maneggiare "handle"	mangiare "eat"
muovere "move"	pulire "clean"
puzzare "stink"	riempire "fill"
rompere "break"	sentire "feel/hear"
sfregare "rub"	spingere "push"
temere "fear"	toccare "touch"
vedere "see"	

Table 1: The "Mitchell" verbs, with English translations

As in Mitchell et al. (2008), in order to find a corpus large enough to provide reliable co-occurrence statistics for our target concepts and the 25 verbs, we resorted to the Web, queried using the Yahoo! API.[6] In particular, we represent each concept by a vector that records how many times it co-occurred with each target verb within a span of 5 words left and right, according to Yahoo! counts. We refer to this corpus-based model as the *yahoo-mitchell* model below.

While manual verb picking has proved effective for Mitchell and colleagues (and for us, as we will see in a moment), ultimately what we are interested in is discovering the most distinctive features of each conceptual category. We are therefore interested in more systematic approaches to inducing corpus-based concept descriptions, and in which of these approaches works best for this task. The alternative models we consider were not extracted from the Web, but from an existing corpus, so that we could rely on pre-existing linguistic annotation (POS tagging, lemmatization, dependency paths), and perform more flexible, annotation-aware queries to collect co-occurrence statistics.

More specifically, we used the *la Repubblica/SSLMIT* corpus[7], that contains about 400 million tokens of newspaper text. From this, we extracted four models where nominal concepts are represented in terms of patterns of co-occurrence with verbs (we collected statistics for the top 20,000 most common nouns in the corpus, including the concepts used as stimuli in the silent nam-

[6] `http://developer.yahoo.com/search/`

[7] `http:://sslmit.unibo.it/repubblica/`

ing experiment, and the top 5,000 verbs). We first re-implemented a "classic" window-based word space model (Sahlgren, 2006), referred to below as *repubblica-window*, where each noun lemma is represented by its co-occurrence with verb lemmas within the maximum span of a sentence, with no more than one other intervening noun. The *repubblica-position* model is similar, but it also records the position of the verb with respect to the noun (so that *X-usare* "X-use" and *usare-X* "use-X" count as different features), analogously to the seminal HAL model (Lund and Burgess, 1996). It has been shown that models that take the syntactic relation between a target word and a collocate feature into account can outperform "flat" models in some tasks (Padó and Lapata, 2007). The next two models are based on the dependency parse of the *la Repubblica* corpus documented by Lenci (2009). We only counted as collocates those verbs that were linked to nouns by a direct path (such as subject and object) or via preposition-mediated paths (e.g., *tagliare* ***con*** *forbici* "to cut **with** scissors"), and where the paths were among the top 30 most frequent in the corpus. In the *repubblica-depfilter* model, we record co-occurrence with verbs that are linked to the nouns by one of the top 30 paths, but we do not preserve the paths themselves in the features. This is analogous to the model proposed by Padó and Lapata (2007). In the *repubblica-deppath* model, we preserve the paths as part of the features (so that *subj-uccidere* "subj-kill" and *obj-uccidere* count as different features), analogously to Lin (1998), Curran and Moens (2002) and others. For all models, following standard practice in computational linguistics (Evert, 2005), we transform raw co-occurrence counts into log-likelihood ratios.

Following the evaluation paradigm of Mitchell et al. (2008), linear models trained on corpus-based features are used to predict the pattern of EEG activity for unseen concepts. This only works if we have a very limited number of features (or else we would have more parameters to estimate than data-points to estimate them). The Repubblica-based models have thousands of features (one per verb collocate, or verb+path collocate). We adopt two strategies to select a reduced number of features. In the *topfeat* versions, we first pick the 50 features that have the highest association with each of the target concepts. We then count in how many of these concept-specific top lists a feature occurs, and we pick the 25 features that occur in the largest number of them. The intuition is that this should give us a good trade-off between how characteristic the features are (we only use features that are highly associated with some of our concepts), and their generalization capabilities (we pick features that are associated with multiple concepts). Randomly selected examples of the features extracted in this way for the various Repubblica models are reported in Table 2.

repubblica-window	*repubblica-position*
abbattere "demolish"	X-ferire "X-wound"
afferrare "seize"	X-usare "X-use"
impugnare "grasp"	dipingere-X "paint-X"
tagliare "cut"	munire-X "supply-X"
trovare "find"	tagliare-X "cut-X"
repubblica-depfilter	*repubblica-deppath*
abbattere "demolish"	con+tagliare "with+cut"
correre "run"	obj+abbattere "obj+demolish"
parlare "speak"	obj+uccidere "obj+kill"
saltare "jump"	intr-subj+vivere "intr-subj+live"
tagliare "cut"	tr-subj+aprire "tr-subj+open"

Table 2: Examples of top features from the *la Repubblica* models

Alternatively, instead of feature *selection* we perform feature *reduction* by means of a Singular Value Decomposition (SVD) of the noun-by-verb matrix. We apply the SVD to matrices that include the top 20,000 most frequent nouns in the corpus (including our target concepts) since the quality of the resulting reduced model should improve if we can exploit richer patterns of correlations among the columns – verbs – across rows – nouns (Landauer and Dumais, 1997; Schütze, 1997). In the *svd* versions of our models, we pick as features the top 25 left singular vectors, weighted by the corresponding singular values. These features do not have a straightforward interpretation, but they tend to group verb meanings that belong to broad semantic domains. For example, among the original verbs that are most strongly correlated with one of the top singular vectors of *repubblica-window* we find *giocare* "play", *vincere* "win" and *perdere* "lose". Another singular vector is associated with *ammontare* "amount", *costare* "cost", *pagare* "pay", etc. One of the top singular vectors of *repubblica-deppath* is strongly correlated with *in+scendere* "descend into", *in+mettere* "put into", *in+entrare* "enter into", though not all singular vectors are so clearly characterized by the verbs they correlate with.

None of the *la Repubblica* models had full coverage of our concept stimulus set (see the second column of Table 3 below), because our extraction method missed some multi-word units, and feature selection led to losing some more items due to data sparseness (e.g., some target words had no collocates connected by the dependency paths we selected). The experiments reported in the next section used all the target concepts available in each model, but a replication using the 50 concepts that were common to all models obtained results that are comparable. For a direct comparison between Yahoo! and *la Repubblica* derived features, we tried collecting statistics for the Mitchell verbs from Repubblica as well, but the resulting model was extremely sparse, and we do not report its performance here.

Finally, it is important to note that any representation yielded by a corpus semantic model does not characterise a concept directly, but is rather an aggregate of the various senses and usages of the noun chosen to represent it. This obvious limitation will persist until comprehensive, robust and computationally efficient word-sense disambiguation techniques become available. However these models are designed to extract semantic (as opposed to syntactic or phonological) properties of words, and as noted in the introduction, have been demonstrated to correlate with behavioural effects of conceptual processing.

4 Predicting EEG patterns using corpus-based models

In Section 2.2 above we showed how we extracted features summarizing the spatial, temporal and frequency distribution of the EEG signal collected while participants were processing each of the target concepts. In Section 3, we described various ways to obtain a compact representation of the same concepts in terms of corpus-derived features. We will now discuss the method we employed to verify whether the corpus-derived features can be used to predict the EEG patterns – that is whether semantics can be used to predict neural activity. Our hope is that a good corpus-based model will provide a decomposition of concepts into meaningful properties, corresponding to coherent subpatterns of activation in the brain, and thus capture generalizations across concepts. For example, if a concept is particularly visually evocative (e.g., *zebra*), we might expect it to be strongly associated with the verb *see*, while also causing particular activation of the vision centres of the brain. Similarly, concepts with strong associations with a particular sound (e.g., *cuckoo*) might be semantically associated with *hear* while also disproportionately activating auditory areas of the brain. It should thus be possible to learn a model of corpus-to-EEG-pattern correspondences on training data, and use it to predict the EEG activation patterns of unseen concepts.

We follow the paradigm proposed by Mitchell et al. (2008) for fMRI data. For each participant and selected EEG feature, we train a model where the level of activation of the latter in response to different concepts is approximated by a linear combination of the corpus features:

$$\vec{f} = \mathbf{C}\vec{\beta} + \vec{\epsilon}$$

where $\vec{f}$ is the vector of activations of a specific EEG feature for different concepts, the matrix $\mathbf{C}$ contains the values of the corpus features for the same concepts (row-normalised to *z*-scores), $\vec{\beta}$ is the weight we must learn for each corpus feature, and $\vec{\epsilon}$ is a vector of error terms. We use the method of least squared errors to learn the weights that maximize the fit of the model. We can then predict the activation of an EEG feature in response to a new concept that was not in the training data by a $\vec{\beta}$-weighted sum of the values of each corpus feature for the new concept. In some cases collinearity in the corpus data (regular linear relationships among the corpus-feature columns) prevented the estimation procedure from finding a solution. In such cases (due to the small number of data, relative to the number of unknowns), the least informative corpus-features (those that correlated on average most highly with other features) were iteratively removed until a solution was reached. All models were trained with between 23 and 25 corpus features.

Again following Mitchell and colleagues, we adopt a leave-2-out paradigm in which a linear model for each EEG feature is trained in turn on all concepts minus 2. For each of the 2 left out concepts, we predict the EEG activation pattern using the trained linear model and their corpus features, as just described. We then try to correctly match the predicted and observed activations, by measuring the Euclidean distance between the model-generated EEG activity (a vector of estimated power levels for the n EEG features selected) and the corresponding EEG activity recorded in the experiment (other distance metrics gave similar results to the ones reported here). Given the 2 left-out concepts a and b, we compute 2 *matched* distances (i.e., distance between predicted and observed pattern for a, and the same for b) and 2 *mismatched* distances (predicted a and observed b; predicted b and observed a). If the average of the matched distances is lower than the average of the mismatched distances, we consider the prediction successful – otherwise we count is as a failed prediction. At chance levels, expected matching accuracy is 50%.

5 Results

Table 3 shows the comparative results for all the corpus models introduced in Section 3. The third column (*all*) shows the overall accuracy in correctly matching predicted to observed EEG activity patterns, and so successfully distinguishing word meanings. The significance of the figures is indicated with the conventional annotation (calculated using a one-way two-sided t-test across the individual participant accuracy figures against an expected population mean of 50%).[8] The second column shows the coverage of each model of the 60 mammal and tool concepts used, which ranged from full (for the *yahoo-mitchell* model) to 51 concepts (for the *depfilter-topfeat* model). The corresponding number of matching comparisons over which accuracy was calculated ranged from 1770 down to 1225.

As suggested by previous work (Murphy et al., 2008), and illustrated by figure 3, coarse distinctions between words in different superordinate categories (e.g., *hammer* vs *armadillo*; *giraffe* vs *nail*) may be easier to detect than those among concepts within the same category (e.g., *hammer* vs *nail*; *giraffe* vs *armadillo*). The fourth and fifth columns give these accuracies, and while between-category discriminations do prove more reliable, they indicate that, for the top rated model at least, finer within-category distinctions are also being captured. Figures from the top two performing models are given for individual participants in tables 4 and 5.

part.	*overall*	*within*	*between*
A	54	53	55
B	54	47	60
C	62	56	67
D	61	56	67
E	68	58	78
F	52	54	51
G	57	51	63

Table 4: Accuracy levels for individual participant sessions, *yahoo-mitchell* web corpus

part.	*overall*	*within*	*between*
A	49	52	46
B	59	57	60
C	60	60	59
D	50	45	55
E	56	53	58
F	64	64	65
G	52	49	55

Table 5: Accuracy levels for individual participant sessions, *repubblica-window-svd*

6 Discussion

Our results show that corpus-extracted conceptual models can be used to distinguish between the EEG activation levels associated with conceptual categories to a degree that is significantly above chance. Though category specific patterns are detectable in the EEG signal alone (as illustrated by the PCA analysis in figure 3), on that basis we cannot be sure that semantics is being detected. Some other property of the stimuli that co-varies with the semantic classes of interest could be responsible, such as visual complexity, conceptual familiarity, lexical frequency, or phonological form. Only by cross-training with individual corpus features and showing that these hold a predictive relationship to neural activity have we been able to establish that EEG patterns encode semantics.

Present evidence indicates that fMRI may provide richer data for training such models than EEG (Mitchell and colleagues obtain an average accuracy of 77%, and 65% for the within category setting). However, fMRI has several clear disadvantages as a tool for language researchers. First of all, the fine spatial resolution it provides (down to 2-3mm), while of great interest to neuroscientists, is not in itself linguistically informative. Its coarse temporal resolution (of the order of several seconds), makes it ill-suited to analysing on-line linguistic processes. EEG on the other hand, despite its low spatial resolution (several centimetres), gives millisecond-level temporal resolution,

[8] On average, the difference seen between matched and mismatched pairs was small, at about 3% of the distance between observed and predicted representations, and was marginally bigger for correct than for incorrect predictions ($p < 0.01$).

model	*coverage*	*all*	*within cat*	*between cat*
yahoo-mitchell	100	58.3** (5.7)	53.6* (3.7)	63.0** (8.9)
repubblica-window-svd	96.7	55.7* (5.6)	54.3 (6.5)	56.9* (5.9)
repubblica-window-topfeat	93.3	52.1 (4.3)	48.7 (3.6)	55.4 (7.0)
repubblica-deppath-svd	93.3	51.4 (8.7)	49.0 (8.0)	54.0 (10.0)
repubblica-depfilter-topfeat	85.0	51.1 (9.6)	49.3 (9.6)	53.1 (10.0)
repubblica-position-topfeat	93.3	50.0 (5.2)	46.0 (4.7)	53.6 (8.0)
repubblica-deppath-topfeat	86.7	49.9 (9.0)	47.0 (9.3)	52.4 (9.6)
repubblica-position-svd	96.7	49.4 (10.2)	46.6 (9.8)	52.3 (11.3)
repubblica-depfilter-svd	93.3	48.9 (11.1)	47.1 (8.9)	50.6 (12.9)

Table 3: Comparison across corpus models, with percentage concept coverage, mean cross-subject percentage prediction accuracy and standard deviation; $*p < 0.05$, $**p < 0.01$

enabling the separate analysis of sequential cognitive processes and states (e.g., auditory processing, word comprehension, semantic representation). fMRI is also prohibitively expensive for most researchers (ca. 300 euros per hour at cost price), compared to EEG (ca. 30 euros per hour). Finally, there is no prospect of fMRI being miniaturised, while wearable EEG systems are already becoming commercially available, making experimentation in more ecological settings a possibility (e.g., playing with a child, meeting at a desk, walking around). In short, while EEG can be used to carry out systematic investigations of categorical distinctions, doing so with fMRI would be prohibitively expensive.

Present results indicate that distinctions between categories are easier than distinctions between category elements; and that selecting the conceptual features by hand gives better results than discovering them automatically. Both of these results however may be due to limitations of the current method. One limitation is that we have been using the same set of features for all concepts, which is likely to blur the distinctions between members of a category more than those between categories. A second limitation of our present methodology is that it is constrained to use very small numbers of semantic features, which limits its applicability. For example it is hard to conceive of a small set of verbs, or other parts-of-speech, whose co-occurrence patterns could successfully characterise the full range of meaning found in the human lexicon. Even the more economical corpus-extracted conceptual models tend to run in the hundreds of features (Almuhareb, 2006). We are currently working on variations in the method that will address these shortcomings.

The web-based model with manually picked features outperformed all *la Repubblica*-based models. However, the results attained with *repubblica-window-svd* are encouraging, especially considering that we are reporting results for an EEG feature configuration optimised for the web data (see footnote 5), and that *la Repubblica* is several orders of magnitude smaller than the web. That data sparseness might be the main issue with *la Repubblica* models is suggested by the fact that *repubblica-window-svd* is the least sparse of them, since it does not filter data by position or dependency path, and compresses information from many verbs via SVD. In future research, we plan to extract richer models from larger corpora. And as the discriminative accuracy of cross-training techniques improves, further insights into the relative validity of corpus representations will be attainable. One research aim is to see if individual corpus semantic properties are encoded neurally, so providing strong evidence for a particular model. These techniques may also prove more objective and reliable in evaluating representations of abstract concepts, for which it is more difficult to collect reliable judgements from informants.

References

A. Almuhareb. 2006. *Attributes in lexical acquisition*. Dissertation, University of Essex.

L. Barsalou. 1999. Perceptual symbol systems. *Behavioural and Brain Sciences*, 22:577–660.

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

J.R. Curran and M. Moens. 2002. Improvements in automatic thesaurus extraction. In *Proceedings of SIGLEX*, pages 59–66.

A. Delorme and S. Makeig. 2003. Eeglab: an open source toolbox for analysis of single-trial dynamics includingindependent component analysis. *Journal of Neuroscience Methods*, 134:9–21.

S. Evert. 2005. *The statistics of word cooccurrences*. Dissertation, Stuttgart University.

Ch. J. Fillmore. 1982. Frame semantics. In Linguistic Society of Korea, editor, *Linguistics in the Morning Calm*, pages 111–138. Hanshin, Seoul.

J. Gilbert, L. Shapiro, and G. Barnes. 2009. Processing of living and nonliving objects diverges in the visual processing system: evidence from meg. In *Proceedings of the Cognitive Neuroscience Society Annual Meeting*.

M. Kiefer. 2001. Perceptual and semantic sources of category-specific effects in object categorization:event-related potentials during picture and word categorization. *Memory and Cognition*, 29(1):100–116.

T. Landauer and S. Dumais. 1997. A solution to Platos problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104(2):211–240.

A. Lenci. 2009. Argument alternations in italian verbs: a computational study. In *Atti del XLII Congresso Internazionale di Studi della Società di Linguistica Italiana*.

D. Lin. 1998. Automatic retrieval and clustering of similar words. In *COLING-ACL98*, Montreal, Canada.

K. Lund and C. Burgess. 1996. Producing high-dimensional semantic spaces from lexical cooccurrence. *Behavior Research Methods, Instruments, and Computers*, 28:203—208.

S. Makeig, A.J. Bell, T. Jung, and T.J. Sejnowski. 1996. Independent component analysis of electroencephalographic data. In *in Advances in Neural Information Processing Systems*, pages 145–151. MIT Press.

A. Martin and L. Chao. 2001. Semantic memory and the brain: structure and processes. *Current Opinions in Neurobiology*, 11:194–201.

T. Mitchell, S. Shinkareva, A. Carlson, K. Chang, V. Malave, R. Mason, and M. Just. 2008. Predicting human brain activity associated with the meanings of nouns. *Science*, 320:1191–1195.

B. Murphy, M. Dalponte, M. Poesio, and L. Bruzzone. 2008. Distinguishing concept categories from single-trial electrophysiological activity. In *Proceedings of the Annual Meeting of the Cognitive Science Society*.

S. Padó and M. Lapata. 2007. Dependency-based construction of semantic space models. *Computational Linguistics*, 33(2):161–199.

G. Pfurtscheller and F. Lopes da Silva. 1999. Event-related EEG/MEG synchronization and desynchronization: Basic principles. *Clinical Neurophysiology*, 110:1842–1857.

F. Pulvermüller. 2002. *The neuroscience of language: on brain circuits of words and serial order*. Cambridge University Press, Cambridge.

F. Pulvermüller. 2005. Brain mechanisms linking language and action. *Nature Reviews Neuroscience*, 6:576–582.

J. Pustejovsky. 1995. *The Generative Lexicon*. MIT Press, Cambridge.

M. Sahlgren. 2006. *The Word-Space Model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces*. Dissertation, Stockholm University.

S. Schulte im Walde. 2008. *Theoretical adequacy, human data and classification approaches in modelling word properties, word relatedness and word classes*. Habilitation, Saarland University.

H. Schütze. 1997. *Ambiguity Resolution in Natural Language Learning*. CSLI, Stanford.