Visually Grounded and Textual Semantic Models Differentially Decode Brain Activity Associated with Concrete and Abstract Nouns

Andrew J. Anderson
Brain & Cognitive Sciences
University of Rochester
aander41@ur.rochester.edu

Douwe Kiela
Computer Laboratory
University of Cambridge
dk427@cam.ac.uk

Stephen Clark
Computer Laboratory
University of Cambridge
sc609@cam.ac.uk

Massimo Poesio
School of Computer Science and Electronic Engineering
University of Essex
poesio@essex.ac.uk

Abstract

Important advances have recently been made using computational semantic models to decode brain activity patterns associated with concepts; however, this work has almost exclusively focused on concrete nouns. How well these models extend to decoding abstract nouns is largely unknown. We address this question by applying state-of-the-art computational models to decode functional Magnetic Resonance Imaging (fMRI) activity patterns, elicited by participants reading and imagining a diverse set of both concrete and abstract nouns. One of the models we use is linguistic, exploiting the recent word2vec skipgram approach trained on Wikipedia. The second is visually grounded, using deep convolutional neural networks trained on Google Images. Dual coding theory considers concrete concepts to be encoded in the brain both linguistically and visually, and abstract concepts only linguistically. Splitting the fMRI data according to human concreteness ratings, we indeed observe that both models significantly decode the most concrete nouns; however, accuracy is significantly greater using the text-based models for the most abstract nouns. More generally this confirms that current computational models are sufficiently advanced to assist in investigating the representational structure of abstract concepts in the brain.

1 Introduction

Since the work of Mitchell et al. (2008), there has been increasing interest in using computational semantic models to interpret neural activity patterns scanned as participants engage in conceptual tasks. This research has almost exclusively focused on brain activity elicited as participants comprehend concrete nouns as experimental stimuli. Different modelling approaches — predominantly distributional semantic models (Mitchell et al., 2008; Devereux et al., 2010; Murphy et al., 2012; Pereira et al., 2013; Carlson et al., 2014) and semantic models based on human behavioural estimation of conceptual features (Palatucci et al., 2009; Sudre et al., 2012; Chang et al., 2010; Bruffaerts et al., 2013; Fernandino et al., 2015) — have elucidated how different brain regions contribute to semantic representation of concrete nouns; however, how these results extend to non-concrete nouns is unknown.

In computational modelling there has been increasing importance attributed to grounding semantic models in sensory modalities, e.g., Bruni et al. (2014), Kiela and Bottou (2014). Andrews et al. (2009) demonstrated that multi-modal models formed by combining text-based distributional information with behaviourally generated conceptual properties (as a surrogate for perceptual experience) provide a better proxy for human-like intelligence. However, both the text-based and behaviourally-based components of their model were ultimately derived from linguistic information. Since then, in analyses of brain data, Anderson et al. (2013) have applied multi-modal models incorporating features that are truly grounded in natural image statistics to further support this claim. In addition, Anderson et al. (2015) have demonstrated that visually grounded models describe brain activity associated with internally induced visual features of objects as the ob-

Transactions of the Association for Computational Linguistics, vol. 5, pp. 17–30, 2017. Action Editor: Daichi Mochihashi.
Submission batch: 2/2016; Revision batch: 7/2016; Published 1/2017.

jects names are read and comprehended.

Having both image- and text-based models of semantic representation, and neural activity patterns associated with concrete and abstract nouns, enables a natural test of Dual coding theory (Paivio, 1971). Dual coding posits that concrete concepts are represented in the brain in terms of a visual and linguistic code, whereas abstract concepts are only represented by a linguistic code. Whereas previous work has demonstrated that image- and text-based semantic models contribute to explaining neural activity patterns associated with concrete nouns, it remains unclear whether either text- or image-based semantic models can decode neural activity patterns associated with abstract words.

We extend previous work by applying image- and text-based computational semantic models to decode an fMRI data set spanning a diverse set of nouns of varying concreteness. The 70-word stimuli for the fMRI experiment (listed in Table 1) are semantically structured according to taxonomic categories and domains embedded in WordNet (Fellbaum, 1998) and its extensions. Participants read the noun and were instructed to imagine a situation that they personally associate with the noun. In this sense, the data solicited was targetting deep thought patterns (deeper than might be anticipated for rapid semantic processing required in conversations and many real time interactions with the world). In the analysis we split the fMRI data set into the most concrete and most abstract words based on behavioural concreteness ratings. Our key contribution is in demonstrating a decoding advantage for text-based semantic models over the image-based models when decoding the more abstract nouns. In line with the previous results of Anderson et al. (2013) and Anderson et al. (2015), both visual and textual models decode the more concrete nouns.

The image- and text-based computational models we use have recently been developed using neural networks (Mikolov et al., 2013; Jia et al., 2014). The image-based model is built using a deep convolutional neural network approach, similar in nature to those recently used to study neural representations of visual stimuli (see Kriegeskorte (2015), although note this is the first application to study word elicited neural activation known to the authors). For decoding we use a recently introduced algorithm (Anderson et al., 2016) that abstracts the decoding task to representational similarity space, and achieve decoding accuracies on par with those conventionally achieved through discriminating concrete nouns (and higher if we combine data to exploit group-level regularities).

Because the fMRI experiments were performed in Italian on native Italians, and because approximately comparable text corpora in content were available in English and Italian (English and Italian Wikipedia), we were able to compare how well English and Italian text-based semantic models can decode neural activity patterns. Whilst Italian Wikipedia could reasonably be expected to be advantaged by supporting culturally appropriate nuances of semantic structure, it is disadvantaged by being considerably smaller than English Wikipedia. Taking inspiration from previous work exploiting cross-lingual resources (Richman and Schone, 2008; Shi et al., 2010; Darwish, 2013) we combined Italian and English text-based models in our decoding analyses in an attempt to leverage the benefits of both.

Although combined language and English models tended to yield marginally better decoding accuracies, there were no significant differences between the different language models. Whilst we expect semantic structure on a grand scale to broadly straddle language boundaries for most concrete and abstract concepts (albeit with cultural specificities), this is proof of principle that cross linguistic commonalities are reflected in neural activity patterns measurable with current technology.

2 Brain Data

We reanalyze the fMRI data originally collected by Anderson et al. (2014), who investigated the relevance of different taxonomic categories and domains embedded in WordNet to the organization of conceptual knowledge in the brain.

2.1 Word stimuli

Anderson et al. (2014) systematically selected a list of 70 words intended to be representative of a broad range of abstract and concrete nouns. These were organised according to the domains of law and music, cross-classified with seven taxonomic categories. They began by identifying low-concreteness

	LAW		MUSIC	
Ur-abstracts	giustizia	justice	musica	music
	liberta'	liberty	blues	blues
	legge	law	jazz	jazz
	corruzione	corruption	canto	singing
	refurtiva	**loot**	punk	punk
Attribute	giurisdizione	jurisdiction	sonorita'	sonority
	cittadinanza	citizenship	ritmo	rhythm
	impunita'	impunity	~~melodia~~	~~melody~~
	legalita'	legality	tonality'	tonality
	illegalita	illegality	intonazione	pitch
Communication	divieto	prohibition	canzone	song
	verdetto	verdict	**pentagramma**	**stave**
	ordinanza	decree	ballata	ballad
	addebito	accusation	ritornello	refrain
	ingiunzione	injunction	sinfonia	symphony
Event/action	arresto	arrest	**concerto**	**concert**
	processo	trial	recital	recital
	reato	crime	assolo	solo
	furto	theft	**festival**	**festival**
	assoluzione	acquital	**spettacolo**	**show**
Person/Social-role	**giudice**	**judge**	**musicista**	**musician**
	ladro	**thief**	**cantante**	**singer**
	imputato	**defendant**	**compositore**	**composer**
	testimone	**witness**	**chitarrista**	**guitarist**
	avvocato	**lawyer**	**tenore**	**tenor**
Location	**tribunale**	**court/tribunal**	**palco**	**stage**
	carcere	**prison**	**auditorium**	**auditorium**
	~~**questura**~~	~~**police-station**~~	**discoteca**	**disco**
	penitenziario	**penitentiary**	**conservatorio**	**conservatory**
	patibolo	**gallows**	**teatro**	**theatre**
Object/Tool	**manette**	**handcuffs**	**violino**	**violin**
	toga	**robe**	**tamburo**	**drum**
	manganello	**truncheon**	**tromba**	**trumpet**
	cappio	**noose**	**metronomo**	**metronome**
	~~**grimaldello**~~	~~**skeleton-key**~~	**radio**	**radio**

Table 1: Italian stimulus words and English translations, divided into law and music domains (columns), and taxonomic categories (groups of 5 rows). The most concrete half of the words are indicated in bold font. Strike-throughs indicate words for which we did not have semantic model coverage.

words in the norms of Barca et al. (2002). They then linked these to WordNet to identify the taxonomic category of the dominant sense of each word. Six taxonomic categories that were heavily populated with abstract words, as well as one unambiguously concrete category, were chosen. All categories supported ample coverage of Law and Music domains (determined according to WordNet Domains (Bentivogli et al., 2004)). Five law words and five music words were selected from each taxonomic category. Taxonomic categories and example stimulus words (translated into English) are as below:

Ur-abstract: Anderson et al.'s term for concepts that are classified as abstract in WordNet but do not belong to a clear subcategory, e.g., *law* or *music*. **Attribute**: A construct whereby objects or individuals can be distinguished, e.g., *legality*, *tonality*. **Communication**: Something that is communicated by, to or between groups, e.g., *accusation*, *symphony*. **Event/action**: Something that happens at a given place and time, e.g., *crime*, *festival*. **Person/Social-role**: Individual, someone, somebody, mortal, e.g., *judge*, *musician*. **Location**: Points or extents in space, e.g., *court*, *theatre*. **Object/Tool**: A class of unambiguously concrete nouns, e.g., *handcuffs*, *violin*.

The full list of stimuli is in Table 1. We split the stimulus nouns into the 35 most concrete and 35 most abstract words according to the behavioural concreteness ratings from Anderson et al. (2014).

2.2 fMRI Experiment

Participants Nine right-handed native Italian speakers aged between 19 and 38 years (3 women) were recruited to take part in the study. Two were scanned after Anderson et al. (2014) to match the number of participants analysed by Mitchell et al. (2008). Scanning had previously been halted at 7 instead of the planned 9 participants for a period due to equipment failure. All had normal or corrected-to-normal vision.

The 70 stimulus words were presented as written words, in 5 runs (all runs were collected in one participant visit), with the order of presentations randomised across runs. In each run, a randomly selected word was presented every 10 seconds, and remained on screen for 3 seconds. On reading a stimulus word, participants thought of a situation that they individually associated with the noun. This process is similar to previous concrete noun tasks, e.g., Mitchell et al. (2008), where participants were instructed to think of the properties of the noun. However, as people encounter difficulties eliciting properties of non-concrete concepts, compared to thinking of situations in which concepts played a role (Wiemer-Hastings and Xu, 2005), the experimental paradigm was adapted to imagining situations.

fMRI acquisition and preprocessing Anderson et al. (2014) recorded fMRI images on a 4T Bruker MedSpec MRI scanner. They used an Echo Planar Imaging (EPI) pulse sequence with a 1000 msec repetition time, an echo time of 33 msec, and a 26° flip angle. A 64×64 acquisition matrix was used, and 17 slices were imaged with a between-slice gap of 1 mm. Voxels had dimensions of 3mm×3mm×5mm. fMRI data were corrected for head motion, unwarped, and spatially normalized to the Montreal Neurological Institute and Hospital (MNI) template. Only voxels estimated to be grey matter were included in the subsequent analysis. For each participant, for each scanning run (where a run is a complete presentation of 70 words), voxel activity was corrected by removing linear trend and transformed to z scores (within each run). Each stimulus word was represented as a single volume by taking the voxel-wise mean of the 4 sec of data offset by 4 sec from the stimulus onset (to account for hemodynamic response).

Voxel selection The 500 most stable grey matter voxels per participant were selected for analysis. This was undertaken within the leave-2-word-out decoding procedure detailed later in Section 4 using the same method as Mitchell et al. (2008): Pearson's correlation of each voxel's activity between matched word lists in all scanning run pairs (10 unique run pairs giving 10 correlation coefficients of 68/70 words, where the other 2 words were test words to be decoded) was computed. The mean coefficient was used as stability measure. Voxels associated with the 500 largest stability measures were selected.

3 Semantic Models

3.1 Image-based semantic models

Following previous work in multi-modal semantics (Bergsma and Van Durme, 2011; Kiela et al., 2014), we obtain a total of 20 images for each of the stimulus words from Google Images[1]. Images from Google have been shown to yield representations that are competitive in quality compared to alternative resources (Bergsma and Van Durme, 2011; Fergus et al., 2005). Image representations are obtained by extracting the pre-softmax layer from a forward pass in a convolutional neural network (CNN) that has been trained on the ImageNet classification task using Caffe (Jia et al., 2014). This approach is similar to e.g., Kriegeskorte (2015), except that we only use the pre-softmax layer, which has been found to work particularly well in semantic tasks (Razavian et al., 2014; Kiela and Bottou, 2014). Such CNN-derived image representations have been found to be of higher quality than traditional bag of visual words models (Sivic and Zisserman, 2003) that were previously used in multi-modal semantics (Bruni et al., 2014; Kiela and Bottou, 2014). We aggregate images associated with a stimulus word into an overall visually grounded representation by taking the mean of the individual image representations.

Image search for abstract nouns The validity and success of the following analyses are dependent on having built the image-based models from a set of images that are indeed relevant to the abstract words. The Google Image searches we used

[1] www.google.com/imghp

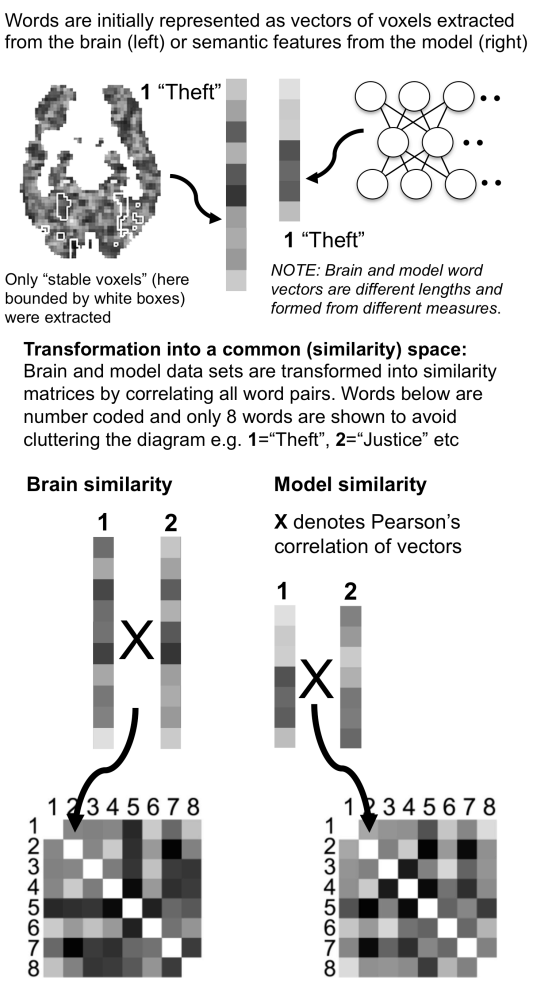


Figure 1: Representing brain and semantic model vectors in similarity space.

to build the image-based models largely returned a selection of images systematically associated with our most abstract nouns. For instance, 'corruption' returns suited figures covertly exchanging money; 'law', 'justice', 'music', 'tonality' return pictures of gavels, weighing scales, musical notes and circles of fifths, respectively. For 'jurisdiction', the image search returns maps and law-related objects. However, there were also misleading cases such as 'pitch' where the image search, whilst returning potentially useful pictures of sinusoidal graphs, was heavily contaminated by images of football pitches. This problem is not exclusive to images, and the current text-based models are also not immune to the multiple senses of polysemous words.

3.2 Text-based semantic models

For linguistic input, we use the continuous vector representations from the skip-gram model of Mikolov et al. (2013). Specifically, we obtained 300-dimensional word embeddings by training a skip-gram model using negative sampling on recent Italian and English Wikipedia dumps (using Gensim with preprocessing from word2vec's demo script). For English, representations were built for the English translations of the 70 stimuli provided by Anderson et al. (2014). The English model was trained for 1 iteration, whereas the Italian was trained for 5, since the Italian Wikipedia dump was smaller (5.2 vs 1.3 billion words respectively).

Following previous work exploiting cross-lingual textual resources (Richman and Schone, 2008; Shi et al., 2010; Darwish, 2013), we also applied Italian and English text-based models in combination. Model combination was achieved at the analysis stage, by fusing decoding outputs of Italian and English models as described in Section 4.1.

4 Representational similarity-based decoding of brain activity

We decoded word-level fMRI representations using the semantic models following the procedure introduced by Anderson et al. (2016). The process of matching models to words is abstracted to representational similarity space: For both models and brain data, words are semantically re-represented by their similarities to other words by correlating all word pairs within the native model or brain space, using Pearson's correlation (see Figure 1). The result is two square matrices of word pair correlations: one for the fMRI data, another for the model. In the similarity space, each word is a vector of correlations with all other words, thereby allowing model and brain words (similarity vectors) to be directly matched to each other.

In decoding, models were matched to fMRI data as follows (see Figure 2). Two test words were chosen. The 500 voxels estimated to have the most stable signal were selected using the strategy described in Section 2.2. Voxel selection was based on the fMRI data of the other 68/70 words. Selection on 68/70 rather than all 70 words was to allay any concern that voxel selection could have

systematically biased the fMRI correlation structure (calculated next) to look like that of the semantic model, and consequently biased decoding performance. However, as similarity-based decoding does not optimise a mapping between fMRI data and semantic model, it is not prone to modelling and decoding fMRI noise as in classic cases of double dipping (Kriegeskorte et al., 2009). Indeed, as we report later in this section, there were no significant differences in decoding accuracy arising from tests using voxel selection on 68/70 versus 70 words.

A single representation of each word was built by taking the voxel-wise mean of all five presentations of the word for the 500 selected voxels. An fMRI similarity matrix for all 70 words was then calculated. Similarity vectors for the two test words were drawn from both the model and fMRI similarity matrices. Entries corresponding to the two test words in both model and fMRI similarity vectors were removed because these values could reveal the correct answer to decoding. The two model similarity vectors were then compared to the two fMRI similarity vectors by correlation, resulting in four correlation values. These correlation values were transformed using Fisher's r to z (arctanh). If the sum of z-transformed correlations between the correctly matched pair exceeded the sum of correlations for the incongruent pair, decoding was scored a success, otherwise a failure. This process was then repeated for all word pairs, with the mean accuracy of all test iterations giving a final measure of success.

Fisher's r to z transform (arctanh) is typically used to test for differences between correlation coefficients. It transforms the correlation coefficient r to a value z, where z has amplified values at the tails of the correlation coefficient (r otherwise ranges between -1 and 1). This is to make the sampling distribution of z normally distributed, with approximately constant variance values across the population correlation coefficient. In the similarity-decoding method used here, z is evaluated in decoding because it is a more principled metric to compare and combine (as later undertaken in Section 4.1)

However, under most circumstances r to z is not critical to the procedure. z noticeably differs from r only when correlations exceed .5, and r to z changes decoding behaviour in select circumstances. Specifically r to z can influence how word labels are as-

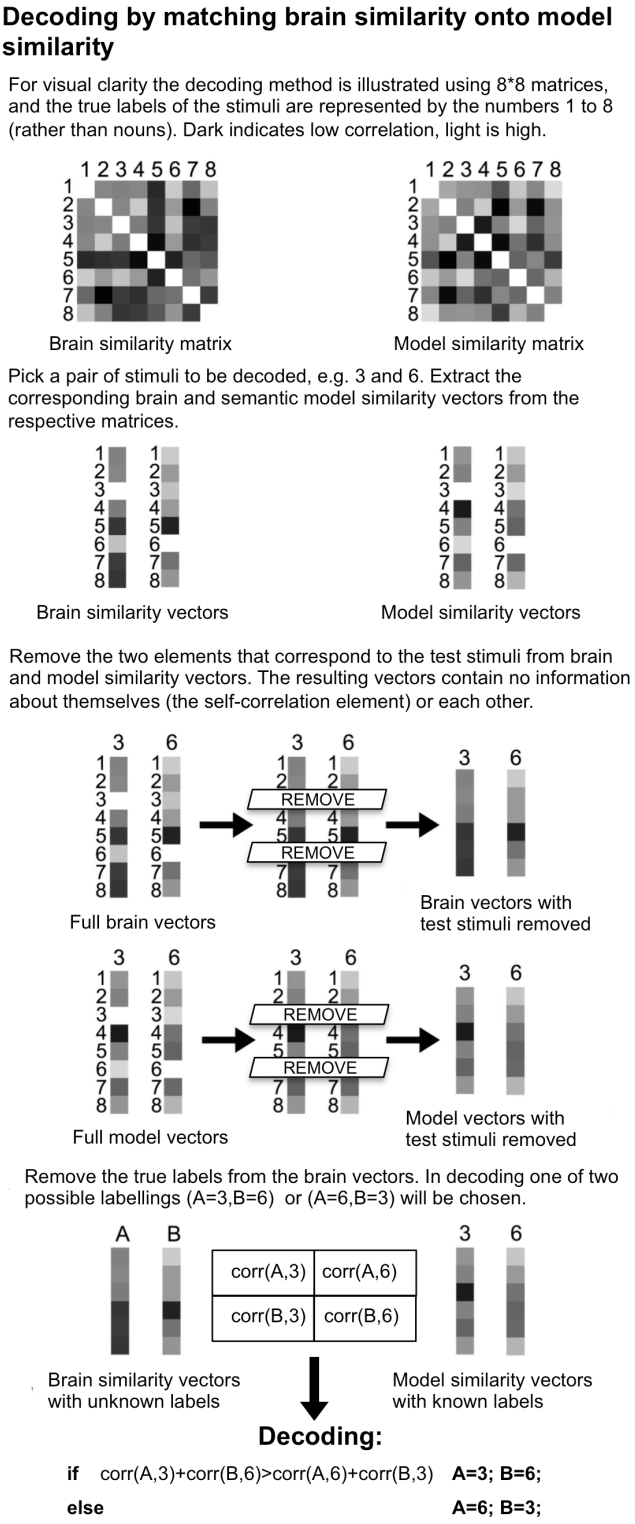


Figure 2: Similarity-decoding algorithm (adapted from Anderson et al. 2016).

signed to similarity vectors by upweighting high value correlation coefficients at the final stage of decoding.

A hypothetical scenario to illustrate the above point is as follows. Let Pearson(X,Y) denote Pearson's correlation of vectors X and Y, and brainA correspond to a brain similarity vector "A" for an unknown word label, and model1 to a semantic model similarity vector for a known word label "1". In the final stage of analysis, there are two decoding alternatives given by (i) Pearson(brainA,model2)=.9 and Pearson(brainB,model1)=.9, which when summed gives 1.8; (ii) Pearson(brainA,model1)=.89, Pearson(brainB,model2)=.91. Here the sum is also 1.8 and therefore (i) and (ii) are identical. Applying the r to z transform would result in selection of (ii) because arctanh(.9)+arctanh(.9)=2.94, whereas arctanh(.89)+arctanh(.91)=2.95.

Statistical significance of decoding accuracy was determined by permutation testing. Decoding was repeated multiple times using the following procedure: creating a vector of word-label indices and randomly shuffling these indices; applying the vector of shuffled indices to reorder both rows and columns of only one of the similarity matrices (whilst keeping the original correct row/column labels so that word-labels now mismatch matrix contents); and repeating the entire pair-matching decoding procedure described above. If word labels are randomly assigned to similarity vectors, we expect a chance-level decoding accuracy of 50%. Repetition of this process (here 10,000 repeats) supplies a null distribution of decoding accuracies achieved by chance. The p-value of decoding accuracy is calculated as the proportion of chance accuracies that are greater than or equal to the observed decoding accuracy.

For permutation testing only, voxel selection was undertaken a single time, per participant, on all 70 words (rather than on 68/70 words in each leave-2-out decoding iteration). This was to reduce computation time that would otherwise have been prohibitive. This is very unlikely to have yielded any discernible difference in outcome. Unlike decoding strategies, that involve fitting a classification/encoding model to fMRI data (and are prone to fitting and subsequently decoding fMRI noise), similarity-based decoding does not learn a mapping between semantic-model and fMRI data and is robust to "double dipping" giving spurious decoding accuracies (see Kriegeskorte et al. (2009) for problems associated with double dipping).

As an empirical demonstration, we reran all of our 21 actual (non-permuted) model-based decoding analyses, that are reported later in Section 5.2, whilst selecting voxels from all 70 words (as opposed to leave-2-out voxel-selection on 68/70 words). Specifically, decoding analyses were repeated for all 7 model combinations, and tested first on all words, then for the most concrete words only, and finally the most abstract words only. Mean decoding accuracies for the 9 participants yielded with and without leave-2-out voxel selection were compared using paired t-tests. There were no significant differences across all 21 tests. The most different (non-significant) individual result was t=1.87, p=.09 (2-tailed), and in this case leave-2-out voxel selection gave the higher accuracy.

4.1 Model combination by ensemble averaging

To test whether the three different semantic models (image-based, Italian/English text-based) carried complementary information, we combined the models in evaluation, thus allowing us to test whether accuracies achieved using model combinations were higher than those achieved with isolated models.

To combine the different models, we used an ensemble averaging strategy and ran the similarity-based decoding analyses as described above in parallel with each of the three semantic models. At each leave-2-out test iteration, this gave three arctanh transformed 2*2 correlation matrices (one for each semantic model) that were used to evaluate decoding. Model combination was achieved by fusing the respective arctanh transformed correlation matrices by summing them together. Evaluation of the resulting 2×2 summation matrix proceeded as previously by first summing the two congruent values on the main-diagonal of the matrix, then summing the two incongruent scores on the counter-diagonal. If the congruent sum was greater than the incongruent sum, decoding was a success, otherwise a failure.

5 Results

We split the stimulus nouns into the 35 most concrete and 35 most abstract words according to the behavioural concreteness ratings from Anderson et al. (2014), and ran analyses on all words combined and these two subsets. Due to limitations in word coverage of the semantic models, 'melody' was missing from the abstract words, and 'skeleton-key' and 'police-station' were missing from the most concrete words (hence 67/70 words were analysed).

5.1 Hypotheses

Dual coding theory (Paivio, 1971) leads to the following hypotheses: (1) The text-based models will decode the more abstract nouns' neural activity patterns with higher accuracy than the image-based model; (2) both image and text-based models will decode the more concrete nouns' neural activity.

We also compared the decoding accuracy for the most concrete nouns achieved using the combined image- and text-based models to the unimodal models in isolation. Whilst previous analyses have observed advantages of multimodal models in describing concrete noun fMRI, it is not clear whether this effect will carry over to our noun data set. One reason is because many of the most concrete half are "less concrete" than those of previous studies: according to Brysbaert et al. (2014)'s concreteness norms (where words were rated on a scale from 1 to 5), the mean $\pm$ SD rating of the 60 concrete nouns analysed by Mitchell et al. (2008) (and subsequently by Anderson et al. (2015)) is 4.87$\pm$.12, whereas the mean $\pm$ SD of the "most concrete" nouns analysed in the current article, when tested with an independent samples t-test, was significantly smaller at 4.42$\pm$.44 ($t = 7.4$, $p < .0001$, 2-tail). A second reason is that the experimental task required participants to imagine a situation associated with the noun, rather than think of object properties. Therefore this analysis was of a more exploratory nature.

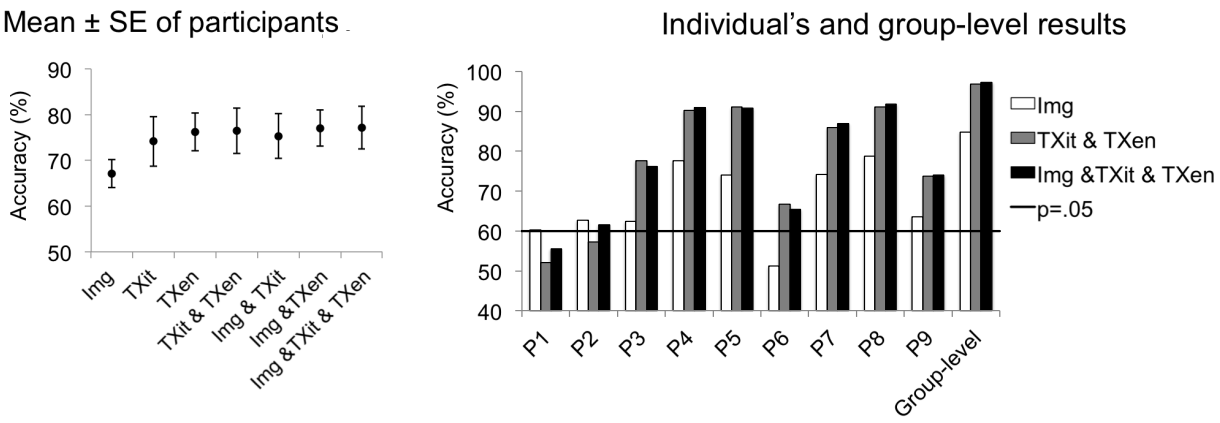
5.2 Decoding Analysis

Decoding analyses were run using the image-based model and Italian and English text-based models in isolation, and also all combinations of these models as described in Section 4. Results are in Figure 3. In this section we use the abbreviations **Img** for the image-based model and **TXit** and **TXen**, for the Italian and English text-based models, respectively.

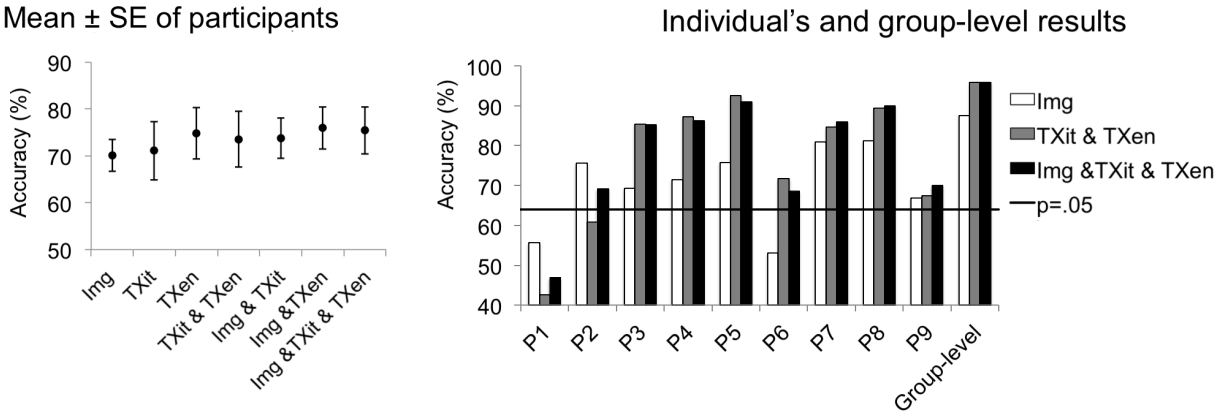
In all tests, chance-level decoding accuracy (the expected accuracy if word labelling is random) is 50%. Mean$\pm$SE accuracies across all participants are displayed in the leftmost column of plots for all 7 model combinations. Individual-level results are displayed for only three model combinations to avoid cluttering the graphs (Img only, the combined TXit&TXen, and the combined Img&TXit&TXen). To simplify the following discussion of results, we mainly focus on these three models. The choice to focus on TXit&TXen, rather than the Italian model, was made following the rationale that the language combination would leverage cultural nuances of semantic structure found in the Italian text-corpora jointly with the more extensive coverage of the larger English Wikipedia. Although TXit&TXen and TXen tended to produce higher decoding accuracies, there were no significant differences between either TXit or TXen tested in isolation, or any model combination incorporating them. Mean results are displayed for all model combinations in Figure 3 and key results are tabulated in Table 2.

5.3 An advantage for the textual model on abstract nouns

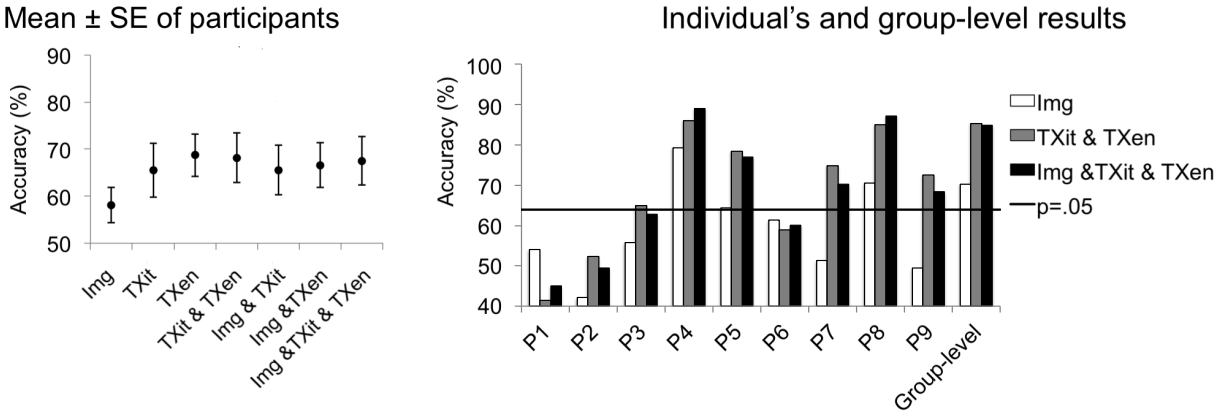
With respect to hypothesis 1 (an advantage for the text-based models decoding abstract neural activity patterns), the key difference to observe in Figure 3 is the drop in relative decoding accuracy between the image-based model and text-based models when decoding the most abstract nouns. The nine participant's mean decoding accuracies for the most abstract nouns were compared between the Img, TXit, TXen and TXit&TXen models using Repeated Measures ANOVA. Combinations of image and text-based models (e.g. Img&TXen) were not directly relevant to this analysis (because they integrate visual and textual data) and consequently these models were excluded. Bartlett's test was used to verify that there was no evidence against homogeneity of variances prior to analysis (χ^2=1.77, $p = .62$). The ANOVA indicated a statistically significant difference between models: F(3,24) = 5.06, $p < .01$. Post hoc comparisons conducted using the Tukey Honest Significant Difference (HSD) test revealed that decoding accuracies achieved using TXen and the



All 67 concrete and abstract words combined



33 most concrete words



34 most abstract words

Figure 3: Results of the decoding analysis from Section 5.2. See also Table 2. p=.05 lines were empirically estimated as described in Section 4 and apply to decoding an individual's fMRI data (not multiple individuals).

	All words combined	Most concrete	Most abstract
Img	67±3%, 7/9 (<.001)	70±3%, 7/9 (<.001)	58±4%, 2/9 (.07)
TXit&TXen	76±5%, 7/9 (<.001)	76±6%, 7/9 (<.001)	68±5%, 6/9 (<.001)
Img&TXit&TXen	77±5%, 8/9 (<.001)	77±5%, 8/9 (<.001)	68±5%, 5/9 (<.001)

Table 2: Key decoding accuracies from Section 5.2 (see also Figure 3). Each cell shows mean±SE decoding accuracy, the number (n) of participants decoded at a level significantly above chance (p<.05), and in round brackets, the cumulative binomial probability of achieving ≥ n significant results at p=.05.

TXit&TXen model were significantly different (and larger than) Img (both $p < .05$). There were no other significant differences (including between Img and TXit). One possible reason for the weaker performance of TXit than TXen is that Italian Wikipedia is a less rich source of information due to being smaller in size than English Wikipedia (despite it presumably containing semantic information that is more relevant to Italian culture).

5.4 Both image and text-based models decode the more concrete nouns

That both image- and text-based models significantly decoded the most concrete nouns is consistent with hypothesis 2. To test for differences between image- and text-based models, mean decoding accuracies for the nine participants on the most concrete nouns were compared for the Img, TXit, TXen and TXit&TXen models using Repeated Measures ANOVA. Combinations of image- and text-based models (e.g. Img&TXen) were not directly relevant to this analysis (because they integrate visual and textual data) and so these models were excluded. Bartlett's test was used to verify homogeneity of variances prior to analysis (χ^2 = 2.86, $p = .41$). The ANOVA detected no statistically significant differences between the models: F(3,24) = 1.56, $p = .22$. Therefore when decoding the most concrete nouns there was no significant difference in accuracy between image-based and any text-based model.

5.5 No overall advantage for multimodal models on the more concrete nouns

The third exploratory test compared the accuracy of the multimodal combination of image- and text-based models to the unimodal models when decoding the more concrete neural activity patterns.

For the most concrete words, the highest scoring combination across all models was Img&TXen (mean±SE=77±4%). Whilst this proved to be significantly greater than Img ($t = 3.13$, $p <= .02$, $df = 8$, 2-tail), it was not significantly greater than TXen ($t = .81$, $p = .44$, $df = 8$, 2-tail). Turning to the analogous case for the Italian models, Img&TXit (mean±SE=75±4%) was not significantly greater than Img ($t = 1.74$, $p = .12$, $df = 8$, 2-tail), or TXit ($t = 1.09$, $p = 0.31$, $df = 8$, 2-tail). Therefore, although multimodal combinations returned higher accuracies than either the image- and text-based models in isolation (for concrete words), decoding accuracy was not significantly higher than either image- or text-based models.

Previous work decoding neural activity associated with concrete nouns has found image-based models to supply complementary information to text-based models (Anderson et al., 2015). We suggest three reasons that image-based models may have been disadvantaged in the current study compared to these past analyses. Firstly, Anderson et al. focused on fMRI data elicited by unambiguously concrete nouns, whereas the experimental nouns analysed in the current article were mostly intended to be 'less than concrete' (of the seven taxonomic categories investigated only 'objects/tools' was designed to be unambiguously concrete). Secondly, Anderson et al. used more images to build noun representations (on average 350 images per noun compared to 20 used here), and nouns in the ImageNet images were segmented according to bounding boxes. Consequently their input may have been less noisy than Google Images (which we used because of its wider coverage). Finally, the experimental task of the previous analyses required participants to actively think about the properties of objects, whereas the current data set was elicited as participants imagined situ-

ations associated with nouns (and hence may have invoked neural representations with more contextual elements).

The lack of a significant increase in decoding accuracy achieved by pairing image- and text-based models allows us to infer that the text-based model contained many aspects of the visual semantic structure found in the image-based model. Of course we expect modal structure in text-based models commensurate with what people are inclined to report in writing; e.g., it is easy to convey in text that both bananas and lemons are yellow and curvy, and lightbulbs and pears have similar shapes. Therefore we would anticipate correspondences in semantic similarities between image and text-based models and for these correspondences to extend to match neural similarities, e.g., as induced by participants viewing pictures of objects (Carlson et al., 2014).

5.6 Group-level decoding analysis

The similarity-based decoding approach we have applied enables group-level neural representations to be built simply by taking the mean similarity matrix over participants. Values in the correlation matrix were r to z (arctanh) transformed prior to averaging, then the average values were back transformed to the original range using tanh. This was because averaging z-transformed values (and back transforming) tends to yield less biased estimates of the population value than averaging the raw coefficients (Silver and Dunlap, 1987). However, in the current analysis results obtained with z-transformation versus without it were virtually identical.

Building group-level representations by averaging correlation matrices side-steps potential problems surrounding the obvious alternative method of averaging data in fMRI space, where anatomical/functional differences between different peoples' brains may result in relatively similar activity patterns being spatially mismatched in the standardised fMRI space. The motivation behind building group-level neural representations is that we might expect these to better match the computational semantic models than individual-level data. This is because the models are also built at group-level, created from the photographs and text of many individuals. However building group-level neural representations will only be beneficial if there exist group-level commonalities in representational similarity (when combining data will reduce noise) as opposed to individual semantic representational schemes.

Accuracies achieved using models to decode the group-level neural similarity matrices are displayed in the final column of the bar charts at the right of Figure 3. Specifically, decoding accuracies were:

For all words combined: Img=84.8%, TXit&TXen=96.9% and Img&TXit&TXen=97.3%.

For the most concrete words: Img=87.5%, TXit&TXen=95.8% and Img&TXit&TXen=95.8%.

For the most abstract words: Img=70.2%, TXit&TXen=85.2% and Img&TXit&TXen=84.8%.

To statistically test whether group-level decoding accuracies surpassed those of the individual-level results, we compared the set of individual-level mean accuracies to the corresponding group-level mean accuracy using one sample t-tests. In all tests (see Table 3) the individual-level accuracies were significantly different (lower) than the group-level accuracy (corrected for multiple comparisons using false discovery rate (Benjamini and Hochberg, 1995)). This is indicative of group-level regularities in semantic similarity for both concrete and abstract nouns and also their combination.

A qualitative observation is that the differences between group and individual-level accuracy appear to be greater for concrete nouns. This could be consistent with participants having a more subjective semantic representation of abstract nouns; however we did not attempt to statistically test this claim. This is because a meaningful comparison would require concrete and abstract words to be controlled by being at least equally discriminable at individual level and this does not appear to be the case with this dataset.

6 Conclusion

This article has demonstrated that neural activity patterns elicited in mental situations of abstract nouns can be decoded using text-based computational semantic models, thus demonstrating that computational semantic models can make a contribution to interpreting the semantic structure of neural activity patterns associated with abstract nouns. Furthermore, by comparing how well visually grounded and textual semantic models de-

	All words combined	Most concrete	Most abstract
Img	-5.6 (.004)	-5.2 (.004)	-3.0 (.02)
TXit&TXen	-4.2 (.007)	-3.6 (.010)	-3.4 (.01)
Img&TXit&TXen	-4.4 (.007)	-3.9 (.008)	-3.4 (.01)

Table 3: Results of one sample t-tests comparing the set of individual-level mean decoding accuracies to the group-level accuracy (see Section 5.6). All tests were 2-tailed with df=8. The first number in each cell is the t-statistic, the second number in round brackets is the p-value (corrected according to false discovery rate).

code brain activity associated with concrete or abstract nouns, we have observed a selective advantage for textual over visual models in decoding the more abstract nouns. This has therefore provided initial model-based brain decoding evidence that is broadly in line with the predictions of dual coding theory (Paivio, 1971). However, results should be interpreted in light of the following two factors.

First, the dataset analysed was for a small sample of 67 words, and it is reasonable to conjecture that some of these words are also encoded in modalities other than vision and language. For example, musical words may be encoded in acoustic and motor features (see also Fernandino et al. (2015)). Future work will be necessary to verify that the findings generalise more broadly to words from domains beyond law and music. In work in progress the authors are undertaking more focused analyses on the current dataset, using textual, visual and newly developed audio semantic modes (Kiela and Clark, 2015) to tease apart linguistic, visual and acoustic contributions to semantic representation and how these vary throughout different regions of the brain.

A second limitation of the current approach, as pointed out by a reviewer, is that the Google image search algorithm (the workings of which are unknown to the authors) may not perform as well for abstract words as it does for concrete words. Consequently, the visual model may have been handicapped compared to the textual model when decoding neural representations associated with more abstract words. We have no current measure of the degree of this effect, but it may be possible to alleviate it in future work, by having participants manually select images that they associate with abstract stimulus words, and using computational representations derived from these images in the analysis.

Secondary results are that we have exploited representational similarity space to build group-level neural representations which better match our inherently group-level computational semantic models. In so doing, this exposes group-level commonalities in neural representation for both concrete and abstract words. Such group-level representations may prove both a useful test-bed for evaluating computational semantic models, as well as a potentially useful information source to incorporate into computational models (see Fyshe et al. (2014) for related work).

Finally we have demonstrated that English and Italian text-based models are roughly interchangeable in our neural decoding task. That the English text-based model tended to return marginally higher results on our Italian brain data than the Italian model provides a cautionary note for future studies wishing to use semantic models from different languages to identify culturally specific aspects of neural semantic representation e.g., as a follow up to Zinszer et al. (2016). However we also note that the English Wikipedia data was larger than the corresponding Italian corpus.

Acknowledgments

We thank three anonymous reviewers for their insightful comments and suggestions, Brian Murphy for his involvement in the configuration, collection and preprocessing of the original dataset, and Marco Baroni and Elia Bruni for early conversations on some of the ideas presented. Stephen Clark is supported by ERC Starting Grant DisCoTex (306920).

References

A. J. Anderson, E. Bruni, U. Bordignon, M. Poesio, and M. Baroni. 2013. Of words, eyes and brains: Correlating image-based distributional semantic models with

neural representations of concepts. In *Proceedings of EMNLP*, pages 1960–1970, Seattle, WA.

A. J. Anderson, B. Murphy, and M. Poesio. 2014. Discriminating taxonomic categories and domains in mental simulations of concepts of varying concreteness. *J. Cognitive Neuroscience*, 26(3):658–681.

A. J. Anderson, E. Bruni, A. Lopopolo, M. Poesio, and M. Baroni. 2015. Reading visually embodied meaning from the brain: Visually grounded computational models decode visual-object mental imagery induced by written text. *NeuroImage*, 120:309–322.

A. J. Anderson, B. D Zinszer, and R. D. S. Raizada. 2016. Representational similarity encoding for fMRI: Pattern-based synthesis to predict brain activity using stimulus-model-similarities. *NeuroImage*, 128:44–53.

M. Andrews, G. Vigliocco, and D. Vinson. 2009. Integrating experiential and distributional data to learn semantic representations. *Psychological Review*, 116(3):463–498.

L. Barca, C. Burani, and L. S. Arduino. 2002. Word naming times and psycholinguistic norms for Italian nouns. *Behavior Research Methods, Instruments, & Computers*, 34:424–434.

Y. Benjamini and Y. Hochberg. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B (Methodological)*, 57(1):289–300.

L. Bentivogli, P. Forner, B. Magnini, and E. Pianta. 2004. Revising the WordNet Domains Hierarchy: Semantics, coverage, and balancing. In *Proceedings of the Workshop on Multilingual Linguistic Resources*, pages 101–108, Geneva, Switzerland.

S. Bergsma and B. Van Durme. 2011. Learning bilingual lexicons using the visual similarity of labeled web images. In *IJCAI*, pages 1764–1769.

R. Bruffaerts, P. Dupont, R. Peeters, S. De Deyne, G. Storms, and R. Vandenberghe. 2013. Similarity of fMRI activity patterns in left perirhinal cortex reflects similarity between words. *J. Neuroscience*, 33(47):18597–18607.

E. Bruni, N. K. Tran, and M. Baroni. 2014. Multimodal distributional semantics. *Journal of Artifical Intelligence Research*, 49:1–47.

M. Brysbaert, A. B. Warriner, and V. Kuperman. 2014. Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior research methods*, 46(3):904–911.

T. A. Carlson, R.A. Simmons, and N. Kriegeskorte. 2014. The emergence of semantic meaning in the ventral temporal pathway. *J. Cognitive Neuroscience*, 26(1):120–131.

K. M. Chang, T. M. Mitchell, and M. A. Just. 2010. Quantitative modeling of the neural representations of objects: How semantic feature norms can account for fMRI activation. *NeuroImage: Special Issue on Multivariate Decoding and Brain Reading*, 56:716–727.

K. Darwish. 2013. Named entity recognition using cross-lingual resources: Arabic as an example. In *Proc. ACL*, pages 1558–1567.

B. Devereux, C. Kelly, and A. Korhonen. 2010. Using fMRI activation to conceptual stimuli to evaluate methods for extracting conceptual representations from corpora. In *Proceedings of the NAACL HLT First Workshop on Computational Neurolinguistics*, pages 70–78, Los Angeles, USA.

C. Fellbaum, editor. 1998. *WordNet: An Electronic Database*. MIT Press, Cambridge, MA.

R. Fergus, L. Fei-Fei, P. Perona, and A. Zisserman. 2005. Learning object categories from Google's image search. In *ICCV*, pages 1816–1823.

L. Fernandino, C. J. Humphries, M. S. Seidenberg, W. L. Gross, L. L. Conant, and J. R. Binder. 2015. Prediction of brain activation patterns associated with individual lexical concepts based on five sensory-motor attributes. *Neuropsycholigia. doi:10.1016/j.neuropsychologia.2015.04.009*.

A. Fyshe, P. P. Talukdar, B. Murphy, and T. M. Mitchell. 2014. Interpretable semantic vectors from a joint model of brain-and text-based meaning. In *Proceedings of ACL*, pages 489–499, Baltimore, MD.

Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. B. Girshick, S. Guadarrama, and T. Darrell. 2014. Caffe: Convolutional architecture for fast feature embedding. In *ACM Multimedia*, pages 675–678.

D. Kiela and L. Bottou. 2014. Learning image embeddings using convolutional neural networks for improved multi-modal semantics. In *Proceedings of EMNLP*, pages 36–45, Doha, Qatar.

D. Kiela and S. Clark. 2015. Multi- and cross-modal semantics beyond vision: Grounding in auditory perception. In *Proceedings of the Empirical Methods in Natural Language Processing Conference (EMNLP 2015)*, pages 2461–2470, Lisbon, Portugal.

D. Kiela, F. Hill, A. Korhonen, and S. Clark. 2014. Improving multi-modal representations using image dispersion: Why less is sometimes more. In *Proceedings of ACL 2014*.

N. Kriegeskorte, W. K. Simmons, P. S. F. Bellgowan, and C. I. Baker. 2009. Circular analysis in systems neuroscience: The dangers of double dipping. *Nature Neuroscience*, 12:535–540.

N. Kriegeskorte. 2015. Deep neural networks: A new framework for modeling biological vision and brain information processing. *Annual Review of Vision Science*, 1:417–446.

T. Mikolov, K. Chen, G. Corrado, and J. Dean. 2013. Efficient estimation of word representations in vector space. In *Proceedings of ICLR*, Scottsdale, Arizona, USA.

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

B. Murphy, P. Talukdar, and T. Mitchell. 2012. Selecting corpus-semantic models for neurolinguistic decoding. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics (*SEM)*, pages 114–123, Montreal, Canada.

A Paivio, editor. 1971. *Imagery and verbal processes*. Holt, Rinehart, and Winston, New York.

M. Palatucci, D. Pomerleau, G. Hinton, and T. Mitchell. 2009. Zero-shot learning with semantic output codes. *Neural Information Processing Systems*, 22:1410–1418.

F. Pereira, M. Botvinick, and G. Detre. 2013. Using Wikipedia to learn semantic feature representations of concrete concepts in neuroimaging experiments. *Artif. Intell.*, 194:240–252.

A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson. 2014. CNN features off-the-shelf: An astounding baseline for recognition. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops 2014*, pages 512–519.

A. E. Richman and P. Schone. 2008. Mining wiki resources for multilingual named entity recognition. In *Proc. ACL*.

L. Shi, R. Mihalcea, and M. Tian. 2010. Cross-language text classification by model translation and semi-supervised learning. In *Proc. EMNLP*.

N. C. Silver and W. P. Dunlap. 1987. Averaging correlation coefficients: Should Fisher's z transformation be used? *J. Applied Psychology*, 72(1):146–148.

J. Sivic and A. Zisserman. 2003. Video google: A text retrieval approach to object matching in videos. In *ICCV*, pages 1470–1477.

G. Sudre, D. Pomerleau, M. Palatucci, L. Wehbe, A. Fyshe, R. Salmelin, and T. Mitchell. 2012. Tracking neural coding of perceptual and semantic features of concrete nouns. *NeuroImage*, 62:451–463.

K. Wiemer-Hastings and X. Xu. 2005. Content differences for abstract and concrete concepts. *Cognitive Science*, 29:719–736.

B. D. Zinszer, A. J. Anderson, O. Kang, T. Wheatley, and R. D. S. Raizada. 2016. Semantic structural alignment of neural representational spaces enables translation between English and Chinese words. *J. Cognitive Neuroscience*, 28(11):1749–1759.