Discriminating Taxonomic Categories and Domains in Mental Simulations of Concepts of Varying Concreteness

Andrew J. Anderson[1], Brian Murphy[2,4], and Massimo Poesio[1,3]

Abstract

■ Most studies of conceptual knowledge in the brain focus on a narrow range of concrete conceptual categories, rely on the researchers' intuitions about which object belongs to these categories, and assume a broadly taxonomic organization of knowledge. In this fMRI study, we focus on concepts with a variety of concreteness levels; we use a state of the art lexical resource (WordNet 3.1) as the source for a relatively large number of category distinctions and compare a taxonomic style of organization with a domain-based model (an example domain is Law). Participants mentally simulated situations associated with concepts when cued by text stimuli. Using multivariate pattern analysis, we find evidence that all Taxonomic categories and Domains can be distinguished from fMRI data and also observe a clear concreteness effect: *Tools* and *Locations* can be reliably predicted for unseen participants, but less concrete categories (e.g., *Attributes*, *Communications*, *Events*, *Social Roles*) can only be reliably discriminated within participants. A second concreteness effect relates to the interaction of Domain and Taxonomic category membership: Domain (e.g., relation to Law vs. Music) can be better predicted for less concrete categories. We repeated the analysis within anatomical regions, observing discrimination between all/most categories in the left mid occipital and left mid temporal gyri, and more specialized discrimination for concrete categories *Tool* and *Location* in the left precentral and fusiform gyri, respectively. Highly concrete/abstract Taxonomic categories and Domain were segregated in frontal regions. We conclude that both Taxonomic and Domain class distinctions are relevant for interpreting neural structuring of concrete and abstract concepts. ■

INTRODUCTION

Data about the organization of conceptual knowledge in the brain coming from patients with semantic deficits (Mahon & Caramazza, 2011; Patterson, Nestor, & Rogers, 2007; Damasio, Tranel, Grabowski, Adolphs, & Damasio, 2004; Vinson, Vigliocco, Cappa, & Siri, 2003; Caramazza & Shelton, 1998; Warrington & Shallice, 1984) or collected from healthy patients using fMRI (Malach, Levy, & Hasson, 2002; Martin & Chao, 2001; Ishai, Ungerleider, Martin, Schouten, & Haxby, 1999) have proven an essential source of evidence for our understanding of conceptual representations, particularly when analyzed using machine learning methods (e.g., Connolly et al., 2012; Chang, Mitchell, & Just, 2010; Just, Cherkassky, Aryal, & Mitchell, 2010; Hanson & Halchenko, 2008; Kriegeskorte, Mur, & Bandettini, 2008; Kriegeskorte, Mur, Ruff, et al., 2008; Mitchell et al., 2008; Shinkareva, Mason, Malave, Wang, & Mitchell, 2008; Kamitani & Tong, 2005; O'Toole, Jiang, Abdi, & Haxby, 2005; Hanson, Matsuka, & Haxby, 2004; Haxby et al., 2001). Most of this work, however, has focused on a narrow range of conceptual categories, primarily concrete concepts such as animals, plants, tools, etc. (i.e., only a small percentage of the range of conceptual categories that form human knowledge). Although there is a substantial body of work investigating the representation of verbs and actions (e.g., Papeo, Rumiati, Cecchetto, & Tomasino, 2012; Peelen, Romagno, & Caramazza, 2012; Tomasino, Ceschia, Fabbro, & Skrap, 2012), until recently only a few studies went beyond concrete concepts to study the representation in the brain of concepts such as *law* or *freedom* (Wilson-Mendenhall, Kyle Simmons, Martin, & Barsalou, 2013; Kranjec, Cardillo, Schmidt, Lehet, & Chatterjee, 2012; Quadflieg et al., 2011; Binder, Westbury, McKiernan, Possing, & Medler, 2005; Friederici, Steinhauer, & Pfeifer, 2002; Grossman et al., 2002; Jessen et al., 2000). Interest has grown recently; for example, some studies have shown that fMRI data contain sufficient information to discriminate between concrete and nonconcrete concepts (Vigliocco et al., 2013; Wang, Baucom, & Shinkareva, 2012; Binder, Desai, Graves, & Conant, 2009; Binder et al., 2005). However, meta-analyses such as Wang, Conder, Blitzer, and Shinkareva (2010) and also Wilson-Mendenhall et al. (2013) showed that fairly different results are obtained depending on the types of nonconcrete concepts under study and that the range of nonconcrete concepts considered remains fairly narrow. The first objective of the

[1]University of Trento, [2]Carnegie Mellon University, [3]University of Essex, [4]Queen's University, Belfast

Journal of Cognitive Neuroscience 26:3, pp. 658–681
doi:10.1162/jocn_a_00508

present work is therefore to broaden the range of nonconcrete concepts under study and to analyze more carefully the effect of concept category.

This type of analysis is however complicated by the fact that the representation and organization of human knowledge about nonconcrete conceptual categories is much less understood than in the case of concrete concepts. Human intuitions about nonconcrete concepts are not very sharp, for example, studies asking participants to specify the defining characteristics of nonconcrete concepts find that this task is much harder than for concrete ones (Wiemer-Hastings & Xu, 2005; McRae & Cree, 2002; Hampton, 1981). On the theoretical side, as well, there is not much agreement on nonconcrete concepts among psychologists, (computational) linguists, philosophers, and other cognitive scientists who proposed theories about the organization of conceptual knowledge. Just about the only point of agreement among such proposals is the need to go beyond the dichotomy "concrete concept"/"abstract concept": human conceptual knowledge includes a great variety of nonconcrete categories of varying degrees of nonconcreteness ranging from knowledge about space and time (e.g., *day*, *country*) to knowledge about actions and events (e.g., *concert*, *robbery*), to knowledge about inner states including emotions (*fear*) and cognitive states (*belief*), to purely abstract concepts (e.g., *art*, *jazz*, *law*). It is also known that many of these categories have their own distinct representation in memory (Binder & Desai, 2011). (These considerations also challenge the notion that concreteness is a matter of degree, an issue that is very relevant to this study—see also Connell & Lynott, 2012.) But there is a lot of disagreement among exactly which categories these different types of nonconcrete concepts belong to (e.g., which category does the concept *law* belong to). These disagreements are reflected by the major differences one can find in the way nonconcrete conceptual knowledge is organized in the large-scale repositories of conceptual knowledge that have appeared in the last 20 years, such as WordNet (Fellbaum, 1998), CYC (Lenat & Guha, 1990), and DOLCE (Gangemi et al., 2002). For instance, in WordNet, the top category "abstract concept" covers attributes, events and actions, temporal entities, and highly abstract concepts such as "law" both in the sense of "collection of all laws" and in the sense of "area of study," whereas locations are considered concrete concepts. In DOLCE, actions and events, attributes, and highly abstract concepts such as propositions are treated as completely unrelated conceptual categories, whereas both temporal and spatial locations are included in the *quality* category.

More fundamental objections have also been raised. Some researchers have questioned whether the traditional distinction among categories each representing objects of different types, originally developed for concrete concepts, is applicable to nonconcrete ones as well. Gentner (1981), Hampton (1981), and others found that, unlike concrete concepts, nonconcrete concepts are mostly characterized in terms of relations to other entities present in a situation. Wiemer-Hastings and Xu (2005) provided further support for this finding and proposed that abstract concepts are "anchored in situations" (Wiemer-Hastings & Xu, 2005, p. 731); in a similar fashion, Barsalou (1999) argued that the representation of abstract concepts is "framed by abstract event sequences." This suggests what we will call here a scenario-based organization for nonconcrete concepts. In this type of organization, nonconcrete concepts are not organized in memory by virtue of their similarity with other concepts of the same "type" (e.g., court and theater are types of location, and judge and musician are types of social role) but in terms of the scenarios in which they play a role. For instance, according to these theories, the conceptual representation of *justice* would not be determined by the fact that it belongs to the same type as other nonconcrete concepts, whatever that type may be, but by the fact that it plays a key role in *law* scenarios, and it is related to other concepts and entities in those scenarios (along with lawyers, court rooms, evidence and verdicts, etc). Or to make another example, the conceptual representation of *jazz* would be determined by its relations to other concepts such as saxophones, guitarists, and songs in *Music* scenarios. Thus, a second key objective of this study was to compare two types of organization for nonconcrete concepts:

- The traditional taxonomic organization: As we are going beyond the repertoire of "uncontroversial" concepts commonly studied in cognitive work on concepts, we could not simply rely on our intuitions concerning their categorization; instead, we followed the distinctions made in the WordNet lexical database (Fellbaum, 1998), release 3.1, at present the largest-scale and most widely used repository of conceptual knowledge.
- A scenario-based organization: Unfortunately, there is no lexical database specifying the scenarios concepts belong to. There is, however, a resource called WordNet Domain (Bentivogli, Forner, Magnini, & Pianta, 2004) that specifies the domains a concept belongs to, e.g., that concept *judge* belongs to the *Law* domain, whereas concept *clarinet* belongs to the *Music* domain. In this first study, we used domains as an approximation of scenarios; we will therefore call the organization we compared to taxonomic organization "domain-based organization."

To summarize, four fundamental questions about the organization of nonconcrete concepts in the mind were addressed in this study: (1) Can taxonomic distinctions and domain distinctions be distinguished from the fMRI data for concrete and nonconcrete concepts? (2) Is there a difference in classification accuracy between taxonomic organization and domain-based organization? (3) Are there commonalities in taxonomic/domain representation across participants? (4) How do taxonomic and domain distinctions interact?

These questions were targeted using a standard multivariate pattern analysis procedure, where a classifier is trained to predict the class membership of unseen fMRI data. We used as stimuli concepts belonging to seven distinct taxonomic categories defined in WordNet, ranging from concrete categories (*Tool*) to more abstract ones (*Location*, *Social Role*, *Event*, *Communication*, *Attribute*, and a category we called *Urabstract* of highly abstract words; see Materials section) and to two different domains in WordNet Domain (*Music* and *Law*). The stimuli were presented in the form of words on the screen. The experiment aimed to activate conceptual representations, so participants were asked to mentally simulate situations that exemplified each stimulus while their brain activity was recorded using fMRI. Multivariate pattern analysis was then used to determine if single stimulus trials could be classified according to their taxonomic category and domain, both within single sessions and across participants. In addition to testing whether nonconcrete concepts of various types could be discriminated using the same types of analysis successfully used with concrete concepts, we aimed to compare the relative ease of decoding taxonomic category versus domain, and so to determine which of these two types of organization is more central to the definition of nonconcrete concepts.

METHODS

Materials

We aimed to use as stimuli a list of words representative of the full range of nonconcrete concepts and also clearly associated with the two domains *Music* and *Law*. To identify the taxonomic categories, we started from the concreteness norms collected by Barca, Burani, and Arduino (2002). We selected the words with the lowest concreteness value. We then looked up these words in the Italian version of WordNet, MultiWordNet (Pianta, Bentivogli, & Girardi, 2002), to determine the taxonomic category of their dominant sense. In this way, we identified the six WordNet categories most commonly found with the "most abstract" words in the Barca et al. norms. These six categories are as follows:

- *Location*, defined in WordNet as "points or extents in space," and including concepts such as *court*, *jail*, and *theatre*. *Location* is considered as concrete objects in WordNet but belongs to a separate category of "qualities" in DOLCE and could therefore be considered concepts in between concrete and abstract.
- Four nonconcrete categories of arguably increasing levels of abstractness (see also Discussion section): *Event/action* ("something that happens at a given place and time"), *Communication* ("something that is communicated by or to or between groups," covering concepts such as *accusation*, *letter*, and *symphony*), *Attribute* ("a construct whereby objects or individuals can be distinguished"), and *Urabstract* (our own term for concepts such as *law* or *jazz*, which are classified as *abstract* in WordNet but do not belong to a clear subcategory of *abstract* such as *Event* or *Attribute*).
- Finally, the WordNet category *person, individual, someone, somebody, mortal*, a great many of whose hyponyms are what we may call *social roles* such as *judge* or *tenor*. Social roles occur frequently among the least concrete concepts in concreteness norms such as Barca et al.'s, which is a good reason to include them in this study, although it is not very clear whether they should be considered concrete or nonconcrete — and indeed, in so-called "generative" theories of the lexicon such as Pustejovsky's (1995), they are considered hybrids (the term used is dot objects). A second reason is their strong association with scenarios and also with domains, which makes their classification very relevant for this study.

In addition, one category of concrete concepts was selected: *Tool*.

As the original words from the Barca norm could not be used as they were for the most part highly ambiguous, the next step was to select 70 words whose unique or most preferred sense belonged to one of these seven categories and that were also representative of the two chosen domains: *Music* and *Law*. This was done using WordNet Domains (Bentivogli et al., 2004), a publically available resource in which every concept in WordNet is annotated with its domain. Lists of candidate words for each domain/taxonomy combination were obtained; from these lists, we eliminated all those words that were either too infrequent or polysemous with senses belonging to different categories. The result was a list of 70 stimuli, 10 stimuli per taxonomic category, of which 5 were classified in WordNet Domains as belonging to the *Music* domain, whereas 5 belonged to the *Law* domain. The full set of words is listed in Table 1.

Tests for Perceptual Confounds

To examine whether there were any perceptual differences between categories of stimulus words, we tested for differences between number of letters, number of phonemes, and number of syllables using two-way ANOVA. For the number of letters, we found no significant difference for Domain, $F(1, 56) = 2.09$, $p = .15$, a significant difference between Taxonomic Category, $F(6, 56) = 2.05$, $p = .03$, and no significant interaction between Domain × Taxonomic category, $F(1, 56) = 1.8$, $p = .12$. Post hoc t tests found only *Attribute* and *Urabstract* to be significantly different: *Urabstracts* have significantly fewer letters than *Attribute* ($t = -2.645$, $df = 18$, $p = .03$). For number of phonemes there were no significant differences between domains, $F(1, 56) = 2.25$, $p = .139$, between Taxonomic categories, $F(6, 56) = 2.08$, $p = .07$,

Table 1. Italian Stimulus Words and English Translations, Divided into Domains (Columns) and Taxonomic Categories (Groups of Five Rows)

Law		*Music*	
Attribute			
giurisdizione	jurisdiction	sonorita'	sonority
cittadinanza	citizenship	ritmo	rhythm
impunita'	impunity	melodia	melody
legalita'	legality	tonalita'	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			
arresto	arrest	concerto	concert
processo	trial	recital	recital
reato	crime	assolo	solo
furto	theft	festival	festival
assoluzione	acquittal	spettacolo	show
Social Role			
giudice	judge	musicista	musician
ladro	thief	cantante	singer
imputato	defendant	compositore	composer
testimone	witness	chitarrista	guitarist
avvocato	lawyer	tenore	tenor
Tool			
manette	handcuffs	violino	violin
toga	robe	tamburo	drum
manganello	truncheon	tromba	trumpet
cappio	noose	metronomo	metronome
grimaldello	skeleton key	radio	radio
Location			
tribunale	court/tribunal	palco	stage
carcere	prison	auditorium	auditorium
questura	police station	discoteca	disco
penitenziario	penitentiary	conservatorio	conservatory
patibolo	gallows	teatro	theater
Urabstracts			
giustizia	justice	musica	music
liberta'	liberty	blues	blues
legge	law	jazz	jazz
corruzione	corruption	canto	singing
refurtiva	loot	punk	punk

and interaction, $F(6, 56) = 1.55$, $p = .18$. Number of syllables did not differ between Domains, $F(1, 56) = 3.92$, $p > .05$, but did differ between Taxonomic categories, $F(6, 56) = 4.33$, $p = .0012$. Interaction was not significant, $F(6, 56) = 1.63$, $p = .155$. Post hoc comparisons using the Tukey HSD test revealed that *Urabstracts* differed from *Attribute*, *Communication*, *Location*, and *Social Role* by having fewer syllables. There were no other significant differences.

Word Norming

Norming ratings for the 70 stimulus words were collected following the procedure proposed by Barca et al. (2002) from 24 Italian native speakers (12 men, mean age = 28.8 years, $SD = 5.2$ years). No significant differences between Domain and Taxonomic Category and no significant interaction were found for *Familiarity* (Domain: $F(6, 56) = 2.23$ by a two-way ANOVA, $p < .1407$; Taxonomic Category: $F(6, 56) = 1.81$, $p < .1127$; Domain × Taxonomic Category: $F(6, 56) = 0.87$, $p = .5211$). The 70 words were generally rated as familiar. Following Barca et al., Concreteness was rated on a scale of 1 as *highly abstract* to 7 as *highly concrete*. Figure 1 plots histograms of participants' concreteness ratings per word, and Figure 2 plots the associated means and standard deviations for each stimulus word. In Figure 2, an increase in the spread of the histograms as the expected concreteness of the category decreases is clear. *Tools* and, to a slightly lesser extent, *Locations* were consistently rated as concrete, whereas ratings for the remaining five taxonomic categories became increasingly inconsistent; 20 words in these categories were rated as both *highly concrete* (7) and *highly abstract* (1). From Figure 2, it can be seen that mean ratings of *Social Roles* were consistently toward the concrete end of the scale; however, the considerable variability within the categories *Event*, *Communication*, *Attribute,* and *Urabstract* is clear. There was also a divide within these four categories between *Music* words and *Law* words: *Music* words tended to be rated as more concrete.

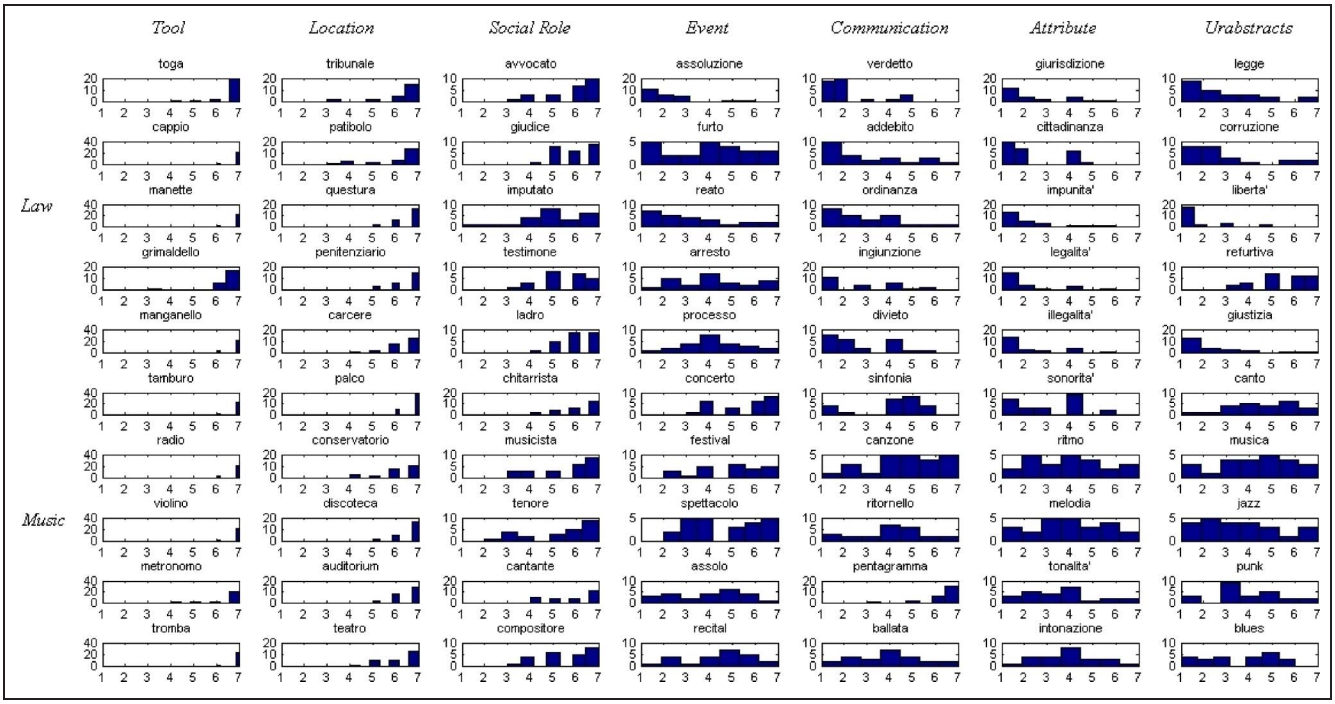

Figure 1. Per word histograms of concreteness ratings (24 participants, 7-point scale, 1 = *highly abstract*, 7 = *highly concrete*).

A two-way ANOVA found significant differences between Domains, $F(1, 56) = 35.15, p < .001$; Taxonomic categories, $F(6, 56) = 65.01$, $p < .001$; and a significant interaction between Domain and Taxonomic category, $F(6, 56) = 4.36$, $p = .001$. Table 2 displays the results of post hoc independent t tests between each of the 21 unique combinations of taxonomic pairs when measured on the 7-point scale. *Tool*, *Location*, and *Social Role* are all highly significantly different ($p < .001$) from each other and the other categories (*Tool* is more concrete than *Location*, which is more concrete than *Social Role*). Except for between *Attribute* and *Event*, combinations of *Attribute*, *Event*, *Communication*, *Urabstract* do not differ significantly in concreteness.

This raises two issues. First, the implications of the inconsistent ratings, which are reminiscent of the difficulties of getting participants to agree on the characteristic features of nonconcrete concepts (Wiemer-Hastings & Xu, 2005). We take this to be strong evidence that concepts are not organized in a "concreteness scale": humans can tell that, say, a *Tool* is more concrete than an *Attribute*, but have no clear intuition as to whether an *Attribute* like "sonority" is more or less concrete than another *Attribute* or a *Communication* concept such as "accusation." This introduces a question of whether it is indeed appropriate to measure concreteness on such a scale, and in light of this, we consider a simple alternative: to binarize the Likert scale to concrete = 1 and abstract = 0. We realize this by dividing the concreteness ratings according to a cutoff of 5 and below as abstract based on observations of Figure 2. As can be seen from the resulting histograms in Figure 3, this reduces ambiguity for *Communication*, *Attribute*, and *Urabstracts*, which are predominantly abstract; *Tool*, *Location* are consistently concrete; however, *Social Role* and *Event* remain ambiguous. Either way, there is some variability in the concreteness of concepts within some

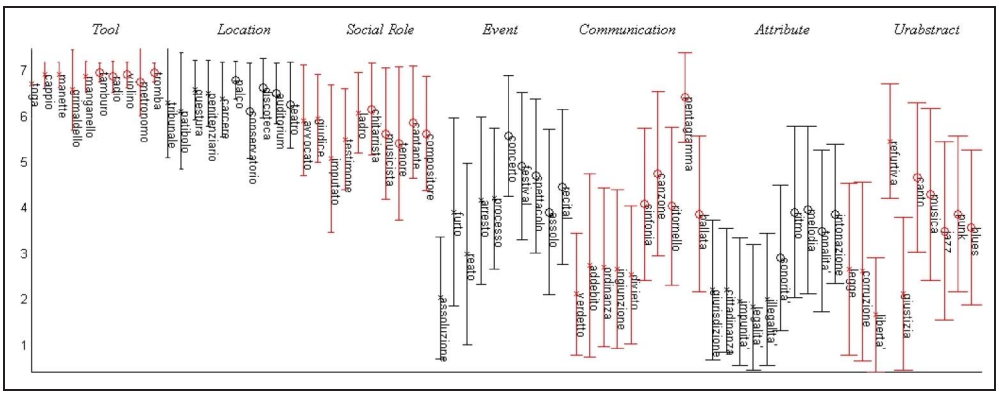

Figure 2. Per word mean and standard deviation concreteness ratings (24 participants, 1 = *highly abstract*, 7 = *highly concrete*). Circles are Law-related words, and Crosses are Music-related words.

Table 2. Results of Independent *t* Tests (Two-Tailed, *df* = 18) Testing for Differences in Concreteness Ratings (Scale [1 7]) between Each of the 21 Taxonomic Category Pairs

	Tool	*Location*	*Social Role*	*Event*	*Communication*	*Attribute*
Tool						
Location	$t = 4.2, p < .001$					
Social Role	$t = 9.6, p < .001$	$t = 5.7, p < .001$				
Event	$t = 14.8, p < .001$	$t = 11.1, p < .001$	$t = 5.7, p < .001$			
Communication	$t = 9.4, p < .001$	$t = 7.7, p < .001$	$t = 4.5, p < .001$	$t = 0.7, p = .5$		
Attribute	$t = 29.7, p < .001$	$t = 20.3, p < .001$	$t = 11, p < .001$	$t = 3.2, p = .005$	$t = 1.3, p = .21$	
Urabstract	$t = 16.0, p < .001$	$t = 12.3, p < .001$	$t = 6.8, p < .001$	$t = 1.2, p = .24$	$t = 0.2, p = .86$	$t = -1.7, p = .1$

categories. Given that the categories were selected on the basis of containing nonconcrete concepts and peoples' experience and estimations naturally differ (apparently widely), we consider this an inevitable consequence of such an experimental design. Therefore, to summarize, our investigation was of the neural relevance of taxonomic/domain categories, with our nonconcrete categories populated by concepts considered on average to be nonconcrete, and with our subsequent cross-validation analyses targeting category discrimination.

Second, the significant difference between domains, which is a nuisance for the interpretation of our classification results, in that domain meaning and concreteness are apparently confounded (if meaning and concreteness can indeed be considered to be separable; also discussed in Discussion section). We address this in Can taxonomic

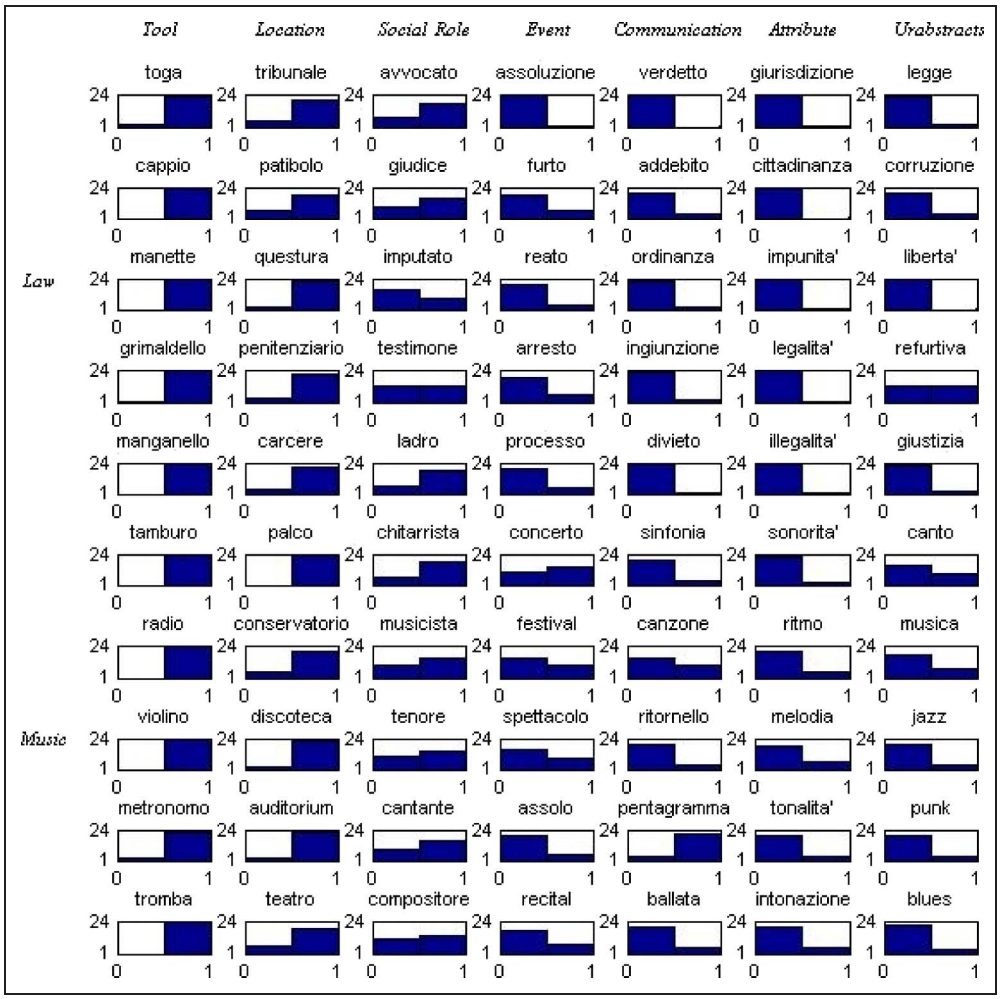

Figure 3. Per word histograms of binarized concreteness ratings (24 participants, 1 = *concrete*, 0 = *abstract*).

category-based and domain-based distinctions be recognized within participants? section by regressing out concreteness trends (as described either with the binary or Likert scale, which we consider dubious). In Which is more strongly encoded, taxonomic category distinctions or domain distinctions? section, where we have many taxonomic category pairs to choose from, we identify categories balanced in concreteness.

Participants

Seven right-handed native Italian speakers aged between 19 and 38 years (three women) were recruited to take part in the study. All had normal or corrected-to-normal vision. Participants received compensation of €15 per hour. The studies were conducted under the approval of the ethics committee at the University of Trento, and participants gave informed consent.

Data Acquisition

fMRI images were recorded on a 4T Bruker MedSpec MRI scanner at the neuroimaging (LNiF) labs of the Centre for Mind/Brain Sciences, University of Trento. An EPI pulse sequence with repetition time = 1000 msec, echo time = 33 msec, and 26° flip angle was used. A 64 × 64 acquisition matrix was used, and 17 slices were imaged with a between slice gap of 1 mm. Voxels had dimensions of 3 mm × 3 mm × 5 mm.

Experimental Paradigm

The names of the 70 concepts were presented to participants in the form of written words on the screen. Stimuli were displayed using bold Arial-Black size 20 font on a gray background. Each stimulus was presented five times, for a total of 350 trials, split in five blocks with the order of presentation being randomized in each block. Participants had the opportunity to pause between blocks, and the overall task time did not exceed 60 min. Each trial began with the presentation of a blank screen for 0.5 sec, followed by the stimulus word of dark gray on a light gray background for 3 sec, and a fixation cross for 6.5 sec. Participants were asked to keep still during the task and during breaks.

In experiments studying concrete concepts, the task of participants is often to think actively about the properties of the object named (see, e.g., Mitchell et al., 2008). However, eliciting properties has been found difficult with nonconcrete concepts (Wiemer-Hastings & Xu, 2005). On the other hand, participants to studies such as Wiemer-Hastings and Xu (2005) and Hampton (1981) appeared able to think of situations in which these concepts played a role and to produce situation-related objects. Our participants were therefore instructed to "think about situations that exemplify the concept the word refers to." The list of concept words was supplied to participants in advance of the experiment, so that they could prepare appropriate situations to simulate consistently.

Preprocessing

Preprocessing was undertaken using the Statistical Parametric Mapping software (SPM99, Wellcome Department of Cognitive Neurology, London, UK). The data were corrected for head motion, unwarped (to compensate for geometric distortions in the image interacting with motion), and spatially normalized to the Montreal Neurological Institute template image and resampled at 3 mm × 3 mm × 6 mm. Only voxels estimated to be gray matter were included in the subsequent analysis. For each participant, the data, per voxel, in each session (presentation cycle of 70 words) were corrected for linear trend and transformed to z scores.

A single volume was computed to represent each stimulus word 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).

Cross-validation Analysis Procedure

Broadly the same cross-validation procedure was followed for each of the analyses targeting the fundamental questions in the introduction. Specifics on variations in the procedure are indicated where relevant in the results. Input and target data pairs were partitioned into training and testing sets (using a leave-n-out approach) to support a number of cross-validation iterations. Target patterns were binary vectors with a single field set to one to uniquely specify the category (e.g., *Law* = [1 0] and *Music* = [0 1]). Input was a masked version of the fMRI gray matter data, retaining the 1000 most stable voxels in the training set according to the following procedure, similar to that used by Mitchell et al. (2008). For each voxel, the set of 70 words from each unique pair of scanning sessions in the training set were correlated, and the mean of the six resulting correlations (from the four scanning sessions used in training) was taken as the measure of stability. The 1000 voxels with highest mean correlations were selected for analysis.

Pattern classification used a single layer neural network with logistic activation functions (MATLAB 2009B, Mathworks, Neural Network toolbox). Weights and biases were initialized using the Nguyen-Widrow algorithm and training used conjugate gradient decent, continued until convergence, with performance evaluated using mean square error, with a goal of 10^{-4} or completion of 2000 training epochs. In each cross-validation iteration, the network was trained using the masked fMRI data and binary target codes in the training set and subsequently tested on the previously unseen masked fMRI data. The Euclidean distance between the network output vectors and target codes was computed, and the target code with the minimum distance was selected as the network output.

Our taxonomic category discrimination is a seven-way classification task, and until recently, the available methods for testing significance in the case of multiclass problems were not entirely satisfactory. Binomial tests are often used to test whether a classifier is predicting randomly (which is most informative when there are two-classes), but their application is severely limited in the multiclass case, because they do not tell us whether the classifier is capable of distinguishing between all test categories or just between subsets of categories. Motivated by these concerns and drawing from the statistical literature on contingency tables, Olivetti, Greiner, and Avesani (2012) developed a test in which Bayesian hypothesis testing techniques are used to estimate the posterior probability of each possible partitioning of distinguishable subsets of test classes in a multiclass problem. Consider for instance a problem in which items in the test set belong to the three classes: Classes 1, 2, and 3. There are five possible partitions. The classifier may be able to distinguish all three classes ([1][2][3]). Alternatively, the classifier may only be able of partial discrimination, as in the partitions [1,2][3] (i.e., the classifier is unable to discriminate between Classes 1 and 2, but can discriminate either 1 or 2 from 3), [1,3][2] (it is unable to distinguish between Classes 1 and 3), and [1][2,3]. Finally, the classifier may be unable to discriminate between any of the classes [1,2,3]. Olivetti et al.'s method assigns a posterior probability to each of these partitions, which can then be used to decide which of the interpretations of the confusion matrix are most likely, where, as a rule of thumb, a probability in excess of $1/K$, where K is the number of hypotheses (i.e., $K = 5$ in the three-class example just discussed), would be seen as informative evidence. We interpret our discrimination results using Olivetti et al.'s technique and Binomial tests.

RESULTS

Can Taxonomic Category-based and Domain-based Distinctions Be Recognized within Participants?

Leave-session-out cross-validation analyses were undertaken for each participant to recognize taxonomic and domain distinctions from the fMRI data. There were five scanning sessions; therefore, training in each of the five cross-validation iterations was on 280 volumes (four replicates of each of the 70 stimulus words) and testing was on the remaining 70 words. In each iteration, the 1000 most stable voxels were selected (see Cross validation analysis procedure section), and to give an impression of the spread of their locations across the cortex, all voxels that contributed to at least two of the five cross-validation iterations were identified per participant. Each voxel was linked to its anatomical region according to the automated labeling of Tzourio-Mazoyer et al. (2002), and a count was made of the number of voxels belonging to each unique region contributing to the analysis. The union of all anatomical regions across participants was taken, and voxel counts per region per participant were summed. The breakdown of number of voxels per anatomical region across participants is in Table 3. A discussion of the possible functional contribution of the different regions is in Discussion section. Figures 4 and 5 show confusion matrices averaging results across all seven participants (and cross-validation iterations within participant) for taxonomic category and domain, respectively. Both confusion matrices have an overlay that separates the taxonomic predictions by *Music* and *Law* components and domain predictions by taxonomic components.

Can Taxonomic Distinctions Be Recognized within Participants?

Mean classification accuracy for the seven-way taxonomic distinctions was ~.3, with chance level at .143. Accuracy was greatest for *Location*, *Tool*, and *Attribute*, and there is a visible diagonal in Figure 4, suggesting all classes can be discriminated. Overall 730/2450 correct classifications were observed, and the probability of achieving this by chance is $p = 2.2 \times 10^{-16}$ (two-tailed Binomial test). Applying Olivetti et al.'s (2012) test to the taxonomic confusion matrix in Figure 4 and sorting all subset partitions in descending order of posterior probability, we find that easily the top ranking partition hypothesis (posterior probability = 0.93) is the one according to which all seven test classes can be discriminated. The highest-ranked three partitions are shown in the caption of Figure 4. (The posterior probabilities rapidly diminish in the remaining 874 partitions that are not displayed; however, any posterior probability $>1/877 = 0.0011$ is regarded as informative. The full list is in the supplementary materials for Figure 4.) *Tool*, *Location*, and *Attributes* are most clearly distinguished, whereas prediction of taxonomic category is weakest for medium-ranked categories on the concreteness scale (*Event* and *Communication*). Indeed in the second partition of Olivetti et al.'s (2012) analysis, these categories aggregate.

Can Domains Be Predicted within Participants?

Mean classification accuracy over participants and cross-validation iterations for the two-way domain distinction was ~.7, with chance level at .5. In total there were 1702/2450 correct classifications, and the probability of achieving this by chance is $p = 2.2 \times 10^{-16}$ (two-tailed Binomial test). The results with Olivetti et al.'s (2012) statistics, presented in detail in the caption of Figure 5, strongly suggest that *Law* and *Music* can be discriminated (posterior probability > .99). Domains are relatively weakly discriminated for *Location* words; the accuracy for *Law*/*Tool* is also comparatively low. These differences between classification strength for the Domain distinction for different Taxonomic categories are further discussed

Table 3. Breakdown of the Number of Stable Voxels by Anatomical Region Contributing to Analysis in Can Taxonomic Category-based and Domain-based Distinctions Be Recognized within Participants? Section, Summed over Participants

	Proportion	*Cumulative*
Temporal_Mid_L	0.05	0.05
Precuneus_L	0.04	0.09
Occipital_Mid_L	0.04	0.13
Precuneus_R	0.04	0.17
Parietal_Inf_L	0.04	0.21
Temporal_Mid_R	0.03	0.24
Frontal_Mid_L	0.03	0.27
Precentral_L	0.02	0.29
Calcarine_R	0.02	0.32
Frontal_Mid_R	0.02	0.34
Parietal_Sup_L	0.02	0.36
Angular_L	0.02	0.39
Postcentral_L	0.02	0.41
Calcarine_L	0.02	0.43
Frontal_Inf_Tri_L	0.02	0.45
Frontal_Sup_L	0.02	0.47
Temporal_Sup_L	0.02	0.49
Temporal_Sup_R	0.02	0.51
Occipital_Mid_R	0.02	0.53
Angular_R	0.02	0.55
Cingulum_Mid_L	0.02	0.56
Frontal_Sup_R	0.02	0.58
Frontal_Inf_Tri_R	0.01	0.59
Precentral_R	0.01	0.61
Lingual_R	0.01	0.62
Supp_Motor_Area_L	0.01	0.63
SupraMarginal_L	0.01	0.65
Postcentral_R	0.01	0.66
Cuneus_L	0.01	0.67
Frontal_Inf_Oper_L	0.01	0.68
Fusiform_R	0.01	0.70
Fusiform_L	0.01	0.71

Anatomical regions are ranked in descending order, and regions contributing the top 70% contribution are shown. In total, there were 7449 voxels over seven participants (the 1000 most stable voxels per participant was taken for each cross-validation iteration).

in Which is more strongly encoded, taxonomic category distinctions or domain distinctions? section.

To counteract the possible effects of differences in concreteness ratings between Domains (see Materials section), the above analysis was repeated after regressing out the concreteness trend from the fMRI data. Specifically for each participant, linear regression was used to estimate the relationship between the mean concreteness score per word and fMRI data for each voxel. This was repeated using mean scores derived from the binarization of the 7-point Likert scale discussed in Materials section and the Likert scale. The regression line was subsequently subtracted from the data, and the classification analysis was repeated. In the binary case, overall classification accuracy summed over participants was reduced from .69 to .62 (1527/2450 correct classifications, where $p = 2.2 \times 10^{-16}$ for a two-tailed Binomial test, and Olivetti et al.'s, 2012 posterior probability that classes can be distinguished is >.99, as opposed to $p = 6.67 \times 10^{-32}$ for the hypothesis that classes are indistinguishable). With the Likert scale, although results were still statistically significant, overall classification accuracy was substantially reduced from .69 to .575 (1408/2450 correct classifications, where $p = 1.495 \times 10^{-13}$ for a two-tailed Binomial test, and Olivetti et al.'s, 2012 posterior probability that classes can be distinguished is >.99, as opposed to $p = 3.60 \times 10^{-11}$ for the hypothesis that classes are indistinguishable). The reduction in accuracy is not surprising given the generally lower concreteness ratings for Law words. However, given the inconsistencies observed between participants' concreteness ratings, the trend removed from neural data may have been inappropriate (i.e., for some participants, the detrending may have removed signal associated with domain rather than concreteness).

Which Is More Strongly Encoded, Taxonomic Category Distinctions or Domain Distinctions?

Having established that both taxonomic categories and domains can be distinguished within participants, we may go on to question whether there are differences in the strength with which taxonomic category and domain are encoded within the words. In addition, we can explore whether such a difference (if any) bears any relationship to the degree of concreteness of the words (e.g., in Figure 5, Domain is relatively weakly predicted for *Location*).

As there are only two domains and seven taxonomic categories (and therefore a mismatch in the amount of respective training data available for each), the experimental design was balanced by selecting all unique taxonomic category pairs (e.g., *Attribute* vs. *Communication*, *Attribute* vs. *Event*, etc., giving 21 pairs of 10 words each in total) and running two classification analyses, first, distinguishing between domains and, second, between taxonomic categories on each of the 21 data sets. The 21 data sets

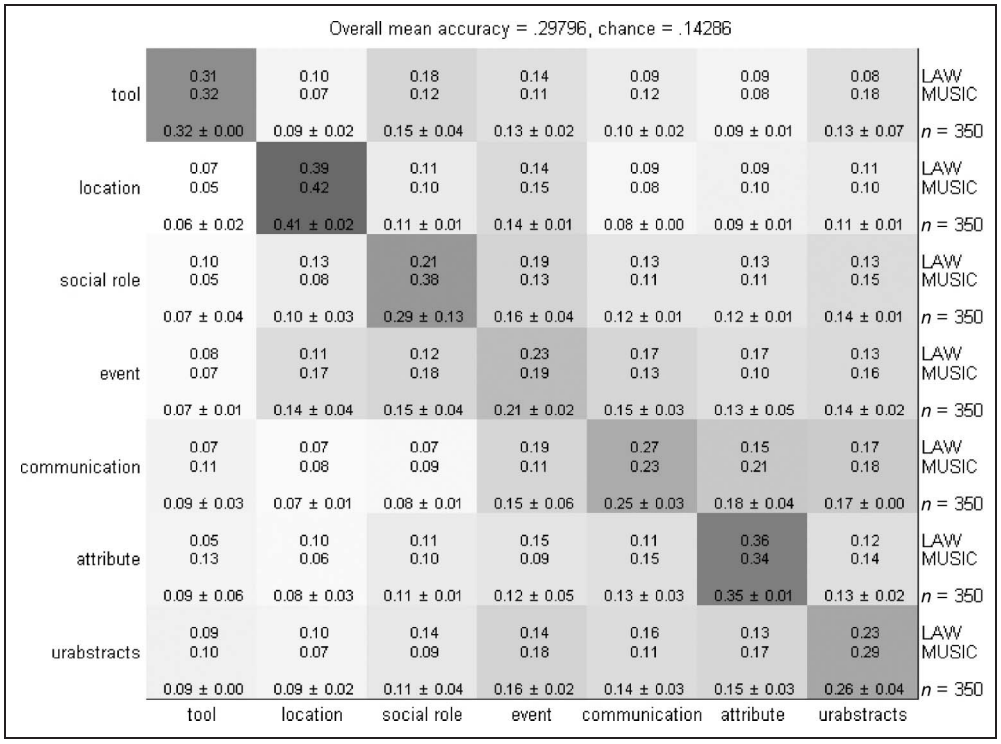

Figure 4. Leave-session-out Taxonomic category classification confusion matrix. Rows are the target labels, and columns are predictions. Numbers overlaid on each cell indicate the proportion of predictions per Law and Music, respectively (as indicated on the right *y* axis) for that row, averaging over seven participants. The numbers on the bottom line of each cell are the mean and standard deviation of predictions. Cell shading is scaled to the range 0–0.41 (0.41 is the maximum mean accuracy per cell displayed). Olivetti et al.'s (2012) test results [T = Tool] [L = Location] [S = Social Role] [E = Event] [C = Communication] [A = Attribute] [U = Urabstracts]: Partition 1: [[T] [L] [S] [E] [C] [A] [U]], postP: 0.93; Partition 2: [[T] [L] [S] [E,C] [A] [U]], postP: 0.04; Partition 3: [[T] [L] [S] [E,U] [C] [A]], postP: 0.02.

therefore contained 100 words: 10 words × 2 taxonomies × 5 replicates. Other than the reduced data set size, each of the 42 analyses followed the same protocol as Can taxonomic category-based and domain-based distinctions be recognized within participants? section (selection of the 1000 most stable voxels and leave-session-out cross-validation analysis).

Confusion matrices summed over participants for each of the 42 experiments (21 domain distinctions, 21 pairwise category distinctions) and subsequently tested according to Olivetti et al. (2012) can be found in the supplementary materials for Figure 6. In all analyses the strongest posterior probability was that the taxonomic categories/domains could be distinguished. The weakest classification

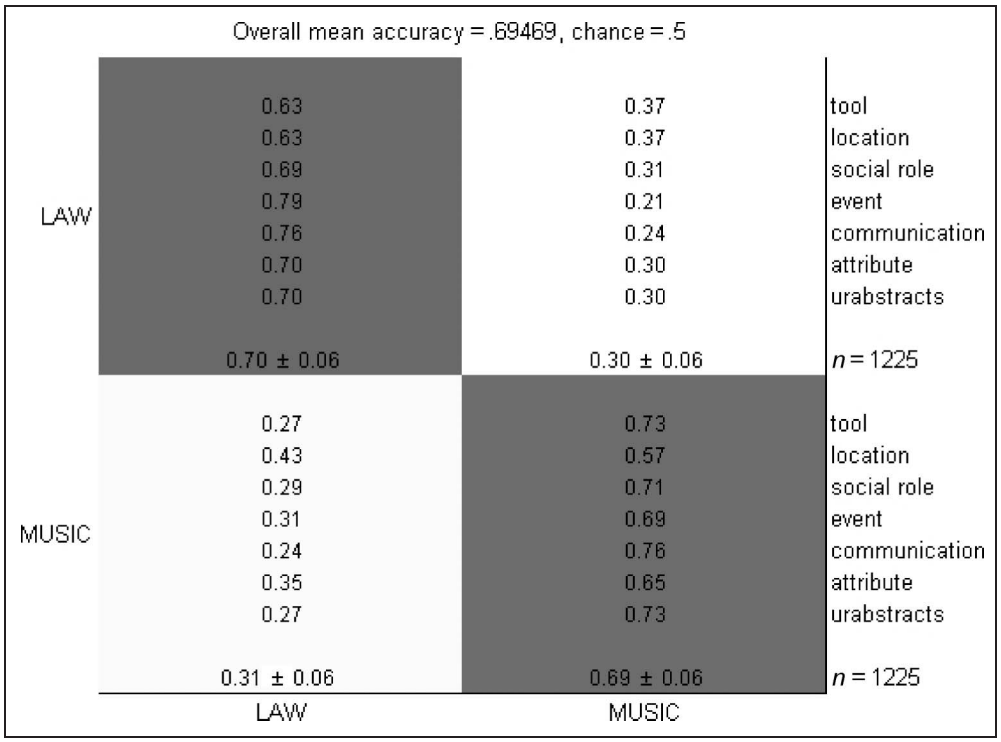

Figure 5. Leave-session-out Domain classification confusion matrix. Rows are the target labels, and columns are predictions. Numbers overlaid on each cell indicate the proportion of predictions per taxonomic category (as indicated on the right *y* axis) for that row, averaging over seven participants. The numbers on the bottom line of each cell are the mean and standard deviation of the predictions. Cell shading is scaled to the range 0–0.7 (0.7 is the maximum mean accuracy per cell displayed). Olivetti et al.'s (2012) test results [L=LAW] [M = MUSIC]: Partition 1: [[L] [M]], postP: >.99; Partition 2: [[L,M]], postP: 5×10^{-82}.

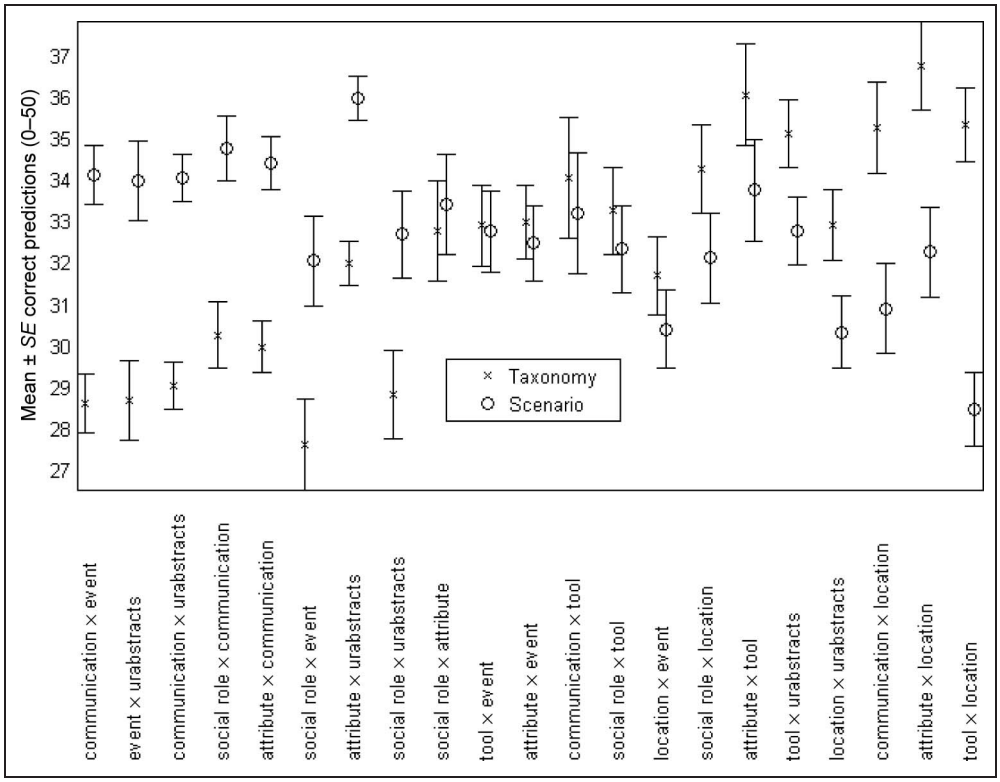

Figure 6. Mean ± *SE* classification accuracy per taxonomic test pair averaged across seven participants.

performance was Taxonomic classification between *Social Role* and *Event*, where there were a total of 387/700 correct classifications and the probability of achieving this at chance is $p = .0058$ (two-tailed Binomial test). Therefore, the probability of achieving the remaining stronger results at chance can be considered highly unlikely.

To test whether taxonomic category or domain could be more accurately predicted, for each participant, the mean accuracy per test (mean of the diagonal of each confusion matrix) was calculated, and the set of means was analyzed using two-way ANOVA. Main effects were Target (taxonomic category vs. domain distinction) and test set (the taxonomic category pair). There was no significant difference between domain and taxonomic category, $F(1, 252) = 1.07$, $p = .3$; there was however a significant difference between taxonomic test pair, $F(20, 252) = 15.44$, $p < .001$, and a significant interaction, $F(20, 252) = 2.8$, $p < .001$. The significant interaction is evidence that there are differences in the strength of Taxonomic category/Domain encoding between Taxonomic category pairs.

The results of post hoc *t* tests testing for difference in domain/taxonomic category classification accuracy in each test set are included in the supplementary materials for Figure 6. Figure 6 plots the mean ± standard error number of correct domain/taxonomic predictions per test over participants. Taxonomic category pairs on the *x* axis are sorted in ascending order of the difference between mean taxonomic category and domain classification accuracy. If differences in Taxonomic category/Domain encoding bear any relationship to concreteness, we would anticipate some distinction between the concreteness of categories visible in the rankings. It is clear that domains are more accurately predicted than taxonomic category in nonconcrete taxonomic category pairings (the left side of the plot) and that taxonomic distinctions tend to be relatively more accurate when one of the pair involves a concrete category. Distinction of domain is greatest between the two least concrete classes (*Attribute* and *Urabstract*). Distinction of taxonomic category is greatest between *Attribute* and *Location*, followed by the distinction between *Attribute* and *Tool*, then *Tool* and *Location*. The relative difference between classification of taxonomic category and domain is greatest for the two most concrete classes (*Tool* and *Location*).

To place the above analysis in the context of which taxonomic class pairs showed significant differences between Domains in concreteness ratings (e.g., pooling all *Tool* words and all *Attribute* words, testing for a significant difference in concreteness between Domains). Twenty-one independent samples *t* tests were computed on each taxonomic category pairing (18 degrees of freedom for each). Measuring concreteness on a binary scale (see Can taxonomic category-based and domain-based distinctions be recognized within participants? section), there were significant differences between *Event* and *Communication* ($t = -2.99$, $p = .008$), *Event* and *Attribute* ($t = -2.96$, $p = .008$), *Communication* and *Attribute* ($t = -2.82$, $p = .01$). Measuring concreteness on the more inconsistent Likert scale (see Materials section and Discussion section), 15/21 comparisons were not significantly different. All *t* statistics and associated *p* values for the Likert scale

tests are in Table 4. All significant differences were observed between pairings of the least concrete classes: *Attribute*, *Urabstracts*, *Communication*, and *Event*. There were, however, no significant differences in concreteness when these classes were paired with *Tool*, *Location*, or *Social Role*. In pairings of *Attribute*, *Urabstracts*, *Communication*, and *Event* with *Social Role*, Domain was more strongly classified. Therefore, adopting either concreteness measurement scale, differences in concreteness between Domains could have contributed to their discrimination in a selection of pairings of low concreteness classes.

In summary, (1) Domain and taxonomic category can be distinguished for each taxonomic category/domain combination, (2) Domain is most accurately distinguished for the least concrete taxonomic categories, and (3) Taxonomy category is most accurately distinguished when one or both categories in the taxonomic pair are concrete/near concrete.

Can Taxonomic Categories/Domains Be Predicted across Participants?

To test whether fMRI representations of taxonomic categories and domains share commonalities across participants, a leave-participant-out cross-validation protocol was adopted. For each participant, all data corresponding to each taxonomic versus domain subclass were averaged (e.g., all *Music/Tool* volumes were averaged across session and word, and all *Law/Tool* volumes were averaged likewise). As a result, 14 volumes per participant were obtained (7 taxonomic categories × 2 domain categories); in each of the seven cross-validation iterations, six participants were used for training and one for testing. Only voxels corresponding to gray matter in all of the spatially normalized volumes were selected for analysis. In contrast to the previous sections, selection of the 1000 most stable voxels was based on a correlation of Taxonomic category × Domain mean volumes (in the training set) rather than correlation of words. As there were six participants used in each training iteration, the 15 rather than 6 correlations (where $n = 14$ rather than 70) were averaged to give the stability score for each voxel.

Can Taxonomic Categories Be Predicted across Participants?

The leave-out-participant taxonomic category confusion matrix is in Figure 7. Mean classification accuracy was ~.37, with chance level at .143. The probability of achieving this result (36/98 correct classifications) at chance is $p = 3.039 \times 10^{-8}$ (two-tailed Binomial test). *Tool* and *Location* were accurately predicted; the other, less concrete, taxonomic categories, however, were conflated. This observation was supported by Olivetti et al.'s test: The results of the highest ranking five posterior probabilities are in the caption of Figure 7 (the remaining partitions are in the supplementary materials for Figure 7).

The highest posterior probability ($p = .1$) suggests that *Tool* and *Location* are distinct, and the remaining five less concrete classes conflate. It should be recognized that this probability is comparatively low compared with results in previous sections (previous seven-way classification results have had posterior probabilities >.9). This can be considered in part because of the lower sample size. We therefore examine also the next four ranking partitions that have posterior probabilities ranging from .032 to .068. There is consistency across partitions, *Location* is distinct in all five partitions and *Tool* is distinct in 4/5 partitions (in the second it forms a subset with *Communication*). The third ranking partition maintains *Social Role* as being distinct. Collating these results, it is reasonable to assume that the most concrete classes are distinguishable and the less concrete classes are confused; however, specifically which classes conflate is unclear.

Can Domains Be Recognized across Participants?

Mean domain classification accuracy was ~.7, with chance level at .5, in the leave-participant-out analyses.

Table 4. Results of Independent *t* Tests (Two-Tailed, $df = 18$) Testing for Differences in Concreteness Ratings (Scale [1 7]) between Domains, per Each of the 21 Taxonomic Category Pairs Corresponding to the Classification Analysis in Which Is More Strongly Encoded, Taxonomic Category Distinctions or Domain Distinctions? Section and Figure 6

	Tool	*Location*	*Social Role*	*Event*	*Communication*	*Attribute*
Tool						
Location	$t = -0.69, p = .50$					
Social Role	$t = -0.22, p = .83$	$t = -0.28, p = .78$				
Event	$t = -0.95, p = .35$	$t = -1.08, p = .29$	$t = -1.32, p = .20$			
Communication	$t = -1.30, p = .21$	$t = -1.45, p = .16$	$t = -1.73, p = .10$	$t = -4.60, p = .0002$		
Attribute	$t = -0.87, p = .40$	$t = -0.96, p = .35$	$t = -1.129, p = .27$	$t = -3.62, p < .0020$	$t = -5.89, p < .0001$	
Urabstract	$t = -0.67, p = .51$	$t = -0.74, p = .47$	$t = -0.85, p = .41$	$t = -2.68, p = .0152$	$t = -3.76, p = .0014$	$t = -3.53, p < .0024$

The top left cell corresponds to the difference in Domain between all 10 Tools and all 10 Locations pooled.

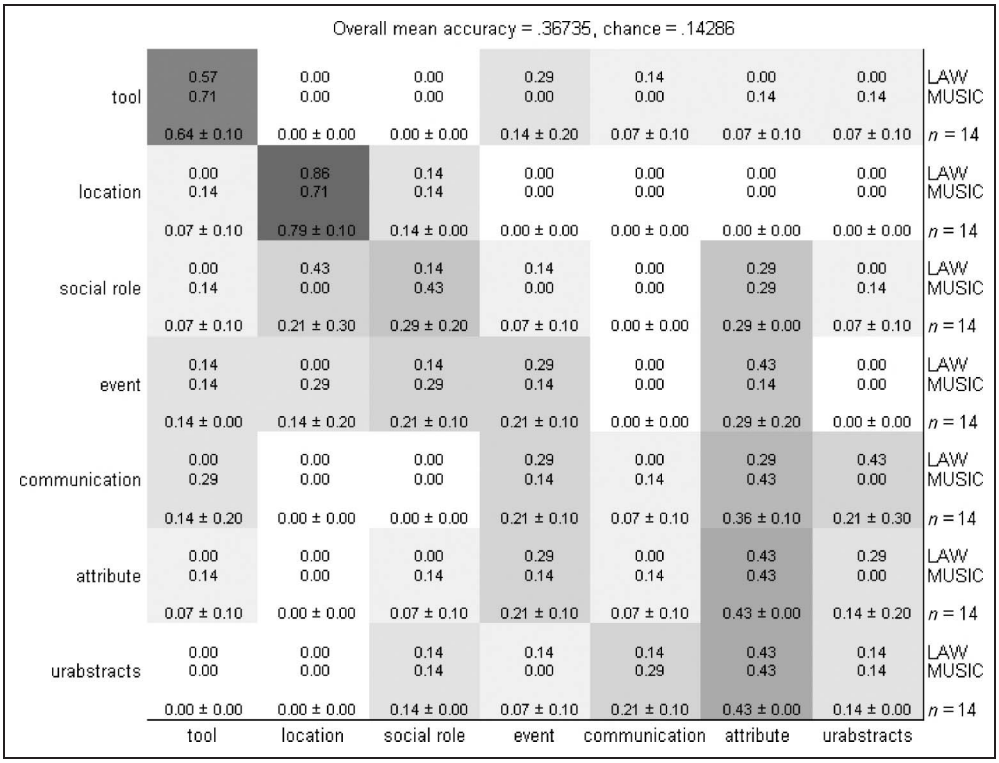

Figure 7. Leave-participant-out Taxonomic category classification confusion matrix. Rows are the target labels and columns are predictions. Numbers overlaid on each cell indicate the proportion of predictions for Law and Music, respectively (as indicated on the right *y* axis), for that row, averaging over seven participants. The numbers on the bottom line of each cell are the mean and standard deviation of predictions. Cell shading is scaled to the range 0–0.79 (0.79 is the maximum mean accuracy per cell displayed). Olivetti et al.'s (2012) test results [T = Tool] [L = Location] [S = Social Role] [E = Event] [C = Communication] [A = Attribute] [U = Urabstracts]: Partition 1: [[T] [L] [S,E,C,A,U]], postP: 0.1; Partition 2: [[T C] [L] [S,E,A,U]], postP: 0.068; Partition 3: [[T] [L] [S] [E,C,A,U]], postP: 0.067; Partition 4: [[T] [L] [S,E,A,U] [C]], postP: 0.062; Partition 5: [[T] [L] [S,E] [C,A,U]], postP: 0.032.

The probability of achieving this result (68/98 correct classifications) at chance is $p = .00015$ (two-tailed Binomial test). A confusion matrix of results is in Figure 8, and the results of Olivetti et al.'s (2012) discrimination test are in the caption, suggesting strongly that *Music* and *Law* can be discriminated (posterior probability > .99). Classification was however inaccurate for *Law*/*Location* and *Music*/*Event* and weak for *Law*/*Tool*. In the within-participant analysis discrimination of Domain from *Law*/*Tool* and *Location* is low (see Figures 2 and 3). There is however no obvious prior reason to anticipate such low prediction from *Music*/*Event*.

The Interrelation of Taxonomic Categories and Domains

Given the variability in the strength of Taxonomic category/Domain encoding observed between different taxonomic category pairs in Which is more strongly encoded, taxonomic category distinctions or domain distinctions? section and the apparent relationship of this to the degree of concreteness, we might expect there to be corresponding differences in the way that representations of Taxonomic category and Domain are organized. Before the study, we identified three models of the respective role of taxonomic categorization and domains in conceptual knowledge. These models, illustrated in Figure 9, include the following: (1) *Taxonomy within domain*: Each domain is independently represented, that is, there are spatially distinct regions in classification space devoted to processing information regarding each domain. Taxonomic categories pertinent to each domain are represented within the respective domain region. (2) *Domain within taxonomy*: There are spatially distinct regions in classification space dedicated to processing information regarding different taxonomic categories. All domain information relevant to a particular taxonomic category is represented within the respective taxonomic region (this is in fact how WordNet Domain is organized). (3) *Domain/taxonomy independent*: There are spatially distinct regions in classification space dedicated to processing all taxonomic and domain categories. There is no overlap between representations of domain and taxonomic class.

The plausibility of each model can be tested by cross-validation analyses, where entire taxonomic categories/domains are left out of the training set and used in testing. The *domain within taxonomy* model can be tested by leaving out a domain when training taxonomic classifiers and using this left out domain for testing. If there is a region encoding taxonomy common to both *Music* and *Law*, we would expect that a taxonomic classifier trained on *Music* words would be able to distinguish unseen *Law*/taxonomic category representations and vice versa. If taxonomic classification is at chance, this suggests that *Music* and *Law* representations of taxonomic categories are separate.

The *taxonomy within domain* model can be tested by leaving out taxonomic categories when training domain

classifiers and using the left out taxonomic category for testing. If the domain of the left-out taxonomic category cannot be predicted, then information pertinent to domain classification is not shared with the other taxonomic categories. For instance, the relationship of musical instruments to the *Music* domain may be encoded by neural representations pertaining to acoustics and complex manual motor patterns, which we would not necessarily expect to be active or relevant for distinguishing *Music Location*. If organization follows the *taxonomy within domain* model, domain classification should be unaffected by leaving out taxonomic categories. If organization follows the *taxonomy/domain independent* model, then either of the above analyses should not disrupt classification.

Leave-one-category-out cross-validation analyses were undertaken on each participant for taxonomic category and domain classification. In taxonomic categorization, there were two cross-validation iterations, where either all *Music* words or all *Law* words were left out of the training set and used in the test set. Training in each iteration was therefore on 175 volumes (five words in seven taxonomic categories and five replicates per word), and testing was also on 175 volumes. In each iteration, the 1000 most stable voxels were selected from the mean correlation of all words in the training set (35), between each unique session pair (5 sessions; therefore, 10 correlation coefficients were averaged per voxel to give the stability score).

For domain classification, there were seven cross-validation iterations, in each of which all words from a taxonomic class were left out of the training set and used in testing. In each iteration, training was therefore on 300 volumes (10 words in six taxonomic categories with five replicates per word), and testing was on 50 words. For each iteration, the 1000 most stable voxels were selected using correlations as above, this time based on 60 words in five sessions, and again 10 correlation coefficients were averaged.

As the above two analyses use a different number of volumes in training, comparative to the leave-out-session analyses in Can taxonomic category-based and domain-based distinctions be recognized within participants? section (leave-Domain-out Taxonomic category classification is based on 175 volumes in contrast to 280, and leave-Domain-Taxonomic-category-out classification is based on 300 rather than 280 volumes), control analyses with data set sizes matching the above analyses were run. Words in the control training and test sets were randomly selected but constrained to span all secondary categories (i.e., in the control for the leave-Taxonomic-category-out test, five words would be left out of training, three that were selected randomly would belong to one domain and the remaining two would be to the other).

Can the Taxonomic Category of Words from an Unseen Domain Be Predicted?

Mean classification accuracy for leave-Domain-out Taxonomic category classification was ~.22, with chance level

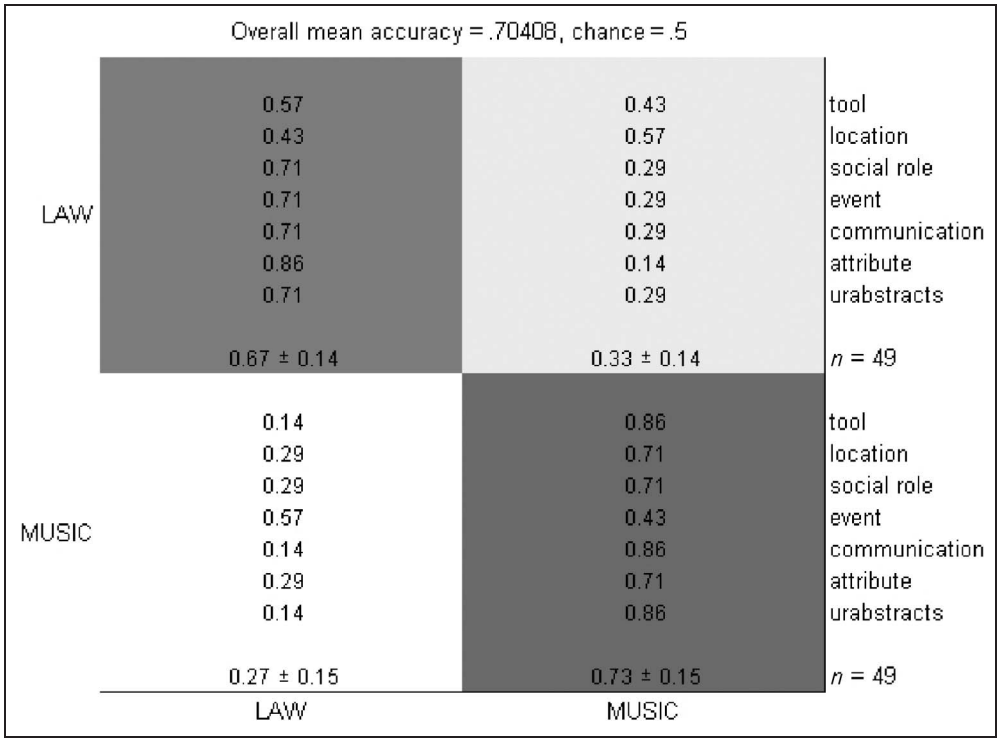

Figure 8. Leave-participant-out Domain classification confusion matrix. Rows are the target labels, and columns are predictions. Numbers overlaid on each cell indicate the proportion of predictions per taxonomic category (as indicated on the right *y* axis) for that row, averaging over seven participants. The numbers on the bottom line of each cell are the mean and standard deviation of the predictions. Cell shading is scaled to the range 0–0.73 (0.73 is the maximum mean accuracy per cell displayed). Olivetti et al.'s (2012) test results [L = LAW][M = MUSIC]: Partition 1: [[L][M]], postP: >0.99; Partition 2: [[L,M]], postP: 0.0015.

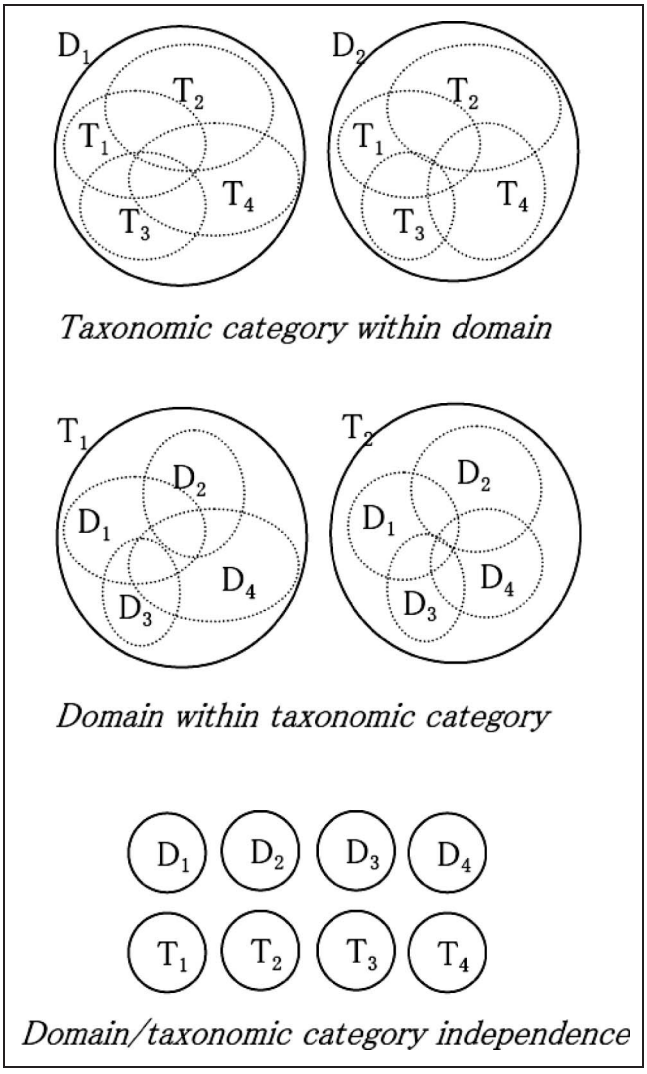

Figure 9. Putative models of organization of Domain/Taxonomic class intersections. Description is in The interrelation of taxonomic categories and domains section. The dashed lines are intended to symbolize a fuzzy class boundary with overlap between neighboring classes.

at .143. In total there were 534/2450 correct classifications, and the probability of achieving this by chance is $p \leq 2.2 \times 10^{-16}$ (two-tailed Binomial test). A confusion matrix of results is in Figure 10. *Tool* and *Location* can be seen to be relatively well predicted; however, the less concrete classes aggregate. The top four ranking partitions from Olivetti et al.'s (2012) discrimination test are in the caption of Figure 10 (the remainder in the supplementary materials for Figure 10). The top ranking partition (posterior probability of .67) concurs with this observation: *Tool* and *Location* are distinct, and all other classes form the same subset. The second to fourth partitions have posterior probabilities ranging from 0.15 to 0.01. Each partition discriminates between either three or four subsets of taxonomic category; however, *Tool* and *Location* are always distinct, and the less concrete classes conflated in certain combinations.

The comparative control confusion matrix is in Figure 11. Mean accuracy in the control was ~.25, there were 606/2450 correct classifications, and the probability of achieving this by chance is $p \leq 2.2 \times 10^{-16}$ (two-tailed Binomial test). Here there also appears to be some ambiguity among the less concrete classes (suggesting that the reduced training set size comparative to Can taxonomic category-based and domain-based distinctions be recognized within participants? section impairs classification); however, *Tool*, *Location*, *Social Role*, and *Attribute* are visibly distinct. The four top ranking partitions from Olivetti et al. apos;s (2012) analysis are in the caption of Figure 11 (the remaining partitions are in the supplementary materials for Figure 11). The top ranking partition (posterior probability = .67) indeed segregates *Tool*, *Location*, *Social Role*, and *Attribute*, whereas *Event*, *Communication*, and *Urabstracts* are aggregated. The following three partitions have posterior probabilities ranging between .13 and .03 and allocate categories to five, four, and six subsets, respectively (with *Tool* and *Location* always distinct and the less concrete classes split into two to four subsets).

It follows that leaving out an entire domain disrupts classification of the less concrete categories, which is consistent with the suggestion that the less concrete concepts more closely adhere to taxonomic category within domain organization. However, it should be recognized that classification accuracy is reduced for six of the seven taxonomic categories (including *Tool* and *Location*). Therefore, it is reasonable to suggest that domain is a less important organizing principle for the more concrete experimental concrete concepts; however, it is not irrelevant.

Can Domain Be Predicted from an Unseen Taxonomic Category?

In the leave-Taxonomic-category-out classification analysis, mean accuracy was ~.7, with chance at .5. There were 1710/2450 correct classifications, and the probability of achieving this by chance is $p \leq 2.2 \times 10^{-16}$ (two-tailed Binomial test). Olivetti et al.'s (2012) test strongly supports that Domains can be distinguished (posterior probability > 0.99). Inspecting Figure 12, it can be seen that prediction of Domain from *Law*/*Tool* is at chance level. Prediction of Domain from *Music*/*Location* is also weak; however, this was also the case in the original analysis (Figure 5). This implies that the information necessary to correctly decode *Law*/*Tool* as being related to *Law* is not contained in the remaining training data. In contrast, the information necessary to predict the relation of *Tool* to *Music* clearly overlaps the other categories (presumably the overlap is with the less concrete categories given the weak prediction of *Music* from *Location*).

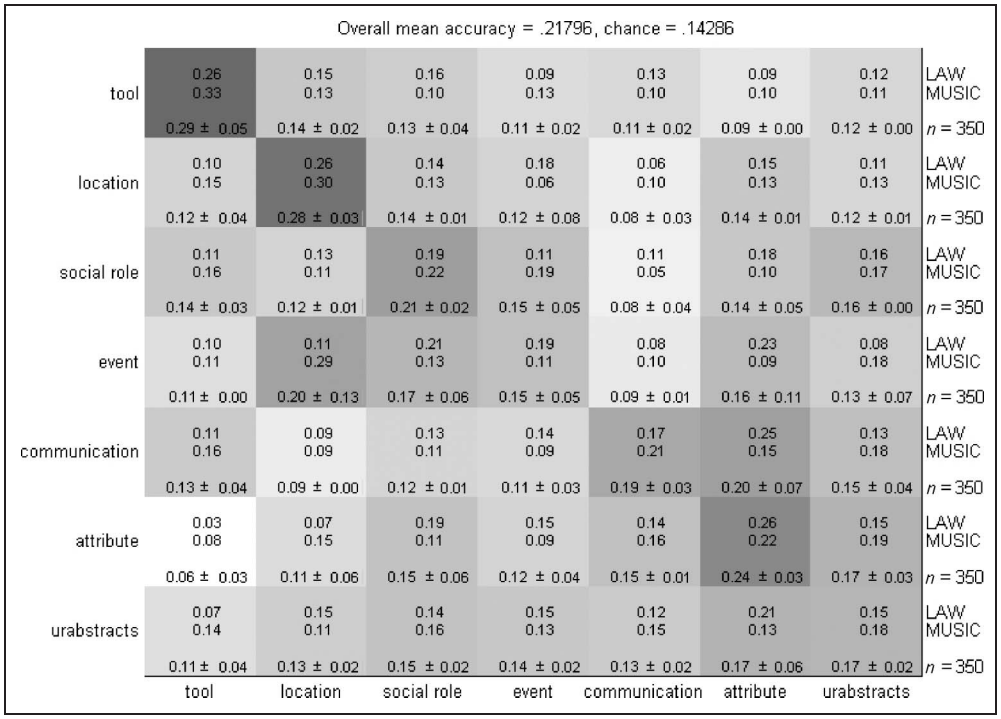

Figure 10. Leave-Domain-out Taxonomic category classification confusion matrix. Rows are the target labels, and columns are predictions. Numbers overlaid on each cell indicate the proportion of predictions per Law and Music, respectively (as indicated on the right *y* axis), for that row, averaging over seven participants. The numbers on the bottom line of each cell are the mean and standard deviation of predictions. Cell shading is scaled to the range 0–0.29 (0.29 is the maximum mean accuracy per cell displayed). Olivetti et al.'s (2012) test results [T = Tool] [L = Location] [S = Social Role] [E = Event] [C = Communication] [A = Attribute] [U = Urabstract]: Partition 1: [[T] [L] [S,E,C,A,U]], postP: 0.67; Partition 2: [[T] [L] [S,E,A,U] [C]], postP: 0.15; Partition 3: [[T] [L] [S] [E,C,A,U]], postP: 0.15; Partition 4: [[T] [L] [S,E] [C,A,U]], postP: 0.01.

The mean accuracy in the control confusion matrix (Figure 13) was ~.71, with chance at 0.5. There were 1732/2450 correct classifications, and the probability of achieving this by chance is $p \leq 2.2 \times 10^{-16}$ (two-tailed Binomial test). The results of Olivetti et al.'s (2012) test, listed in the caption of Figure 13, give strong evidence (posterior probability > 0.99) that Domain can be discriminated. Discrimination of *Law*/*Tool* and *Music*/

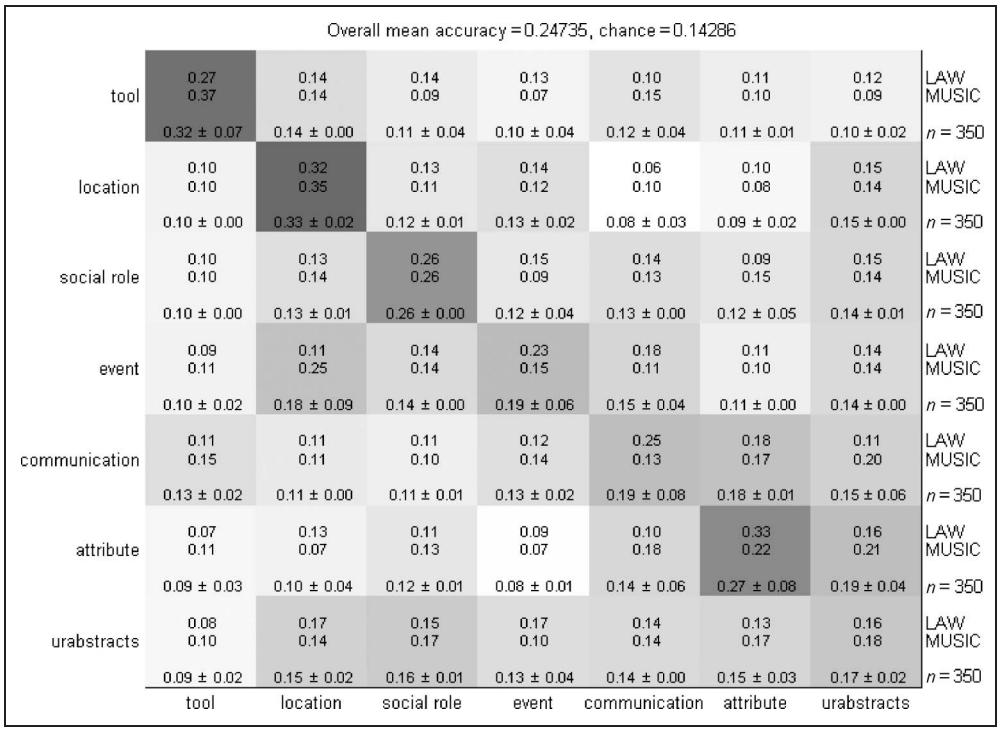

Figure 11. Control for leave-Domain-out Taxonomic category classification confusion matrix. Rows are the target labels, and columns are predictions. Numbers overlaid on each cell indicate the proportion of predictions per Law and Music, respectively (as indicated on the right *y* axis) for that row, averaging over seven participants. The numbers on the bottom line of each cell are the mean and standard deviation of predictions. Cell shading is scaled to the range 0–0.33 (0.33 is the maximum mean accuracy per cell displayed). Olivetti et al.'s (2012) test results: [T = Tool] [L = Location] [S = Social Role] [E = Event] [C = Communication] [A = Attribute] [U = Urabstract]: Partition 1: [[T] [L] [S] [E,C,U] [A]], postP: 0.77; Partition 2: [[T] [L] [S] [E] [C,A,U]], postP: 0.13; Partition 3: [[T] [L] [S,E,C,U] [A]], postP: 0.03; Partition 4: [[T] [L] [S] [E] [C,U] [A]], postP: 0.03.

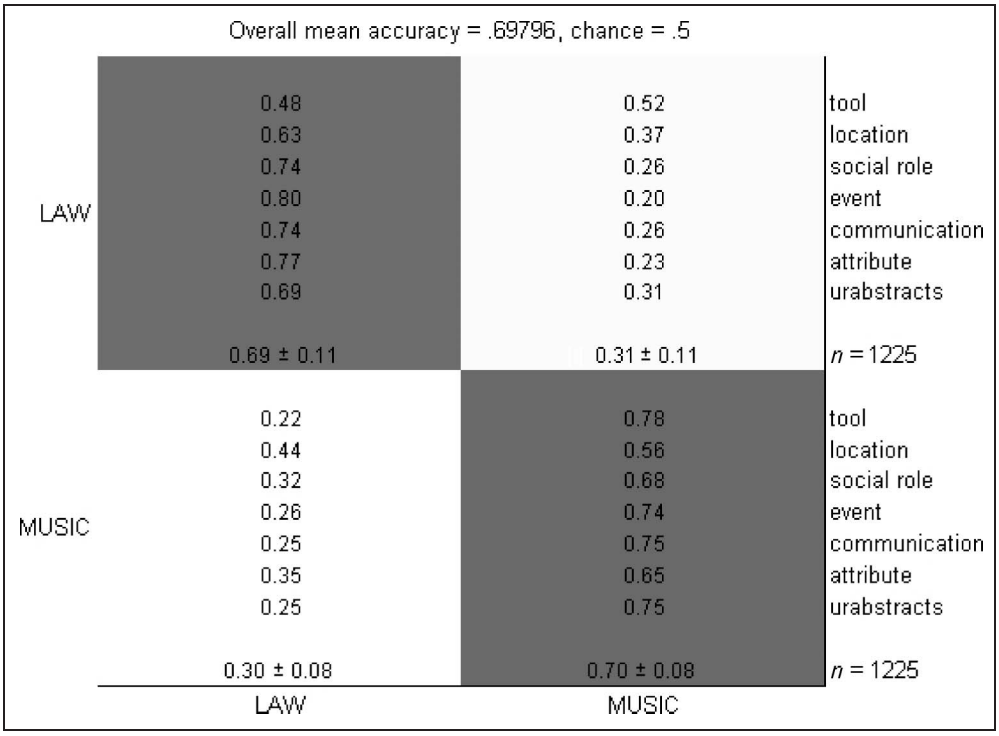

Figure 12. Leave-Taxonomic-category-out Domain classification confusion matrix. Rows are the target labels, and columns are predictions. Numbers overlaid on each cell indicate the proportion of predictions per taxonomic category (as indicated on the right *y* axis) for that row, averaging over seven participants. The numbers on the bottom line of each cell are the mean and standard deviation of the predictions. Cell shading is scaled to the range 0–0.7 (0.7 is the maximum mean accuracy per cell displayed). Olivetti et al.'s (2012) test results [L = LAW] [M = MUSIC]: Partition 1: [[L][M]], postP: >0.99; Partition 2: [[L,M]], postP: 6.6×10^{-85}.

Location is also observably improved from the leave-Taxonomic-category-out test (in fact we might expect better classification than Can taxonomic category-based and domain-based distinctions be recognized within participants? section given the greater training set size here).

In summary: (1) Testing the taxonomic category classifiers on an unseen Domain reduces classification accuracy per se and disrupts distinction between less concrete concepts. Taxonomic category within Domain organization appears more appropriate to describe nonconcrete concepts. (2) Testing Domain classifiers on unseen

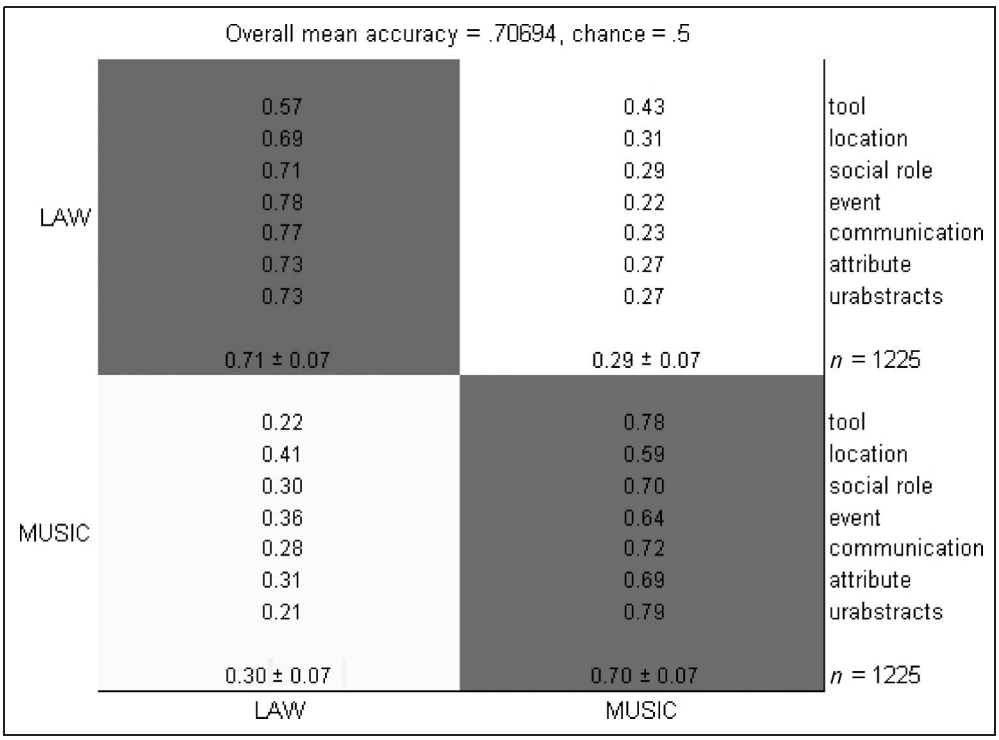

Figure 13. Control for Leave-Taxonomic-category-out Domain classification confusion matrix. Rows are the target labels, and columns are predictions. Numbers overlaid on each cell indicate the proportion of predictions per taxonomic category (as indicated on the right *y* axis) for that row, averaging over seven participants. The numbers on the bottom line of each cell are the mean and standard deviation of the predictions. Cell shading is scaled to the range 0–0.71 (0.71 is the maximum mean accuracy per cell displayed). Olivetti et al.'s (2012) test results [L = LAW][M = MUSIC]: Partition 1: [[L][M]], postP: 1>0.99; Partition 2: [[L,M]], postP: 4.1×10^{-93}.

taxonomic categories selectively ablates identification of the relationship between *Law*/*Tool* and *Law*. This suggests at least for *Location* that some Domain information is represented within the Taxonomic category.

Anatomic Regions Discriminating Taxonomic Categories and Domains

Our major focus thus far has been establishing that the categories tested are cognitively relevant, we close with an analysis probing which anatomical regions independently support classifications. The taxonomic/domain classification analyses of Can taxonomic category-based and domain-based distinctions be recognized within participants? section were repeated for each participant on each of 116 anatomical ROIs, defined by Tzourio-Mazoyer et al. (2002)'s Automated Anatomical Labeling scheme (e.g., as per Wang et al., 2012). All voxels in each region contributed to analysis; however, not all 116 areas were detected for each participant. Otherwise the cross-validation procedure was the same. ROIs for which at least one Taxonomic category could be discriminated when results are summed over participants (according to the most likely hypothesis from Olivetti et al.'s, 2012 analysis) are displayed in Figure 14 and listed in Table 5. Full results are in the supplementary material for Table 5.

Twelve of the 15 ROIs are located in the left hemisphere as would be expected of a language-related task. All seven categories could be discriminated from the left middle occipital gyrus, with an overall accuracy of .25 (.5 less than the whole-brain analysis). *Tool*, *Location*, *Social Role*, and *Attribute* can be discriminated from the left middle temporal gyrus, and three categories (combinations of *Tool*, *Location*, *Social Role*, and *Attribute*) were distinguished from the left precuneus and left inferior and superior parietal regions. More specialized distinctions for specific concrete categories were located in the left precentral gyrus (*Tools* only) and fusiform gyri (*Location* only). The left inferior frontal gyrus distinguishes *Attributes* and *Tools* (i.e., arguably the most abstract class, from the most concrete class, and all other intermediate classes are conflated), and the neighboring left middle frontal gyrus distinguishes *Attributes* from the other categories.

ROIs for which Domains could be discriminated at a level of .58 or over (all distinguishable according to Olivetti et al.'s, 2012 analysis), are displayed in Figure 15 and listed in Table 6 (full results are in the supplementary

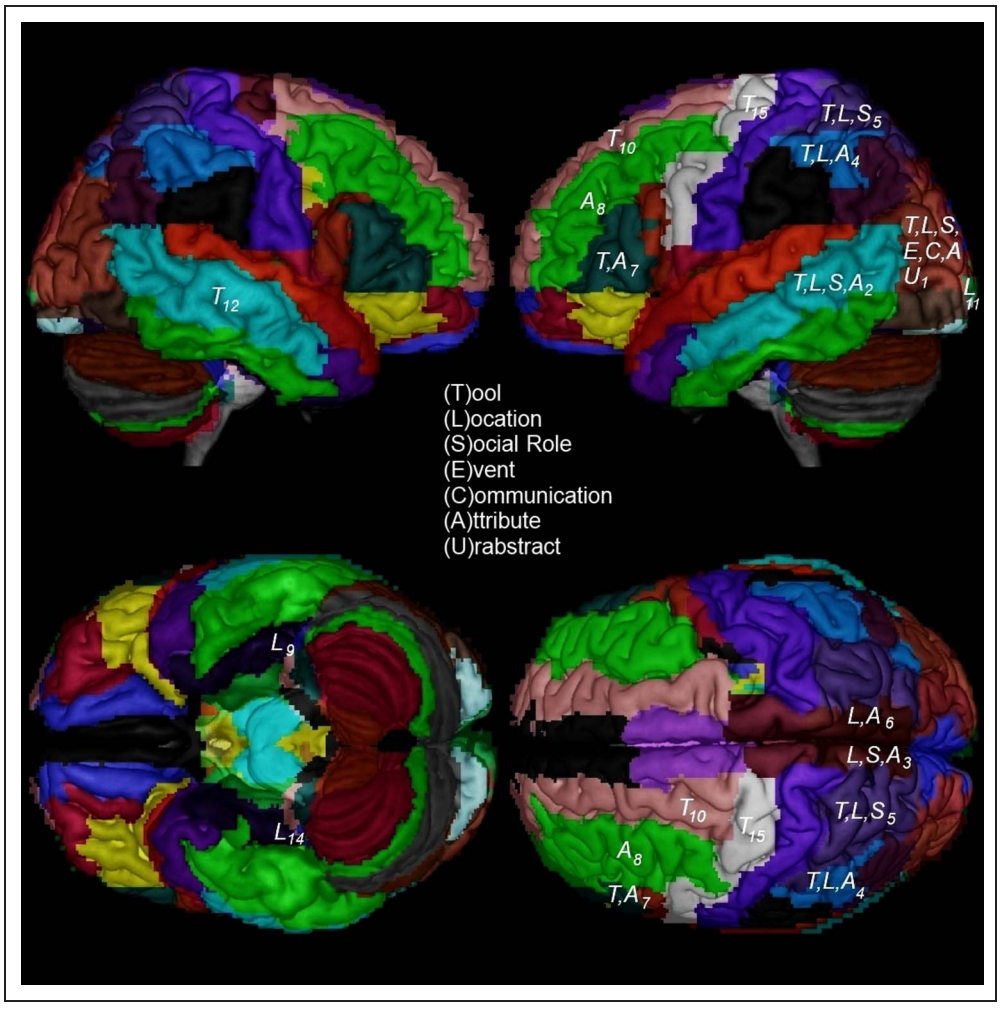

Figure 14. Depiction of within anatomical region Taxonomic category classification results listed in Table 5. AAL regions are color coded, and categories that could be discriminated within each region are overlaid. The numeric subscript links regions to Table 5.

Table 5. Taxonomic Classification Accuracy within Anatomical Regions, Summed over All Seven Participants (Accuracy Corresponds to the Diagonal of the Confusion Matrix)

	Tool	*Location*	*Social Role*	*Event*	*Communication*	*Attribute*	*Urabstracts*	*Mean Accuracy*	*Mean # Voxels*	*Std # Voxels*
Occipital_Mid_L_1	**108**	**103**	**75**	**73**	**80**	**100**	**71**	0.25	428.29	10.21
Temporal_Mid_L_2	**100**	**88**	**90**	64	79	**92**	52	0.23	610.43	56.62
Precuneus_L_3	76	**113**	**73**	69	71	**79**	67	0.22	426.57	18.27
Parietal_Inf_L_4	**98**	**84**	69	73	71	**83**	47	0.21	297.57	8.34
Parietal_Sup_L_5	**88**	**102**	**72**	52	64	77	69	0.21	225.71	11.19
Precuneus_R_6	73	**100**	71	61	62	**88**	56	0.21	409.43	12.69
Frontal_Inf_Tri_L_7	**92**	62	69	59	78	**76**	66	0.2	315.14	5.67
Frontal_Mid_L_8	77	80	61	47	69	**90**	60	0.2	603.57	13.9
Fusiform_L_9	63	**104**	78	70	54	55	58	0.2	246.14	51.7
Frontal_Sup_L_{10}	**91**	61	64	63	60	82	58	0.2	421.86	14.87
Calcarine_L_{11}	76	**94**	69	57	45	75	57	0.19	310.29	10.05
Temporal_Mid_R_{12}	**88**	61	64	63	53	77	63	0.19	612.86	16.96
Frontal_Inf_Oper_L_{13}	**94**	66	63	48	63	71	55	0.19	134.57	2.07
Fusiform_R_{14}	66	**90**	71	55	60	57	56	0.19	272.86	30.99
Precentral_L_{15}	**89**	65	52	50	68	74	50	0.18	395.71	14.5

Full results are in the supplementary material for Table 5. Anatomical regions are shown for which at least one category could be discriminated according to Olivetti et al.'s (2012) statistics. Categories that can be distinguished according to the most likely hypothesis arising from Olivetti et al.'s test are highlighted in **bold**. The two rightmost columns show the mean and standard deviation of the number of voxels per region. The subscripted numbers by each region's name link to Figure 14, where results are displayed on the cortical surface.

material for Table 6). There are 15 ROIs matching this criteria (which is the number of areas supporting Taxonomic distinctions), 10 of which are in the left hemisphere. Best distinction was in the left middle temporal gyrus; however, relatively good accuracy was also observed in the right superior and middle temporal gyri, left superior temporal gyrus, and left precuneus. Notably, the superior temporal gyri contain the auditory cortex, which may be relevant to discriminating *Music*. Absent from the list in Table 6 is the left middle occipital gyrus, for which accuracy was .57 (and still statistically supported, see supplementary material); however, the left angular gyrus (implicated by Binder et al., 2009, to integrate complex information) does appear, as also do five areas in the

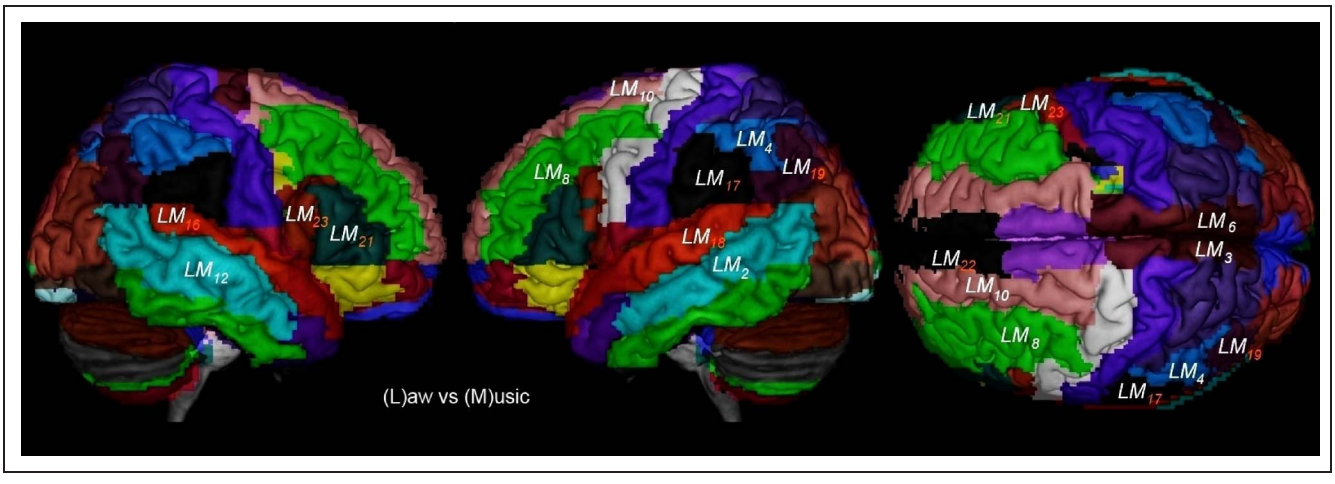

Figure 15. Depiction of within anatomical region Domain category classification results listed in Table 6. AAL regions are color coded, and the overlay indicates categories where Domain discrimination was possible. The numeric subscript links regions to Table 6, regions where also Taxonomic categories were distinguishable have a White numeric subscript. Note that the cingulate gyrus is not visible here (see Table 6).

Table 6. Domain Classification Accuracy within Anatomical Regions, Summed over All 7 Participants

	Law	*Music*	*Mean Accuracy*	*Mean # Voxels*	*SD # Voxels*
Temporal_Mid_L_2	753	789	0.63	610.43	56.62
Temporal_Sup_R_{16}	739	771	0.62	424.14	12.52
Precuneus_L_3	754	750	0.61	426.57	18.27
SupraMarginal_L_{17}	752	744	0.61	152.57	2.88
Temporal_Sup_L_{18}	751	744	0.61	311.71	16.48
Temporal_Mid_R_{12}	749	734	0.61	612.86	16.96
Frontal_Mid_L_8	723	733	0.59	603.57	13.90
Parietal_Inf_L_8	739	712	0.59	297.57	8.34
Angular_L_{19}	735	704	0.59	151.29	5.25
Cingulum_Post_L_{NA}	703	736	0.59	48.57	2.70
Frontal_Inf_Tri_R_{21}	714	725	0.59	269.57	6.40
Frontal_Sup_L_{10}	729	707	0.59	421.86	14.87
Precuneus_R_6	705	727	0.58	409.43	12.69
Frontal_Sup_Medial_L_{22}	701	723	0.58	330.43	11.52
Frontal_Inf_Oper_R_{23}	724	697	0.58	173.43	5.56

Full results are in the supplementary material for Table 6. Following Olivetti et al. (2012), Law and Music can be distinguished in all of the ROIs displayed. The two rightmost columns show the mean and standard deviation of the number of voxels per region. The subscripted numbers by each region's name link to Figure 15, where results are displayed on the cortical surface.

frontal lobe (again including the left inferior and middle gyri).

DISCUSSION

Taxonomic Category/Domain Classification

Our primary finding is that both types of conceptual organization are encoded in fMRI signals. Our evidence suggests that fMRI recordings contain sufficient information to discriminate between all taxonomic categories tested here: In other words, the distinctions between nonconcrete categories proposed in state-of-the-art models of conceptual knowledge such as WordNet are supported to a certain extent by brain data. We found evidence for domain-based organization, as it was also possible to discriminate between Domains. However, we also observed a concreteness effect, most clearly demonstrated in the pairwise tests of Which is more strongly encoded, taxonomic category distinctions or domain distinctions? section, where Domain is most accurately distinguished for the least concrete classes, whereas taxonomic category is better distinguished when one or both classes are concrete/near-concrete.

Subjectivity of Conceptual Representations

A second key finding is that these distinctions have similarities in encoding across participants: in particular, *Tool* and *Location* could be reliably predicted for unseen participants. But all other (less concrete) classes aggregated. This result is in agreement with intuition, in that an individual's experience of abstract entities is likely to be more subjective than that of concrete entities, which will tend to be sensed and interacted with in similar ways. As such, they are grounded in a shared perceptual reality as many perceptual and motor processing networks are mapped in broadly the same way across healthy people. As the features underlying nonconcrete concepts are not clear, we cannot assume similar anatomical mappings across people and for them to be embodied within such expansive areas of cortex as those dedicated to perception and action. It follows that commonalities in representation may be found at a smaller volumetric scale than that investigated here. If the lack of cross-participant shared representations was to be confirmed also at a smaller scale, the question would be raised of how people manage to understand each other when discussing nonconcrete concepts, given the role that common ground plays in communication (see, e.g., Clark, 1996). But the need for a theory of communication that does not rely on assuming "perfect" understanding on the part of the speakers has already been made abundantly clear by recent psycholinguistic work suggesting, for example, that speakers only construct "good enough" representations of each other's sentences (see, e.g., Ferreira, Ferraro, & Bailey, 2002), although the exact form such a theory would take is by no means clear (for an attempt, see, e.g., Poesio & Rieser, 2010).

Models of Concept Organization

Three organizational models addressing how joint Domain/Taxonomic category membership might be arranged were considered. Again we observed a concreteness distinction, in that Domain appears to be a more important organizational principle for less concrete words (it proved difficult to predict the taxonomic category of nonconcrete words from unseen Domains) whereas the more concrete categories *Tool* and *Location* could be identified.

Anatomical Regions of Interest

Repeating the taxonomic and domain discrimination on segmented anatomical ROIs revealed a number of different areas supporting classification. Some were generalist in the sense that it was possible to discriminate all or most Taxonomic/Domain categories (e.g., left middle occipital gyrus and left middle temporal gyrus), others appeared to be more specialized (e.g., left precentral gyrus discriminating *Tools*). It is immediately clear that widespread cortical activity supports discrimination, so we are left with a question over what neural processes our classifiers were actually discriminating within the different regions. We suggest that the strongest case of discrimination, observed in the left middle occipital gyrus, was exploiting differences in grounded visual representations activated in mental imagery. This region has previously been found to discriminate neural activation associated with mental imagery of different patterns, and these self-generated representations were shown to be similar to those observed when patterns were supplied as perceptual input (Stokes, Thompson, Cusack, & Duncan, 2009; Stokes, Thompson, Nobre, & Duncan, 2009). Additionally Anderson, Bruni, Bordignon, Poesio, and Baroni (2013) have begun to accrue evidence suggesting that concrete concept elicited neural activity in occipital areas can be explained by natural image statistics. The precuneus is believed to be involved in self-centered mental imagery and episodic memory retrieval (e.g., see Cavanna & Trimble, 2006), the latter of which would be relevant for recall of the rehearsed concept simulations. Recent evidence suggests that representation of objects in this area is at least in part amodal (Fairhall & Caramazza, 2013), so it may be safer to suggest that the precuneus could assist in configuring the imaginary-spatial layout of a scenario reconstructed in perceptual regions rather than being the site of reconstruction. Areas around the intraparietal sulcus are commonly implicated to have roles in spatial attention (e.g., see Gillebert et al., 2011). We consider that attentional processes could be a precursor to mental imagery,[1] triggering activity in perceptual systems that are appropriate to simulating the concept. This could be seen as relating to goal-oriented behavior where attention modulates perception to attend to specific target features, for example, looking for the "horizontal red line" (see also Stokes, Thompson, Cusack, et al., 2009; Stokes, Thompson, Nobre, et al., 2009).

Binder et al. (2009) suggest that the left middle temporal gyrus subserves a role in supramodal integration and processing perceptual information about objects and their attributes (further evidence coming from Fairhall & Caramazza, 2013), which is consistent with discrimination of the more concrete categories *Tools*, *Locations*, and possibly *Social Roles*. In addition Wang et al. (2010, 2012) and Rodríguez-Ferreiro, Gennari, Davies, and Cuetos (2011) have connected this area with abstract word processing, which is in keeping with the additional discrimination of *Attributes*. Discrimination in the left precentral gyrus (*Tools* only) is presumably via neurons associated with motor control. In similar experimental paradigms, using line drawings as stimuli, the left precentral and fusiform gyri, respectively, were associated with "tools" and "dwellings" (Shinkareva et al., 2008) or using text as stimuli, "manipulation" and "shelter" (Just et al., 2010). Following this interpretation, it is perhaps unexpected that the left angular gyrus, which Binder et al. (2009) suggest has a role in complex information integration and knowledge retrieval only, segregated Domains but not Taxonomic categories. However, this may be a byproduct of the relatively meager voxel coverage of this area (mean = 151 voxels) comparative to, for example, the left middle occipital gyrus (mean = 428 voxels) and the left middle temporal gyrus (mean = 610 voxels).

Conversely, discrimination in frontal areas is more likely to be based on more abstract symbolic/linguistic semantic information. The inferior frontal gyrus has been implicated in abstract word processing (e.g., Wang et al., 2010, 2012; Rodríguez-Ferreiro et al., 2011; Binder et al., 2009). Here we find that the left inferior frontal gyrus distinguishes *Attributes* and *Tools*. The neighboring left middle frontal gyrus distinguishes *Attributes* from the other categories, and the left superior frontal gyrus distinguishes *Tools*. Taken all together, this is a distinction between arguably the most abstract class and the most concrete class, and all other intermediate classes are conflated. Binder et al. (2009) have suggested the left inferior gyrus has a role biased toward syntactic processing and is beneficial but not essential to semantic processing and the superior frontal gyrus may be responsible for coordinating knowledge retrieval.

In summary, we tentatively allude to a mechanism where frontal regions serve goal-directed control, combining with parietal attention processes, to initiate and configure mental simulations instantiated in perceptual and motor areas. Amodal representations in the temporal cortex could be activated in interpreting stimuli words, generating content for mental simulations (e.g., object representations), monitoring the imaginary scenario, or all three.

Particular questions remaining for the future concern how modal/supramodal representations concerning concepts change within and between anatomical regions

(see also the discussion in Shinkareva et al., 2008). For instance, (1) multiple representations of a concept set organized in the same way could support independent processing in different neural regions, guard against damage, and allow easy mapping from one modality to another; (2) replications of a concept set with different organizational schemes could facilitate neural queries by allowing different methods of indexing information and different ways to combine concepts in planning (e.g., in a hypothetical visual feature map "snake" and "belt" could share a similar locus based on appearance, whereas in a semantic map this would be unlikely, or of relevance, to this article, we might speculate our concept set is jointly represented according to both *Taxonomy within Domain* and *Domain within Taxonomy* structures). Note that Anzellotti, Mahon, Schwarzbach, and Caramazza (2011) have observed qualitative differences in computation between regions; (3) a concept set may be replicated at multiple-nested spatial scales of representation (with either similar or different representations), potentially allowing efficient parallel processing at different "volumetric" frequency bands.

Concreteness Ratings

One of our goals has been to broaden the range of nonconcrete concepts investigated and move beyond the concrete/abstract dichotomy. The inconsistencies we observed in peoples' concreteness ratings (Materials section) raise doubts about the value of rating concepts on a concrete–abstract continuum. On the flipside, through demonstrating systematic subdivisions of nonconcrete concepts exist, our results warn against lumping all abstract concepts together as a superclass, which risks concealing meaningful class divisions. The implications of our results are to encourage consideration of how different semantic categorization schemes may be exploited to elucidate neural organization (e.g., with our classes, in behavioral experiments, we might anticipate within taxonomic/domain class processing advantages). The experimental value of concreteness ratings has also recently come under criticism also from Connell and Lynott (2012). They collected perceptual strength norms in sound, taste, touch, smell, and vision (e.g., Lynott & Connell, 2009) and demonstrated that strength in the dominant modality had more explanatory value than concreteness in behavioral tests traditionally used to demonstrate concreteness effects. Additionally, Kousta, Vigliocco, Vinson, Andrews, and Del Campo (2011) have argued that the emotional content of words (abstract words tend to be more valenced) plays an important role in explaining such behavioral differences. In future work, it may prove interesting to relate our class divisions to multimodal perceptual norms and also valence norms and to examine whether these can describe different organization of concepts between anatomical regions that we have conjectured earlier this section.

Scenario-based Conceptual Organization

Through basing our experiment on mentally simulating situations, the success of our results provide support for theories considering concepts to be grounded in situations (Barsalou & Wiemer-Hastings, 2005). These theories expect concrete concepts to be linked to relatively narrow range of situations than abstract concepts, and as such might predict them to be more discriminable, and indeed there was some evidence of this. It also follows that abstract concept representation might be more variable. On testing for a relationship between variability in concept representation and concreteness, there was found to be no correlation.[2] However, this also could be a byproduct of our experimental design, where participants rehearsed mental simulations in advance, or differences could have been invisible to fMRI, for example, hidden by the low sampling rate.

Reprint requests should be sent to Andrew Anderson, Centro Interdipartimentale Mente/Cervello, University of Trento, Palazzo Fedrigotti, Corso Bettini 31, 38068 Rovereto, Italy, or via e-mail: andrew.anderson@unitn.it.

Notes

1. We know of no reason to expect confounds associated with perceptual characteristics of the stimulus words to have driven discrimination (only *Urabstracts*, which were quite difficult to discriminate, are separable in this respect; see Materials section).

2. For each participant, voxels that contributed to at least 3/5 cross-validation iterations were selected, and the voxelwise standard deviation was estimated over the five replicates of the word. The mean standard deviation over participants was taken per word, and results were correlated with the concreteness ratings from the norms. Pearson's correlation revealed a weak correlation that was not significant ($r = .2$, $p = .08$, $n = 70$).

REFERENCES

Anderson, A. J., Bruni, E., Bordignon, U., Poesio, M., & Baroni, M. (2013). Of words, eyes and brains: Correlating image-based distributional semantic models with neural representations of concepts. *Proceedings of EMNLP 2013* (pp. 1960–1970). Seattle, WA.

Anzellotti, S., Mahon, B. Z., Schwarzbach, J., & Caramazza, A. (2011). Differential activity for animals and manipulable objects in the anterior temporal lobes. *Journal of Cognitive Neuroscience, 8,* 2059–2067.

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

Barsalou, L. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences, 22,* 577–660.

Barsalou, L. W., & Wiemer-Hastings, K. (2005). Situating abstract concepts. In D. Pecher & R. A. Zwaan (Eds.), *Grounding cognition: The role of perception and action in memory, language, and thinking* (pp. 129–163). Cambridge: Cambridge University Press.

Bentivogli, L., Forner, P., Magnini, B., & Pianta, E. (2004). Revising WordNet domains hierarchy: Semantics, coverage, and balancing. In *Proceedings of COLING 2004 Workshop*

on "Multilingual Linguistic Resources," Geneva, pp. 101–108.
Binder, J. R., & Desai, R. H. (2011). The neurobiology of semantic memory. *Trends in Cognitive Science, 15,* 527–536.
Binder, J. R., Desai, R. H., Graves, W. W., & Conant, L. L. (2009). Where is the semantic system? A critical review and meta-analysis of 120 functional neuroimaging studies. *Cerebral Cortex, 19,* 2767–2796.
Binder, J. R., Westbury, C. F., McKiernan, K. A., Possing, E. T., & Medler, D. A. (2005). Distinct brain systems for processing concrete and abstract concepts. *Journal of Cognitive Neuroscience, 17,* 905–917.
Caramazza, A., & Shelton, J. R. (1998). Domain specific knowledge systems in the brain: The animate/inanimate distinction. *Journal of Cognitive Neuroscience, 10,* 1–34.
Cavanna, A. E., & Trimble, M. R. (2006). The precuneus: A review of its functional anatomy and behavioural correlates. *Brain, 129,* 564–583.
Chang, K. M., Mitchell, T. M., & Just, M. A. (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.
Clark, H. H. (1996). *Using language*. Cambridge: Cambridge University Press.
Connell, L., & Lynott, D. (2012). Strength of perceptual experience predicts word processing performance better than concreteness or imageability. *Cognition, 125,* 452–465.
Connolly, A. C., Guntupalli, J. S., Gors, J., Hanke, M., Halchenko, Y. O., Wu, Y. C., et al. (2012). The representation of biological classes in the human brain. *Journal of Neuroscience, 32,* 2608–2618.
Damasio, H., Tranel, D., Grabowski, T., Adolphs, R., & Damasio, A. (2004). Neural systems behind word and concept retrieval. *Cognition, 92,* 179–229.
Fairhall, S., & Caramazza, A. (2013). Brain regions that represent amodal conceptual knowledge. *Journal of Neuroscience, 33,* 10552–10558.
Fellbaum, C. (1998). *WordNet*. Cambridge, MA: MIT Press.
Ferreira, F., Ferraro, V., & Bailey, K. G. D. (2002). Good-enough representations in language comprehension. *Current Directions in Psychological Science, 11,* 11–15.
Friederici, A. D., Steinhauer, K., & Pfeifer, E. (2002). Brain signatures of artificial language processing: Evidence challenging the critical period hypothesis. *Proceedings of the National Academy of Sciences, U.S.A., 99,* 529–534.
Gangemi, A., Guarino, N., Masolo, C., Oltramari, A., Oltramari, R., & Schneider, L. (2002). Sweetening ontologies with DOLCE. *AI Magazine, 24,* 13–24.
Gentner, D. (1981). Some interesting differences between verbs and nouns. *Cognition and Brain Theory, 4,* 161–178.
Gillebert, C. R., Mantini, D., Thijs, V., Sunaert, S., Dupont, P., & Vandenberghe, R. (2011). Lesion evidence for the critical role of the intraparietal sulcus in spatial attention. *Brain, 134,* 1694–1709.
Grossman, M., Smith, E. E., Koenig, P., Glosser, G., DeVita, C., Moore, P., et al. (2002). The neural basis for categorization in semantic memory. *Neuroimage, 17,* 1549–1561.
Hampton, J. A. (1981). An investigation in the nature of abstract concepts. *Memory and Cognition, 9,* 149–156.
Hanson, S. J., & Halchenko, Y. O. (2008). Brain reading using full brain support vector machines for object recognition: There is no "face" identification area. *Neural Computation, 20,* 486–503.
Hanson, S. J., Matsuka, T., & Haxby, J. V. (2004). Combinatorial codes in ventral temporal lobe for object recognition: Haxby (2001) revisited: Is there a "face" area? *Neuroimage, 23,* 156–166.
Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., & Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science, 293,* 2425–2430.
Ishai, A., Ungerleider, L. G., Martin, A., Schouten, J. L., & Haxby, J. V. (1999). Distributed representation of objects in the human ventral visual pathway. *Proceedings of the National Academy of Sciences, U.S.A., 96,* 9379–9384.
Jessen, F., Heun, R., Erb, M., Granath, D., Klose, U., & Papassotiropoulos, A. (2000). The concreteness effect: Evidence for dual-coding and context availability. *Brain and Language, 74,* 103–112.
Just, M. A., Cherkassky, V. L., Aryal, S., & Mitchell, T. M. (2010). A neurosemantic theory of concrete noun representation based on the underlying brain codes. *PLoS ONE, 5,* e8622.
Kamitani, Y., & Tong, F. (2005). Decoding the visual and subjective contents of the human brain. *Nature Neuroscience, 8,* 679–685.
Kousta, S. T., Vigliocco, G., Vinson, D. P., Andrews, M., & Del Campo, E. (2011). The representation of abstract words: Why emotion matters. *Journal of Experimental Psychology: General, 140,* 14–34.
Kranjec, A., Cardillo, E. R., Schmidt, G. L., Lehet, M., & Chatterjee, A. (2012). Deconstructing events: The neural bases for space, time, and causality. *Journal of Cognitive Neuroscience, 24,* 1–16.
Kriegeskorte, N., Mur, M., & Bandettini, P. (2008). Representational similarity analysis—Connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience, 2,* 4.
Kriegeskorte, N., Mur, M., Ruff, D. A., Kiani, R., Bodurka, J., Esteky, H., et al. (2008). Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron, 60,* 1126–1141.
Lenat, D. B., & Guha, R. V. (1990). Building large knowledge-based systems: Representation and inference in the {CYC} Project. Addison-Wesley, Reading, Massachusetts.
Lynott, D., & Connell, L. (2009). Modality exclusivity norms for 423 object properties. *Behavior Research Methods, 41,* 558–564.
Mahon, B. Z., & Caramazza, A. (2011). What drives the organization of object knowledge in the brain? *Trends in Cognitive Science, 15,* 97–103.
Malach, R., Levy, I., & Hasson, U. (2002). The topography of high-order human object areas. *Trends in Cognitive Science, 6,* 176–184.
Martin, A., & Chao, L. L. (2001). Semantic memory and the brain: Structure and processes. *Current Opinion in Neurobiology, 11,* 194–201.
McRae, K., & Cree, G. S. (2002). Factors underlying category-specific semantic deficits. In E. M. E. Forde & G. W. Humphreys (Eds.), *Category-specificity in brain and mind* (pp. 211–249). East Sussex, UK: Psychology Press.
Mitchell, T. M., Shinkareva, S. V., Carlson, A., Chang, K.-M., Malave, V. L., Mason, R. A., et al. (2008). Predicting human brain activity associated with the meaning of nouns. *Science, 320,* 1191–1195.
Olivetti, E., Greiner, S., & Avesani, P. (2012). Testing multiclass pattern discrimination. In *IEEE International Workshop on Pattern Recognition in NeuroImaging* (pp. 57–60). London: IEEE.
O'Toole, A. J., Jiang, F., Abdi, H., & Haxby, J. V. (2005). Partially distributed representations of objects and faces in ventral

temporal cortex. *Journal of Cognitive Neuroscience, 17,* 580–590.
Papeo, L., Rumiati, R. I., Cecchetto, C., & Tomasino, B. (2012). On-line changing of thinking about words: The effect of cognitive context on neural responses to verb reading. *Journal of Cognitive Neuroscience, 24,* 2348–2362.
Patterson, K., Nestor, P. J., & Rogers, T. T. (2007). Where do you know what you know? The representation of semantic knowledge in the human brain. *Nature Reviews Neuroscience, 8,* 976–987.
Peelen, M. V., Romagno, D., & Caramazza, A. (2012). Independent representations of verbs and actions in left lateral temporal cortex. *Journal of Cognitive Neuroscience, 24,* 2096–2107.
Pianta, E., Bentivogli, L., & Girardi, C. (2002). MultiWordNet: Developing an aligned multilingual database. In *Proceedings of the First International Conference on Global WordNet*. Mysore, India.
Poesio, M., & Rieser, H. (2010). Completions, coordination, and alignment in dialogue. *Dialogue and Discourse, 1,* 1–89.
Pustejovsky, J. (1995). *The generative lexicon*. Cambridge, MA: MIT Press.
Quadflieg, S., Etzel, J. A., Gazzola, V., Keysers, C., Schubert, T. W., Waiter, G. D., et al. (2011). Puddles, parties, and professors: Linking word categorization to neural patterns of visuospatial coding. *Journal of Cognitive Neuroscience, 10,* 2636–2649.
Rodríguez-Ferreiro, J., Gennari, S. P., Davies, R., & Cuetos, F. (2011). Neural correlates of abstract verb processing. *Journal of Cognitive Neuroscience, 23,* 106–118.
Shinkareva, S. V., Mason, R. A., Malave, V. L., Wang, W., & Mitchell, T. M. (2008). Using fMRI brain activation to identify cognitive states associated with perception of tools and dwellings. *PLoS ONE, 3,* e1394.
Stokes, M., Thompson, R., Cusack, R., & Duncan, J. (2009). Top–down activation of shape-specific population codes in visual cortex during mental imagery. *Journal of Neuroscience, 29,* 1565–1572.
Stokes, M., Thompson, R., Nobre, A. C., & Duncan, J. (2009). Shape-specific preparatory activity mediates attention to targets in human visual cortex. *Proceedings of the National Academy of Sciences, U.S.A., 106,* 19569–19574.
Tomasino, B., Ceschia, M., Fabbro, F., & Skrap, M. (2012). Motor simulation during action word processing in neurosurgical patients. *Journal of Cognitive Neuroscience, 24,* 736–748.
Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., et al. (2002). Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *Neuroimage, 15,* 273–289.
Vigliocco, G., Kousta, S.-T., DellaRosa, P. A., Vinson, D. P., Tettamanti, M., Devlin, J. T., et al. (2013). The neural representation of abstract words: The role of emotion. *Cerebral Cortex*. doi:10.1093/cercor/bht025.
Vinson, D. P., Vigliocco, G., Cappa, S., & Siri, S. (2003). The breakdown of semantic knowledge along semantic field boundaries: Insights from an empirically-driven statistical model of meaning representation. *Brain and Language, 86,* 347–365.
Wang, J., Baucom, L. B., & Shinkareva, S. V. (2012). Decoding abstract and concrete concept representations based on single-trial fMRI data. *Human Brain Mapping, 34,* 1133–1147.
Wang, J., Conder, J. A., Blitzer, D. N., & Shinkareva, S. V. (2010). Neural representation of abstract and concrete concepts: A meta-analysis of neuroimaging studies. *Human Brain Mapping, 31,* 1459–1468.
Warrington, E. K., & Shallice, T. (1984). Category-specific semantic impairment. *Brain, 107,* 829–854.
Wiemer-Hastings, K., & Xu, X. (2005). Content differences for abstract and concrete concepts. *Cognitive Science, 29,* 719–736.
Wilson-Mendenhall, C. D., Kyle Simmons, W., Martin, A., & Barsalou, L. W. (2013). Contextual processing of abstract concepts reveals neural representations of nonlinguistic semantic content. *Journal of Cognitive Neuroscience, 25,* 920–935.