Oxford Handbooks Online

Neural Representations of Concept Knowledge

Andrew J. Bauer and Marcel A. Just

The Oxford Handbook of Neurolinguistics
Edited by Greig I. de Zubicaray and Niels O. Schiller

Print Publication Date: May 2019 Subject: Linguistics, Language and Cognition
Online Publication Date: Mar 2019 DOI: 10.1093/oxfordhb/9780190672027.013.21

Abstract and Keywords

This chapter reviews progress made by brain-reading (neurosemantic) studies that use multivariate analytic methods to delineate the nature, content, and neuroanatomical distribution of the neural representation of concept knowledge in semantic memory. Concept knowledge underlies almost all human thought, communication, and daily activity. The chapter describes how neurosemantic research has provided initial answers to such prominent questions as: What types of information are encoded in a given neural concept representation? To what extent are neural concept representations common across different people? Do neural concept representations evoked by pictures differ from those evoked by language? How are abstract versus concrete concepts represented in the brain? How does the neural representation of a concept evolve while a new concept is being learned? What are the properties and implications of the data analytic techniques that are used in this research area? The initial answers to these questions illuminate how the properties of brain organization impose a structure on the neural representations of concepts.

Keywords: neural representation, concept representation, concept, semantic memory, fMRI, MVPA, machine learning

Introduction

A key goal of cognitive neuroscience is to delineate the nature, content, and neuroanatomical distribution of the neural representation of concept knowledge, which underlies human thought, communication, and daily activities, from small talk about well-worn topics to the learning of quantum physics. Accordingly, research that identifies the neural systems that underlie different categories of concept knowledge (e.g., concepts of animals, tools, and numbers) has made significant advances, particularly research on object concepts (see Martin, 2007). Earlier methods that investigated the neural representation of concept knowledge included the study of deficits in concept knowledge in brain-damaged patients, as well as univariate analyses of activation in the healthy brain using functional magnetic resonance imaging (fMRI).

More recent neuroimaging research is uncovering the fine-grained spatial patterns of brain activation (e.g., multi-voxel patterns) evoked by individual concepts. These brain-reading or neurosemantic studies have generally shown that the spatial pattern of activation that is the neural signature of the concept is distributed across multiple brain regions, where the regions are presumed to encode or otherwise process different aspects of a concept. Conventional linear model-based univariate analyses of activation levels (e.g., statistical parametric mapping in fMRI; Friston et al., 1994), which do not take account of multi-voxel patterns, have typically detected only a small number of brain areas involved in concept representation. Although the idea of multivariate pattern analysis of activation data is not new (Cox & Savoy, 2003), neurosemantic methods have enabled a paradigm shift in studying how concepts are neurally represented.

This chapter summarizes some key research findings that have characterized where and how different types of concept knowledge are represented in the brain. The focus of this chapter is on studies of the neural representations of concepts, rather than on (p. 520) the brain regions that support and mediate semantic processing (see Binder, Desai, Graves, & Conant, 2009, for a meta-analytic review of the neural systems that underlie semantic processing; for other approaches, see, in this volume, Musz & Thompson-Schill, Chapter 22, and Garcea & Mahon, Chapter 23). Because most of the neuroimaging research reviewed here used blood oxygenation level-dependent (BOLD) fMRI (see Heim & Specht, Chapter 4 in this volume), *brain activation* henceforth refers to data collected using fMRI unless stated otherwise (e.g., magnetoencephalography, or MEG; see Salmelin, Kujala, & Liljeström, Chapter 6 in this volume).

The majority of this chapter details how neurosemantic research has illuminated various prominent questions in ways that build on the results of conventional data analytic methods. Some of these questions are the following: Do neural concept representations evoked by pictures differ from those evoked by words? What types of information are encoded in a given neural concept representation? To what extent are neural representations common across different people? What are the differences between the neural representations of abstract versus concrete concepts? The chapter includes a brief survey of the neurosemantic methods that are used to anatomically localize and characterize the kinds of information that are neurally represented.

The chapter ends by summarizing the results of neurosemantic studies that characterize the changes in neural concept representations during the learning of new concepts, a topic that has received little attention. The findings from these studies provide a foundation for cognitive neuroscience to trace how a new concept makes its way from the words and pictures used to teach it, to a neural representation of that concept in a learner’s brain. Monitoring the growth of a new neural concept representation has the potential for further illuminating how concepts are stored and processed in the brain.

The Spatially Distributed Nature of a Neural Concept Representation

Human beings are capable of thinking about a vast number of concepts at various levels of abstraction. This variety of ideas and abstractions is reflected in everyday vocabulary and technical terminology (although not all concepts are necessarily expressible in language). One approach to characterizing concepts in terms of semantically related words is to construct a lexical database, as the authors of WordNet have done. WordNet is an English lexical database in which nouns, verbs, adjectives, and adverbs are grouped into sets of synonyms, where each set constitutes a distinct individual concept. It is a semantic network that consists of 117,659 concepts, each of which is connected to other concepts through a chain of semantic relations (WordNet 3.1, http://wordnet.princeton.edu). WordNet was originally created to be consistent with hierarchical propositional theories of semantic memory, which postulate that concepts are organized hierarchically from general to specific concepts (e.g., Collins & Quillian, 1972). The most (p. 521) common type of semantic connection between words is the hierarchical "is-a" relation. The concept *chair*, for example, is related to *furniture* by an "is-a" connection. As indicated in WordNet, concepts that human beings can think about range from thoughts of physical objects, such as organisms and geological formations, to abstractions, such as psychological states and mathematical entities.

The ability to study how this range of concepts is represented neurally has only recently become possible since the development of data analytic methods that can detect a correspondence between a distributed brain activation pattern and an individual concept. Neurosemantic research initially focused on only a small fraction of this range of concepts, namely animate and inanimate object concepts such as animals, faces, and tools and other manmade objects (e.g., Haxby et al., 2001; Mitchell et al., 2003; for an account of early neurosemantic research, see Haxby, 2012). These choices of concept categories were motivated by previous clinical studies of object category-specific agnosia, and also by univariate analysis-based neuroimaging findings that elaborated on the clinical results. Clinical studies found that relatively selective cortical damage was associated with a disproportionate deficit in concept knowledge for one of a small set of categories (e.g., animals or tools; for a review of the clinical literature, see Capitani, Laiacona, Mahon, & Caramazza, 2003). This body of work suggested that concept categories were subserved by only a few brain regions. However, mapping large brain areas to single-concept categories does not provide an account of neural concept representations that scales to the huge number of concepts that must be represented. A more efficient scheme that can accommodate vast numbers of concepts would be a pattern-encoding scheme, such that the neural representation of a concept corresponds to a spatial pattern of activation of many individual voxels, each displaying a level of activation that is characteristic of the concept. Early neurosemantic research provided the empirical basis for pattern encoding by indicating that concept knowledge might be represented in neural populations distributed over a large number of brain areas.

Since the early fMRI research on concrete object concepts, neurosemantic research has replicated the finding that neural concept representations span multiple brain regions, and has revealed the activation patterns associated with other types of concept knowledge, such as emotions (Baucom, Wedell, Wang, Blitzer, & Shinkareva, 2012; Kassam, Markey, Cherkassky, Loewenstein, & Just, 2013), numbers (Damarla & Just, 2013 Eger et al., 2009), personality traits (Hassabis et al., 2013), and social interactions (Just Cherkassky, Buchweitz, Keller, & Mitchell, 2014). In one study, close to 2,000 individual object and action concepts were each localized to multiple brain areas (Huth, Nishimoto, Vu, & Gallant, 2012). Figure 21.1 shows that the neural representations of these object and action concepts each reside in multiple areas distributed throughout the brain. The figure contains color-coded mappings between various concepts and their representations in various areas. For example, concepts related to communication (cyan) were found to be represented in auditory sensory cortex in the temporal lobe and a frontal area that includes Broca’s area, a canonical language region.

The main theoretical interpretation regarding spatially distributed neural representations is that the multiple brain areas that conjointly represent a given concept (p. 522) correspond to the brain systems that are involved in the physical and mental interaction with the concepts’ referents. For example, the concept of a knife entails what it looks like, what it is used for, how one holds and wields it, and so on, resulting in a neural representation distributed over sensory, perceptual, motor, and association areas.

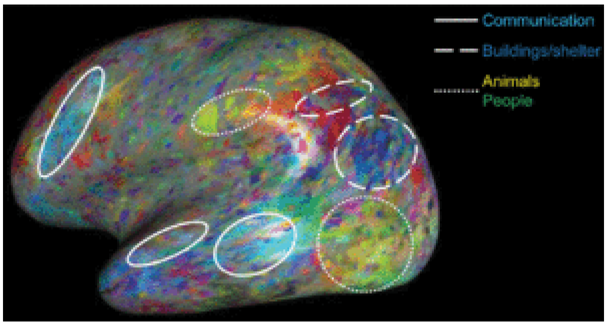

Figure 21.1. Neural concept representations are distributed throughout the brain. In Huth et al. (2012), 1,705 individual object and action concepts that appeared in movies were each found to be represented over multiple brain areas. Indicated by the three types of ellipses are the major brain areas associated with some of the superordinate categories of these object and action concepts. Auditory sensory cortex in the temporal lobe and a frontal language area were associated with communication; postcentral gyrus (sensation and movement) and occipitotemporal cortex (visual) were associated with biological entities; and parietal (spatial) and occipital areas were associated with buildings and shelter.

Source: The figure corresponds to one participant’s inflated brain, and was extracted from http://gallantlab.org/semanticmovies.

These findings from multivariate analyses build on and are consistent with past univariate analysis-based research. For example, nouns that refer to physically manipulable objects such as a knife have been shown to activate left premotor cortex in right-handers (Lewis, 2006). In addition to left premotor cortex, activation has been observed in additional regions but at lower magnitudes, a result that hinted at the greater spatial distribution of concept knowledge in the brain (Chao, Weisberg, & Martin, 2002).

In behavioral cognitive science, a concept is often treated as a mental representation that specifies some of the dimensions of a real-world phenomenon (e.g., visual or tactile properties of an object), in addition to the relations among those dimensions (see Barsalou, 1992, for a discussion of the nature of concept representation). Consistent with this approach is the finding that multiple brain regions, which encode different dimensions, collectively contain the information about a single concept. For example, the concept *cat* might include dimensions of cats that are common across different (p. 523) breeds, such as general body shape, locomotion, diet, temperament, and so on. These properties should be detectable in the brain activation pattern associated with the concept *cat*. Several studies have used regression models to predict the activation pattern of a given object concept, based on how different voxels are tuned to various dimensions of objects and on how important those dimensions are to defining a given object concept. Accurate predictions have been made using properties of objects generated by human participants (Chang, Mitchell, & Just, 2011) or extracted from text corpora such as web-based articles (Mitchell et al., 2008; Pereira, Botvinick, & Detre, 2013). In addition, an MEG study that used properties generated by participants predicted the spatial pattern of evoked magnetic fields associated with an object concept (Sudre et al., 2012).

The discussion here has assumed that sensorimotor systems in the brain store or otherwise process information that is integral to the comprehension of a concept, particularly object concepts. In this view, the representations of some concepts entail body-object interaction information; that is, they are *embodied* representations (Barsalou, Santos, Simmons, & Wilson, 2008). Some alternative theories hold that the brain activation observed in sensorimotor regions reflects imagery or simulated motion that occurs only after conceptual processing, and that fundamental concept meaning is encoded only in association areas such as the anterior medial temporal lobe (for a review of the competing theories, see Mahon & Caramazza, 2008; Meteyard, Cuadrado, Bahrami, & Vigliocco, 2010; and Kiefer & Pulvermüller, 2012). However, several studies using words referring to concrete objects have shown that sensorimotor activity evoked by the words occurs too early to originate from imagery explicitly generated by the participants (e.g., Kiefer, Sim, Herrnberger, Grothe, & Hoenig, 2008; see also Martin, 2007), providing some evidence for the embodied view of the representations of certain concepts.

Characterizing the Semantic Dimensions That Underlie Neural Concept Representations

A central objective of neurosemantic research is to determine some of the main semantic dimensions that underlie the neural representation of a given concept. This objective can be reached by reducing an activation pattern's dimensionality (often consisting of the activation levels of tens to hundreds of voxels) to determine the main factors underlying the representation. For example, in a study of emotions (e.g., anger, disgust, envy, fear, happiness, lust, pride, sadness, and shame), factor analysis of the activation data indicated that each emotion was represented with respect to four underlying dimensions:

valence, arousal, sociality, and *lust* (Kassam et al., 2013). Each of these dimensions was further localized to plausible networks of brain regions. *Arousal*, for example, was localized to basal ganglia and precentral gyrus, which have previously (p. 524) been implicated in action preparation. The *sociality* dimension (which was not previously recognized as a core dimension of emotions) was traced to anterior and posterior cingulate cortex, two default mode network regions previously shown to be involved in social cognition. Although most of the dimensions were traced to brain areas previously implicated by univariate analysis-based research, it is notable that this single neurosemantic study uncovered results comparable to the results of multiple conventional neuroimaging studies.

A similar analysis of the activation patterns evoked by 60 object concepts identified three key dimensions: *manipulation* (e.g., tools and other manipulable objects), *eating* (e.g., vegetables, kitchen utensils), and *shelter* (e.g., dwellings, vehicles) (Just, Cherkassky, Aryal, & Mitchell 2010). *Manipulation* was associated with left postcentral/supramarginal gyrus and left inferior temporal gyrus, which have previously been implicated in the processing of tool concepts (Lewis, 2006). The *eating* dimension was traced to left inferior temporal gyrus and left inferior frontal gyrus, which revealed a link between representations of tool concepts (namely kitchen utensils) and representations of face- and jaw-related actions (Hauk, Johnsrude, & Pulvermüller, 2004). Finally, the *shelter* dimension was traced to bilateral parahippocampal gyrus and bilateral precuneus. The parahippocampal gyrus is well known to activate in response to information about dwellings and scenes (e.g., Epstein & Kanwisher, 1998), and the precuneus areas were anatomically close to retrosplenial cortex, which is thought to be involved in the comprehension of a scene within a larger environment (for a review on retrosplenial cortex, see Vann, Aggleton, & Maguire, 2009).

One study greatly expanded the range of concepts whose neural representations were uncovered by collecting activation data as participants watched several hours of movies (Huth et al., 2012). The investigators used WordNet to label 1,364 common objects (nouns) and actions (verbs) that appeared in the movies, and an additional 341 superordinate categories were inferred using the hierarchical relationships in WordNet (e.g., *canine* and *mammal* were added if *wolf* was an object that appeared in a movie). Data reduction yielded four dimensions that were interpretable: *mobility/animacy, sociality* (e.g., words about people and communication), *civilization* (e.g., people, man-made objects, vehicles), and *biological entities*. Interestingly, these dimensions partially overlap with the dimensions revealed by the two other neurosemantic studies that separately investigated object and emotion concepts (described earlier). Thus, different studies using different methodologies appear to be converging on a common set of underlying neural dimensions of representation.

Integration among a Concept's Semantic Dimensions

A central aspect of a concept is how the different semantic dimensions relate to each other. Although some of the underlying dimensions of certain types of concept knowledge are being identified, much less is known about the relations among the dimensions. For example, some of the key dimensions of concrete objects appear to be *manipulation, eating*, and *shelter*. Is a neural representation of an object concept then anything (p. 525) more than the sum of the representations of the concept's individual dimensions? For example, would there be some indication in a neural representation that a gingerbread house is fundamentally different from a cafeteria building, even though both involve information related to *eating* and *shelter*? It is unclear whether relations among the dimensions of a concept are represented in brain areas that are spatially distinct from the locations of the individual dimensions (e.g., convergence zones; Damasio, 1989), or whether relational information is somehow encoded in a distributed way across the set of areas that also represent the individual dimensions.

A multivariate analysis of the activation in a visual perception task investigated which brain areas encode the conjunction of separate dimensions (Seymour, Clifford, Logothetis, & Bartels, 2009). The dimensions that were combined were the color and direction of motion of a set of dots, which were either green or red and rotated either clockwise or counterclockwise. The specific conjunction of color and direction of motion of a given item was found to be represented in multiple areas of early visual cortex that also encode the individual dimensions, indicating that these areas contain both a representation of the individual dimensions and the relation between the dimensions. This integration of information might be a critical aspect of the representation of a cohesive percept.

On the other hand, there is also evidence that the relations among a concept's component dimensions are represented exclusively within specific high-order brain areas or convergence zones. An anterior temporal lobe region has been suggested as a site of dimension integration because it is innervated by different sensory modalities, and because abnormal functioning of this region is associated with impairments to semantic processing, but not to the performance of non-semantic cognitive tasks (Pobric, Jefferies, & Ralph, 2007; Patterson, Nestor, & Rogers, 2007). Coutanche and Thompson-Schill (2014) demonstrated that the conjunction of an object's dimensions was encoded in a multi-voxel pattern in anterior temporal lobe, but not in the areas that separately represent the individual dimensions. Specifically, the depicted objects were fruits and vegetables, and the dimensions were color and shape (whose representations were investigated in fusiform gyrus and occipitotemporal cortex, respectively). Furthermore, the representation of the conjunction was detectable by the investigators only when each dimension's representation could be detected, strengthening the evidence for the conclusion that the anterior temporal region represents the integration of individual components.

Thus the evidence is mixed as to whether the integration of information about separate dimensions is represented in high-order brain regions versus in the network of areas that underlie the individual dimensions. It is also possible that integrated information is encoded in both representational formats.

Neurosemantic Methodology

Neurosemantic data analyses attempt to detect multi-voxel patterns of brain activation that correspond to the thoughts of concepts. One virtue of this approach is that it adheres to the fundamental principle that thinking is a network function involving (p. 526) multiple brain systems, including thinking about a concept. A second advantage of this approach is that it bestows a greater sensitivity for discovering the underlying phenomenon, by virtue of concurrently assessing the activations of many voxels with similar activation patterns for the stimuli at hand, regardless of the voxels' proximity to each other. One phenomenon whose discovery has benefited from this greater sensitivity is the representation of different concepts that are in the same superordinate semantic category. The greater sensitivity of multi-voxel analysis enables researchers to distinguish between the activation patterns of such concepts (e.g., distinguishing between different animal concepts such as a primate and bird; Connolly et al., 2012); that is, the newer methods can compare patterns of activation between individual concepts in spatially distributed voxels. With the use of neurosemantic methods, a finding of various related concepts eliciting unique yet similar activation patterns over a set of brain areas constitutes suggestive evidence that those brain areas store or otherwise process the meanings of the concepts.

By contrast, conventional univariate analyses often report the magnitudes of activation of individual voxels that are averaged over a region of interest, requiring spatial proximity among the voxels that are grouped together (see Poldrack, 2007). Univariate analysis is sometimes not sensitive enough to distinguish between similar experimental conditions because the mean activation level of a set of voxels is often equivalent between similar conditions. Univariate analysis is useful for identifying the brain areas that are *involved* in the processing of some class of concepts, by determining whether an area's activation level is elevated. Figure 21.2 depicts a hypothetical scenario in which the greater sensitivity of multivariate analysis enables distinguishing between two similar conditions, whereas univariate analysis finds no difference between the conditions but establishes for each condition an elevation in mean activation level (for a detailed comparison of the methods, see O'Toole et al., 2007, and Mur, Bandettini, & Kriegeskorte, 2009). Of course, "sensitivity" is assessed with respect to the phenomenon of interest, and there are doubtless phenomena other than multi-voxel activation patterns corresponding to concepts for which univariate analyses may be more sensitive (Coutanche, 2013; Davis et al., 2014).

One commonly used technique in neurosemantic studies is discriminative multivariate pattern classification analysis (MVPA). A classifier is an algorithm that is trained to associate an activation pattern with each of the stimuli (or classes of stimuli) and is subsequently reiteratively tested (using a procedure called cross-validation) on an independent data set (for a tutorial, see Pereira, Mitchell, & Botvinick, 2009, and Norman, Polyn, Detre, & Haxby, 2006). Logistic regression is an example of a discriminative classifier. The main strength of MVPA is its concurrent consideration of the activations of multiple voxels, regardless of their relative locations in the brain. MVPA has been used to discover neural representations of various types of information, such as covert intentions in prefrontal cortex (Haynes et al., 2007), visual imagery of simple shapes in occipitotemporal cortex (Stokes, Thompson, Cusack, & Duncan, 2009), and episodic memories in the hippocampus (Chadwick et al., 2010). The accuracy of the classification is a measure of the discriminability of the stimuli (or classes of stimuli), sometimes (p. 527) computed as the rank accuracy, or the normalized percentile rank of a correct stimulus in the classifier's ranked output (Mitchell et al., 2004). Here, chance level is a normalized rank accuracy of 0.5, where the correct classification response occupies the middle rank among all possible responses. The obtained accuracy can then be compared to a distribution of accuracies that would be obtained by chance (typically obtained by Monte Carlo simulations).

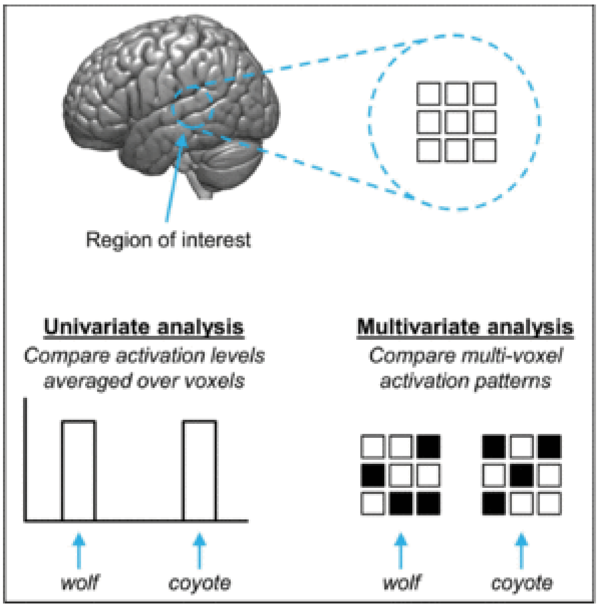
Region of interest
Univariate analysis
Compare activation levels averaged over voxels
Multivariate analysis
Compare multi-voxel activation patterns
wolf
coyote
wolf
coyote

Figure 21.2. Comparison of univariate and multivariate data analysis. A hypothetical scenario in which multivariate analysis (*right*) reveals that the multi-voxel *pattern* of activation levels in left primary auditory cortex differs between two animal concepts (i.e., two animals that make similar sounds). Univariate analysis (*left*) shows that the level of activation *averaged* over the set of voxels is the same between the two concepts, but establishes an elevation in mean activation for both concepts.

Methods That Assess the Semantic Content of Neural Representations

Importantly, what is neurally represented is not just the meaning of a concept, but also the perceptual form of the word or picture that refers to the concept. The perceptual form is represented in primary and secondary sensory brain regions. Thus, simply obtaining accurate classification of an activation pattern does not ensure that the pattern encodes concept meaning. To substantiate that a given activation pattern corresponds to the meaning of a concept, it is sometimes useful to show that the set of correlations among the activation patterns bears a clear relation to behavioral judgments of similarity (p. 528) among the concepts. A statistically reliable correlation between the two sets of inter-item similarities would provide converging evidence that concept meaning underlies the systematic differences in the activation data. Another way to ensure that an MVPA is identifying the representation of a concept—and not the representation of the picture or word that evokes it—is to exclude sensory and early perceptual area voxels from the analysis.

Although discriminative classifiers are extremely useful for associating brain areas with stimuli, they do not easily lend themselves to predictive or generative modeling, as a generative classifier can do. A generative classifier is useful if there is a need to predict the activation that will be evoked by a new stimulus. The central property of a generative classifier is its postulation of a set of intermediate variables between the stimulus and activation that modulate the activation as a function of the properties of the stimulus. The classic method for predictive modeling is regression, which can be used to predict the activation of a yet unseen stimulus, based on how its properties modulated the activation of the stimuli on which the model was trained. Predictive regression models can provide converging evidence for the neural representation of concept meaning in terms of the postulated semantic dimensions that underpin the neural representation of a concept. A model is first estimated of how a set of voxels is tuned to different dimensions (e.g., the size or animacy of an object). A prediction is then made of a concept's activation pattern based on the weighted importance of the dimensions in the representation of that concept. The generalizability of the model can be assessed by testing the predicted activation pattern of any concept that is definable by the dimensions included in the model (Naselaris, Kay, Nishimoto, & Gallant, 2011). As mentioned previously, it is possible to accurately predict an object concept's activation pattern using properties of objects generated by human participants (Chang et al., 2011; Sudre et al., 2012). This approach has also been used to investigate how different voxels in visual brain areas are tuned to different visual features (Kay, Naselaris, Prenger, & Gallant, 2008). The general goal of this approach is to relate concept properties to one or more areas of activation.

Another methodology that can help characterize the neural representation of a concept is representational similarity analysis (RSA), which assesses the neural similarity between all pairs of items and relates the resulting similarity structure to the activation patterns (see Musz & Thompson-Schill, Chapter 22 in this volume). Some researchers use the idea of an *n*-dimensional representational space, in which the distance between a given pair of concepts approximates the degree of similarity between the concepts' activation patterns. If the multi-voxel activation pattern for two concepts is known, then the similarity (or distance) between them can be computed and the full set of inter-item distances can be used to specify the space. In this approach, the set of inter-concept distances can reveal the kinds of information that underlie the representations (see Kriegeskorte, Mur, & Bandettini, 2008, for a tutorial on RSA). For example, Figure 21.3 shows the similarity structure of six biological species concepts, corresponding to two different brain areas. The similarity structure of the concepts differs between the two areas, indicating that the information that is encoded or otherwise processed differs between the areas. The information represented in occipitotemporal cortex is organized (p. 529) with respect to species category, given that the neural dissimilarity is lowest between the two primates, between the two birds, and between the two insects. On the other hand, the information in early visual cortex seems to encode visual properties of the concepts that do not correlate with species category. In this way, representational similarity structures have the potential to reveal the underlying dimensions along which concepts are organized in the brain.

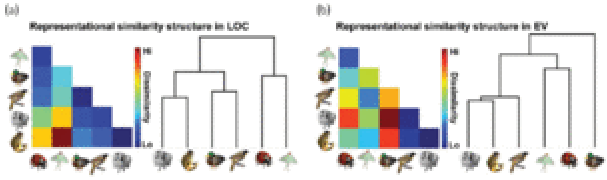

Figure 21.3. Representational similarity structures in different brain areas reveal differences in the types of information neurally represented therein. Shown for two different brain areas are the dissimilarities between all pairs of biological concepts' activation patterns, and the hierarchical structure that emerges from these dissimilarities. The representational dissimilarity between any two concepts was computed as 1 – the correlation between their activation patterns. (A) Lateral occipital complex (LOC), a high-order visual brain area that here represents information corresponding to species category. (B) Early visual cortex (EV), a collection of brain areas that process low-level visual features, which seems to encode visual properties of the concepts that do not correlate with species category.

Source: The figure was adapted with permission from Connolly et al. (2012).

The capacity of RSA to reveal the semantic content encoded in neural representations depends in part on the measures of dissimilarity between a pair of vectors of activation levels, such as correlational measures (e.g., 1—the Pearson correlation between multi-voxel activation patterns), or geometric distances (e.g., Euclidean or Mahalanobis distance). Another type of measure of neural dissimilarity is the classification accuracy in a classifier's confusion matrix. A comparison of these dissimilarity measures for RSA has revealed that continuous distances (i.e., correlation and geometric distances) produce more reliable results than classification accuracies, largely because classification accuracies are obtained from binary decisions that discard continuous dissimilarity information (Walther et al., 2016). Furthermore, dissimilarity measures that are cross-validated across subsets of activation data provide an interpretable zero point against noise.

Apart from RSA, other data-driven, exploratory methods are used to characterize the informational content contained in neural representations. To identify key underlying dimensions from a large set of voxels spanning multiple brain regions, dimension reduction techniques, such as principal or independent components analysis or factor analysis, are useful (see Heim & Specht, Chapter 4 in this volume). These dimension reduction methods can separate the activation patterns into smaller sets of voxels (which (p. 530) maximize the amount of total or shared variance explained in the data), where each set is associated with one or more of the dimensions (e.g., Just et al., 2010; Kassam et al., 2013). If some of the voxels associated with one of the dimensions are localized primarily in the motor cortex, for example, then it is likely that motor action constitutes part of the semantic content of that dimension.

Even with the use of advanced neurosemantic methods, caution may be needed in concluding that the activation pattern in a set of brain regions represents the meaning of a particular concept because of the notorious difficulty in distinguishing representation from process (Anderson, 1978). It is sometimes difficult to distinguish whether an activation pattern corresponds to where and how information is *stored*, versus corresponding to the processes that operate on the representation. Measurement of a neural concept representation requires evoking an activation pattern, thus potentially conflating representation and process; for example, the content of a neural concept representation might be a facet of processing related to selective attention. Selective attention has been shown to change the tuning characteristics of occipitotemporal and frontoparietal cortex for the objects shown in a movie (Çukur, Nishimoto, Huth, & Gallant, 2013). A way to address this type of difficulty might be to test whether characteristics of the activation patterns vary as a function of the nature of the task that the participants perform. It will be useful for future research to characterize neural concept representations in a way that takes into account the nature of the processing that evokes that concept, for example in sentence comprehension (Poeppel, 2012), story comprehension (Wehbe et al., 2014), and problem-solving (Anderson & Fincham, 2014).

Methods of Evoking a Neural Concept Representation

Various methods have been used to evoke the brain activation that underlies a concept. They vary with respect to three characteristics: (1) the amount of time allotted for a participant to process and think about a concept; (2) the nature of the task that the participant is asked to perform; and (3) the modality of the stimulus used to evoke a concept.

Each method of evoking a concept has its own profile of advantages and disadvantages; for example, allotting more time to think about a concept can yield more robust signal in the activation data and thus greater classification accuracy. However, if there is no instruction to think about a concept in a certain way, greater thinking time may result in variation in the activation across the repetitions of a concept, due to the different ways in which a participant may think about a concept. Furthermore, an instruction to think about a concept in a particular way or context may induce an unrepresentative instantiation of the concept (e.g., thinking about *dog* as a participant in a race).

A study in which the participants were presented the same emotion concepts under different task instructions provided evidence of the commonality of the neural representation across two very different task conditions (Kassam et al., 2013). In one condition, participants were presented with emotion words (such as "anger") and were instructed to evoke thoughts and feelings associated with each emotion. In a different condition, (p. 531) participants passively viewed pictures that evoked disgust. A classifier that was trained on the activation evoked by the emotion words (which included "disgust") was then able to identify the disgust evoked by the pictures with good accuracy. This finding provided evidence that, at least in this case, the brain activation patterns corresponding to disgust in these two very different conditions were fairly similar to each other. It would be useful to see many other concept representations compared, under many different conditions, to determine which facets of a neural representation are always activated and which are modulated by the nature of the evoking task.

Influences of Language on Neural Concept Representations

One task effect of long-standing interest involves the difference in the content of a neural concept representation depending on whether the evoking stimulus is a word versus a picture. For example, is a picture of a screwdriver more likely than the word "screwdriver" to evoke a specific, potentially unrepresentative instantiation of the concept *screwdriver*, especially if the picture is richly detailed? A resolution of this issue would provide a theoretical framework to account for the results of numerous studies that use words, pictures, movies, or other stimuli to evoke a concept.

A neurosemantic study uncovered suggestive evidence that the central aspects of a neural concept representation are to a large extent independent of the stimulus used to evoke the concept (Shinkareva, Malave, Mason, Mitchell, & Just, 2011). In this study, it was possible to classify the activation pattern of an object concept cued by the noun naming the object with a classifier trained on the activation pattern of the same concept evoked by a simple line drawing, and vice versa. Specifically, the classifier determined whether a given object concept referred to a tool or dwelling. Although this study assessed only a small number of items from only two categories, it is suggestive of a common core neural representation that is evoked regardless of the stimulus modality.

In Shinkareva et al. (2011), it was possible to classify the words or pictures using activation from the language system (left inferior frontal gyrus) and also from sensorimotor brain regions. These results are consistent with the Language and Situated Simulation theory of semantic processing (akin to the embodied cognition approach), which holds that a concept activates both the language system and the same sensorimotor regions that are active during actual interaction with the concepts' referents (Barsalou et al., 2008; Simmons, Hamann, Harenski, Hu, & Barsalou, 2008).

Despite there being a shared core of the neural representation between pictures and words referring to a concept, there is evidence of differences in the neural concept representations. A possible asymmetry between pictures and words is that pictures evoke not only a concept's core meaning, but also some detailed instantiation of the concept as it is depicted in a picture. A picture generally contains a greater amount of information (p. 532) about an object than does a word (e.g., the shape of a screwdriver's handle and its tip). Words, on the other hand, tend to evoke only the most generic properties of a concept. In the study of the cross-stimulus modality classification described earlier (Shinkareva et al., 2011), the classification accuracy was higher when the classifier was trained on word-cued activation and tested on pictures, versus when it was trained on pictures and tested on words. This result suggests that although the neural representations of words and pictures are similar to each other, pictures activate additional information that is specific to the picture. The classifier that was trained on pictures and tested on words apparently extracted some information unique to the pictures, leading to a classification accuracy that was lower than when the classifier was trained on words and extracted generic information common to the picture and word representations.

In sum, there is evidence of overlap in the semantic content between word- and picture-cued neural representations. However, additional research is needed to characterize the distinctions between the content of representations evoked by words versus pictures. Can identical representations be evoked between words and pictures by manipulating the information that is directly expressed in either presentation modality? For example, would the addition of modifiers to a word increase the amount of information about the evoked concept (e.g. "short, yellow Phillips screwdriver")? Similarly, can the neural representation of a concept evoked by a picture be made more similar to the one evoked by a word by making the picture completely schematic, thereby removing the extra information conveyed by the picture? Empirical studies that address such issues would enable a refinement of theories of how concept knowledge is neurally stored and activated (e.g., dual-coding theory; Paivio, 1986).

Neural Representations of Lexical and Grammatical Categories

There has not yet been found a clear difference in activation patterns evoked by different word classes (e.g., verbs versus nouns). Studies that have addressed this question have been hampered by the problem that concepts corresponding to different grammatical categories are inherently different. For example, verbs are typically associated with concrete actions, whereas nouns typically refer to objects (see Vigliocco, Vinson, Druks, Barber, & Cappa, 2011, for a review of the brain activation underlying nouns and verbs). These grammatical categories also differ in terms of lexical stress and ortho-phonological typicality, which complicate matters further (Arciuli, McMahon, & de Zubicaray, 2012). Consequently, any differences in activation found between nouns and verbs might plausibly be attributed to differences in semantic content, rather than to (non-semantic) differences in lexical category or grammatical structure. One study compared the activation evoked by abstract nouns and verbs (e.g., "idea" versus "think") because abstract nouns and verbs are not associated with concrete objects or concrete actions and thus are not inherently different in the same way that concrete nouns and (p. 533) verbs are (Moseley & Pulvermüller, 2014). This study found no activation location difference between abstract verbs and nouns, whereas there was a difference between concrete verbs and nouns. The authors concluded that there are no word class-specific processing centers in the brain. However, the analysis focused on only a small set of regions of interest; a whole-brain comparison was not conducted, thus leaving the question open to additional investigation.

Other studies have reported differences in the activation locations evoked by pseudo-nouns versus pseudo-verbs (Shapiro et al., 2005). Pseudo-nouns (nonsense words with morphological cues to lexical category, such as ending in *–age*, cuing a noun) elicited greater activation than pseudo-verbs bilaterally in temporal regions, whereas pseudo-verbs (e.g., those ending in *–eve*) evoked greater activity in left-lateralized frontal areas. In another study, pseudo-verbs (e.g., ending in *–eve*) elicited greater activity in motor cortex than pseudo-nouns (de Zubicaray, Arciuli, & McMahon, 2013). Such differential activation evoked by stimuli that are devoid of meaning suggests that a word's lexical class is a possible dimension of lexical organization in the brain.

In a study using multivariate analysis, it was possible to distinguish between the activation patterns associated with semantically equivalent but grammatically different sentences (Allen, Pereira, Botvinick, & Goldberg, 2012). Specifically, a classifier could determine whether a sentence was ditransitive (e.g., "Mike brought a book to Chris") or dative (e.g., "Mike brought Chris a book"), despite the fact that the two grammatical constructions convey the same core information. The classifier used activation from left-lateralized brain areas involved in language processing, such as left inferior frontal gyrus. The result suggests that grammatical category is neurally represented independent of semantic meaning, but further research is needed that identifies the grammatical aspects of ditransitive and dative sentences that might be neurally represented. At the same time, it remains unclear whether grammatically different sentences with the same core meaning still have subtle differences in semantic content that can be detected in activation patterns.

If nouns and verbs tend to evoke differing semantic content (i.e., object- and action-related content, respectively), then some neurosemantic theories would posit that this content is primarily represented in association areas that integrate among different sensorimotor modalities, such as the anterior medial temporal lobe or angular gyrus (e.g., Patterson et al., 2007; for a review of these and other theories, see Mahon & Caramazza, 2008; Meteyard et al., 2010; and Kiefer & Pulvermüller, 2012). In support of this view, abnormal functioning of these brain regions is associated with impairments to semantic processing, but not to the performance of non-semantic cognitive tasks (e.g., Pobric et al., 2007). Because many of the activation differences observed between nouns and verbs are *not* constrained to these association areas, these theories might predict that activation differences between lexical categories could instead reflect non-conceptual linguistic processing. For example, greater activity observed in motor cortex in response to verbs versus nouns could reflect verb-specific ortho-phonological properties (de Zubicaray et al., 2013). However, there is also reason to believe that sensorimotor activation elicited by words instantiates the concepts to which the words refer, which may (p. 534) constitute a form of semantic processing (Mahon & Caramazza, 2008). Thus, additional research that demarcates the semantic system in the brain may provide some answers regarding which brain areas underlie syntactic processing.

Finally, another prominent question addressed by the neurosemantic approach is whether neural concept representations differ between different languages, assuming that the words or phrases are good translation equivalents of each other. There is suggestive evidence that the neural representation of a concept is largely the same, regardless of which language is used to evoke it. In two neurosemantic studies of bilinguals, it was possible to identify the activation pattern associated with an object concept cued in one language based on the activation pattern of the same concept denoted in another language (Buchweitz, Shinkareva, Mason, Mitchell, & Just, 2012; Correia et al., 2014). However, there are subtle clues that neural representations of word classes differ between speakers of different languages due to differences in the semantic content associated with the classes. For example, the most frequently used class of verbs in Spanish refers to the *path* of an object (see Goldstone & Kersten, 2003), whereas English verbs often refer to the manner of an object's *motion*. Thus, neural concept representations might differ between the two languages in this respect, despite a core commonality in the representations.

Commonality of Neural Concept Representations across Different Individuals

One of the most dramatic findings in neurosemantics is that the fine-grained activation pattern corresponding to a given concept is largely common across people. When two people think about the concept *apple*, their activation patterns are distributed over the same brain locations and are very similar. When a classifier is trained on the activation patterns from a set of participants (whose activation data were spatially aligned to a common anatomical template), it can reliably predict which concept a left-out test participant is contemplating. This phenomenon of commonality has been demonstrated for the neural representations of concrete objects (Just et al., 2010), emotions (Kassam et al., 2013), numbers (Damarla & Just, 2013), and social interactions (Just et al., 2014).

The ability to accurately classify concepts across people suggests that some of the same properties of a given concept are evoked in many or most individuals. Behavioral studies in which participants generate properties of objects report that there are some properties that are commonly associated with a given object concept (e.g., Cree & McRae, 2003; Nelson, McEvoy, & Schreiber, 1998). Moreover, a neurosemantic study uncovered suggestive evidence that the most defining properties of a concept are automatically activated during any instance of evoking that concept, even during tasks for which that information is irrelevant (Hsu, Schlichting, & Thompson-Schill, 2014).

(p. 535) Accurate cross-individual classification is somewhat surprising given the uniqueness and variety of personal experiences and associations that might underlie a concept, in addition to the varied levels of experience with a concept. Although the commonality of neural representations of concepts has been demonstrated, the unique aspects of neural representations have yet to be characterized. The unique components could be similar in kind to the common aspects (which constitute semantic memory), or they could be tagged in some way as part of one's autobiographical memory (Charest, Kievit, Schmitz, Deca, & Kriegeskorte, 2014). It will be interesting to determine the characteristics of concept knowledge that are unique, although to do so will be challenging precisely because any emerging pattern of results will be difficult to aggregate over participants. If successful, such research may enable an understanding of how different properties of a neural representation, such as its particular pattern of activity or its anatomical distribution, are shaped by individual factors (e.g., unique experience, genetic predisposition) versus shared, cross-individual factors (e.g., cultural values, evolutionarily conserved biases toward processing certain types of information, and inherent neural constraints; Sadtler et al., 2014). Representational commonality might indicate that there exist category-specific brain networks that process specific kinds of information that are important to survival, such as information about food or shelter (Mahon & Caramazza, 2003).

Methods That Assess Representational Commonality by Abstracting Away from Person-Specific Patterns of Neural Activity

There is another approach that makes it possible to compare neural representations across people, namely cross-individual comparison of the similarity relations among the concepts' activation patterns (Raizada & Connolly, 2012). This approach uses RSA (described earlier) to abstract the activation data away from voxel space (i.e., activation patterns corresponding to different brain locations) to representational similarity space (i.e., correlations among the activation patterns). Thus, the method does not need to warp different participants' activations to a common anatomical template. Classification is not performed directly on the activation patterns. Rather, the classifier determines whether the similarity relations among the concepts' activation patterns are similar across individuals. Cross-individual classification in representational similarity space may produce greater classification accuracy because the method does not need to account for individual differences in brain anatomy. Thus, although there may be slight differences in the precise brain locations of two participants' representations of the same concept, the neural similarity between the two representations is robust to these differences.

Another cross-individual classification method is the mapping of each individual's activation data from original voxel space to a common, high-dimensional space over (p. 536) all the participants (Haxby et al., 2011; Haxby, Connolly, & Guntupalli, 2014). The dimensions in this new common space are not individual voxels, but rather distinct response-tuning functions defined by their commonality across the different brains. This method also results in greater cross-individual classification accuracy.

Yet another cross-individual classification method encodes both activation location and magnitude in a graph structure and is robust to anatomical differences among people (Takerkart, Auzias, Thirion, & Ralaivola, 2014). Thus, the warping of activation data to align with a common anatomical template might lead to an underestimation of the commonality of the semantic content in neural representations.

Intriguingly, one study that used RSA assessed commonality of neural representations between human beings and macaque monkeys (Kriegeskorte, Mur, Ruff, et al., 2008; Kriegeskorte, 2009). This study indicated that pictures of various animate and inanimate objects elicited activity patterns that had similar representational similarity between the humans and monkeys in homologous inferotemporal cortex, a set of high-order visual brain areas. The human data consisted of fMRI activation, and the monkey data were single-neuron electrophysiological recordings from two macaque monkeys (Kiani, Esteky, Mirpour, & Tanaka, 2007). The commonality between the neural representations belonging to the two species is illustrated in Figure 21.4. To the extent that the inferotemporal activation reflected semantic processing rather than perceptual processing of the picture stimuli themselves, this result may bear on the profound question of the nature of thought in other species. The result also motivates future research that uses similar methods to compare neural representations between humans and monkeys within other brain areas, and for other concept categories such as numbers (e.g., Beran, Johnson-Pynn, & Ready, 2011).

Neural Representations of Abstract and Concrete Concepts

Previous behavioral research suggests that how a concept is stored and otherwise processed depends on how concrete or abstract it is. For example, words that refer to concrete concepts (e.g., "ball") are more quickly recognized (e.g., Schwanenflugel & Harnishfeger, 1988), and knowledge of concrete concepts is more resistant to brain damage (see Coltheart, Patterson, & Marshall, 1980). Several fMRI studies have shown that words that refer to either abstract concepts (e.g., "blame") or concrete concepts activate overlapping but partially distinct brain networks, with abstract concepts eliciting greater activation in the frontal language system (e.g., Binder, Westbury, McKiernan, Possing, & Medler, 2005; Friederici, Opitz, & von Cramon, 2000; Noppeney & Price, 2004). In several studies, abstract words were defined as those with low imageability and concreteness ratings, and vice versa for the concrete words. The overlapping portion of activation between the two word classes consisted of left-lateralized areas that (p. 537) receive inputs from multiple sensory modalities, such as angular gyrus. In one of the studies, concrete concepts, in contrast to abstract concepts, elicited greater activation in right-lateralized multimodal areas (Binder et al., 2005). Abstract concepts produced greater activation primarily in left inferior frontal gyrus, an important area in the language system. Thus, the findings from these univariate analysis-based studies suggest that abstract (versus concrete) concepts evoke other verbal concepts and involve less sensorimotor knowledge than concrete concepts.

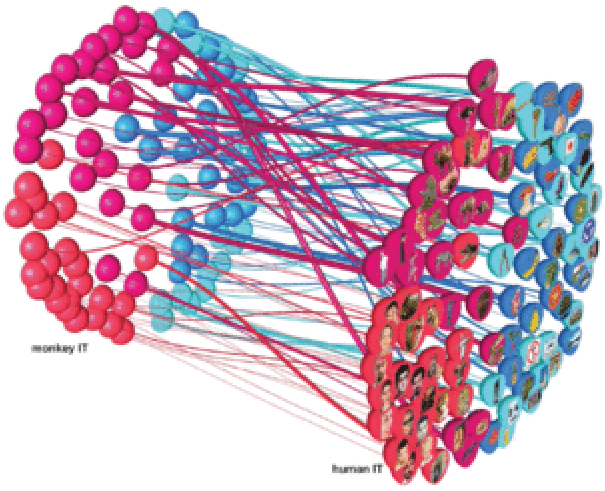

Figure 21.4. Neural representations of concrete objects are similar between monkeys and humans. Arrangements of the same picture stimuli separately for monkeys and humans, such that the distance between any two pictures reflects the dissimilarity between their activity patterns (1 – the spatial correlation) in IT (inferotemporal cortex), a set of high-order visual brain areas.

Monkey data: 674 single-neuron electrophysiological recordings from two macaque monkeys (Kiani et al., 2007). Human data: fMRI activation in 316 voxels (Kriegeskorte, Mur, Ruff, et al., 2008). Different categories: face (red), body (magenta), natural object (blue), artificial object (cyan). The lines connect the same pictures between monkey and human; thick lines indicate that the neural representations were dissimilar between monkey and human.

Source: The figure was adapted from Kriegeskorte (2009) freely under the terms of the Creative Commons Attribution License.

The only multivariate analysis-based study to date that compares the neural representations between abstract and concrete concepts documented findings similar (p. 538) to those of the univariate analysis-based studies (Wang, Baucom, & Shinkareva, 2013). For example, left inferior frontal gyrus was among a small set of regions that by itself enabled a classifier to recognize the activation patterns corresponding to abstract concepts. Taken together, the results in this research area support the dual-coding theory of semantic processing (Paivio, 1986), which postulates that abstract concepts are neurally represented primarily as lexical items, whereas concrete concepts are additionally stored as sensorimotor representations.

The findings of left inferior gyrus involvement in the representation of abstract concepts leave open the question of what type of information is represented there. Apart from its critical role in language processing, this region has been associated with phonological working memory (Burton, 2001), and so activation patterns detected in this region might contain sustained phonological representations of words, while knowledge related to the word meaning is being retrieved from other brain locations. Left inferior frontal gyrus has also been suggested to mediate conflicts in the retrieval of knowledge among competing alternatives (Thompson-Schill, D’Esposito, & Kan, 1999). Thus, activation detected in this region might reflect various instances of mediation among competing requests for the retrieval of knowledge related to the concept currently being thought about. According to either interpretation, activation patterns in this region would not appear to encode the knowledge per se associated with a concept.

The evidence uncovered thus far suggests that an abstract concept evokes a set of verbal or lexical representations associated with that concept, more so than does a concrete concept. This lexical information might also include concrete words, whose meaning is neurally represented in sensorimotor brain areas. Regression models might be used to discover sensorimotor activation patterns associated with abstract concepts by accounting for any hidden concrete factors that underpin the representations.

Scientific concepts are a specific type of abstract concept learned only through formal education. The neural signatures of scientific abstract physics concepts (e.g., *gravity, torque, frequency*) can be decomposed into meaningful underlying neural and semantic dimensions, despite their abstractness. Mason and Just (2016) used factor analysis to uncover the underlying dimensions of the neural representation of 30 physics concepts. The four main dimensions underlying the neural representation of these abstract concepts were *causality, periodicity, algebraic representation* (a sentence-like statement of the quantitative relations among concepts), and *energy flow*, all of which are dimensions that are used for representing familiar concrete concepts. For example, a concept like *frequency* has a strong *periodicity* component. (The brain locations corresponding to this factor included bilateral superior parietal gyrus, left postcentral sulcus, left posterior superior frontal gyrus, and bilateral inferior temporal gyrus.) The applicability of these underlying dimensions was assessed in terms of a classification model that used the factor-related brain locations to accurately classify the 30 abstract concepts based on their neural signature. The findings suggest that abstract scientific concepts are represented by repurposing neural structures that originally evolved for more general purposes. The underlying brain capabilities that form the basis for physics concepts existed long before physics knowledge was developed.

(p. 539) Changes in Neural Concept Representations with Learning

The ability to track the growth of a neural concept representation speaks to one of the foundational goals of cognitive neuroscience research, namely to understand the neural basis of knowledge acquisition. The study of concept learning also promises to enable a greater understanding of how concept knowledge is represented and processed in the brain. However, little is known about the changes that occur in a neural concept representation as a new concept is being learned.

Much of the existing research on concept learning has focused on changes in which brain regions show heightened activation between pre- and post-learning. For example, after a session of learning how to manipulate novel tool-like objects, activation to pictures of the objects was found to shift predominantly to motor cortex compared to pre-learning

(Weisberg, van Turennout, & Martin, 2007). Another study showed that after participants were verbally instructed about the kind of motion or sound that was associated with novel living objects, the activation elicited by the object pictures was localized to motion-specific or auditory cortex (James & Gauthier, 2003). These studies showed that the brain regions that became active after learning corresponded to the kinds of information that were taught. However, the univariate analyses used in these studies did not permit a determination of how each individual new concept became encoded in a distributed neural representation within the new sites of activation.

A multivariate study of concept learning documented the emergence of the neural representations of individual new concepts (Bauer & Just, 2015). Specifically, the growth of the representations of new animal concepts was monitored as two properties of each animal were taught, namely an animal's habitat and its diet or eating habits. The learning of information about each of these dimensions was demonstrated by an increase in the accuracy of classifying the animal identities based on the brain areas associated with the dimension that had been learned. For example, after participants had learned about the habitats of some animals, it was possible to classify which animal they were thinking about by training a classifier on the activation patterns in regions associated with shelter information. This study provides a novel form of causal evidence that newly acquired knowledge comes to reside in the brain regions previously shown to underlie a particular type of concept knowledge.

Another neurosemantic study examined the changes in the neural representations of complex mechanical concepts as they were being learned, and found that different stages of learning are associated with different sets of brain regions that encode the emerging knowledge (Mason & Just, 2015). Specifically, the study demonstrated how incremental instruction about the workings of several mechanical concepts (e.g., bathroom scale, automobile braking system) gradually changed the neural representations of the systems. The representations progressed through different states that reflected different learning stages, starting with the visual properties of (p. 540) the concept encoded from the display, mental animation of mechanical components, generation of causal hypotheses associated with the animation, and determination of how a person would interact with the mechanical system. Research on intermediate stages of learning has lagged behind studies that focus only on final outcomes of learning (Karuza, Emberson, & Aslin, 2014). The results in Mason and Just (2015) raise the possibility that the neural representations of familiar concepts (which is the only type of concept that most previous studies have investigated) may fail to reveal the constructive processes by which the neural representations become established. The constructive processes may reveal some fundamental properties of neural concept representations.

The neurosemantic research on concept learning provides a foundation for brain research to trace how new knowledge makes its way from the words and graphics used to teach it, to a neural concept representation in a learner's brain. It might foreshadow an era in which brain imaging and neurosemantic methods are used to diagnose which aspects of a concept a student misunderstands or lacks, in a way that might be more fundamental and accurate than conventional behavioral testing. An fMRI study in which real-time measurement of brain activation identified mental states that were either “prepared” or “unprepared” for encoding a new stimulus lends credence to this possibility (Yoo et al., 2012).

The study of how the learning process changes neural concept representations promises to enable a greater understanding of the kinds of information encoded in neural representations. Just as neurosemantic methods have been useful in determining where and how the different dimensions of a concept are encoded, these methods might eventually be used to track the developmental trajectory of neural representations as a function of various factors of interest, such as a person’s previous experience or knowledge, or elapsed time between learning episodes. Perhaps a comparison of representations at different stages of knowledge expertise would aid in deciphering the kinds of information that are encoded in the representations. For example, chess experts can remember large configurations of chess pieces on a board by representing various relationships among the chess pieces (Gobet & Simon, 1996). Comparisons between the information that is neurally encoded in a domain expert versus a novice might illuminate the process of building complex neural representations.

Conclusion

Neurosemantic methods have enabled enormous advances in uncovering how various types of concept knowledge are neurally represented, and also in characterizing the information contained in the representations. The ability to study how different concepts are neurally represented has only recently become possible since the development of data analytic methods that can detect a correspondence between a distributed activation pattern and an individual concept. The key virtues of the (p. 541) neurosemantic approach over older methods are that it generally permits greater sensitivity to uncovering the underlying phenomenon, and it adheres to the fundamental principle that concept information is encoded in neural populations distributed throughout the brain. The approach promises to illuminate a number of prominent questions; for example, the field is better equipped to determine whether abstract concepts are neurally encoded as lexical representations, or whether abstract thoughts are underpinned by sensorimotor factors as revealed by organized patterns of activation in these brain regions. The neurosemantic paradigm provides the tools for forging discoveries in areas of daunting complexity, such as how the relations among a concept’s underlying semantic dimensions are neurally encoded and thereby represent a cohesive concept, and how learning establishes and shapes new representations.

Acknowledgments

This work was supported by the National Institute of Mental Health Grant MH029617 and the Office of Naval Research Grant N00014-16-1-2694.

References

Allen, K., Pereira, F., Botvinick, M., & Goldberg, A. E. (2012). Distinguishing grammatical constructions with fMRI pattern analysis. *Brain and Language, 123*(3), 174–182. doi: 10.1016/j.bandl.2012.08.005

Anderson, J. R. (1978). Arguments concerning representations for mental imagery. *Psychological Review, 85*, 249–277. doi: 10.1037/0033- 295X.85.4.249

Anderson, J. R., & Fincham, J. M. (2014). Discovering the sequential structure of thought. *Cognitive Science, 38*(2), 322–352. doi: 10.1111/cogs.12068

Arciuli, J., McMahon, K., & de Zubicaray, G. (2012). Probabilistic orthographic cues to grammatical category in the brain. *Brain and Language, 123*(3), 202–210. doi: 10.1016/j.bandl.2012.09.009

Barsalou, L. W. (1992). Frames, concepts, and conceptual fields. In E. Kittay & A. Lehrer (Eds.), *Frames, fields, and contrasts: New essays in semantic and lexical organization* (pp. 21–74). Hillsdale, NJ: Lawrence Erlbaum Associates.

Barsalou, L. W., Santos, A., Simmons, W. K., & Wilson, C. D. (2008). Language and simulation in conceptual processing. In M. De Vega, A. M. Glenberg, & A. C. Graesser (Eds.), *Symbols, embodiment, and meaning* (pp. 245–283). Oxford: Oxford University Press.

Baucom, L. B., Wedell, D. H., Wang, J., Blitzer, D. N., & Shinkareva, S. V. (2012). Decoding the neural representation of affective states. *NeuroImage, 59*(1), 718–727. doi: 10.1016/j.neuroimage.2011.07.037

Bauer, A. J., & Just, M. A. (2015). Monitoring the growth of the neural representations of new animal concepts. *Human Brain Mapping, 36*(8), 3213–3226. doi: 10.1002/hbm.22842

Beran, M. J., Johnson-Pynn, J. S., & Ready, C. (2011). Comparing children's *Homo sapiens* and chimpanzees' *Pan troglodytes* quantity judgments of sequentially presented sets of items. *Current Zoology*, *57*(4), 419–428.

(p. 542) 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*(12), 2767–2796. doi: 10.1093/cercor/bhp055

Binder, J. R., Westbury, C. F., McKiernan, K., Possing, E. T., & Medler, D. (2005). Distinct brain systems for processing concrete and abstract concepts. *Journal of Cognitive Neuroscience, 17*(6), 905–917.

Buchweitz, A., Shinkareva, S. V, Mason, R., Mitchell, T. M., & Just, M. A. (2012). Identifying bilingual semantic neural representations across languages. *Brain and Language, 120*(3), 282–289. doi: 10.1016/j.bandl.2011.09.003

Burton, M. W. (2001). The role of inferior frontal cortex in phonological processing. *Cognitive Science, 25*, 695–709.

Capitani, E., Laiacona, M., Mahon, B., & Caramazza, A. (2003). What are the facts of semantic category-specific deficits? A critical review of the clinical evidence. *Cognitive Neuropsychology, 20*, 213–261. doi: 10.1080/02643290244000266

Chadwick, M. J., Hassabis, D., Weiskopf, N., Maguire, E. A. (2010). Decoding individual episodic memory traces in the human hippocampus. *Current Biology, 20*, 544–547.

Chang, K. K., Mitchell, T., & Just, M. A. (2011). Quantitative modeling of the neural representation of objects: How semantic feature norms can account for fMRI activation. *NeuroImage, 56*(2), 716–727. doi: 10.1016/j.neuroimage.2010.04.271

Chao, L. L., Weisberg, J., & Martin, A. (2002). Experience-dependent modulation of category-related cortical activity. *Cerebral Cortex, 12*(5), 545–551.

Charest, I., Kievit, R. A., Schmitz, T. W., Deca, D., & Kriegeskorte, N. (2014). Unique semantic space in the brain of each beholder predicts perceived similarity. *PNAS, 111*(40), 14565–14570. doi: 10.1073/pnas.1402594111

Collins, A., & Quillian, M. R. (1972). Experiments on semantic memory and language comprehension. In L. W. Gregg (Ed.), *Cognition in learning and memory* (pp. 117–147). New York: John Wiley & Sons.

Coltheart, M., Patterson, K., & Marshall, J. (1980). *Deep dyslexia*. London: Routledge & Kegan Paul.

Connolly, A. C., Guntupalli, J. S., Gors, J., Hanke, M., Halchenko, Y. O., Wu, Y.-C., Abdi, H., & Haxby, J. V. (2012). The representation of biological classes in the human brain. *The Journal of Neuroscience, 32*(8), 2608–2618. doi: 10.1523/JNEUROSCI.5547-11.2012

Correia, J., Formisano, E., Valente, G., Hausfeld, L., Jansma, B., & Bonte, M. (2014). Brain-based translation: fMRI decoding of spoken words in bilinguals reveals language-independent semantic representations in anterior temporal lobe. *Journal of Neuroscience, 34*(1), 332–338. doi:10.1523/JNEUROSCI.1302-13.2014

Coutanche, M. N. (2013). Distinguishing multi-voxel patterns and mean activation: Why, how, and what does it tell us? *Cognitive, Affective & Behavioral Neuroscience*, *13*(3), 667–73. doi: 10.3758/s13415-013-0186-2

Coutanche, M. N., & Thompson-Schill, S. L. (2014). Creating concepts from converging features in human cortex. *Cerebral Cortex*, *25*, 2584–2593. doi: 10.1093/cercor/bhu057

Cox, D. D., & Savoy, R. L. (2003). Functional magnetic resonance imaging (fMRI) “brain reading”: Detecting and classifying distributed patterns of fMRI activity in human visual cortex. *NeuroImage, 19*(2), 261–270. doi: 10.1016/S1053-8119(03)00049-1

Cree, G. S., & McRae, K. (2003). Analyzing the factors underlying the structure and computation of the meaning of chipmunk, cherry, chisel, cheese, and cello (and many other such (p. 543) concrete nouns). *Journal of Experimental Psychology: General, 132*(2), 163–201. doi: 10.1037/0096-3445.132.2.163

Çukur, T., Nishimoto, S., Huth, A. G., & Gallant, J. L. (2013). Attention during natural vision warps semantic representation across the human brain. *Nature Neuroscience, 16*(6), 763–770. doi: 10.1038/nn.3381

Damarla, S. R., & Just, M. A. (2013). Decoding the representation of numerical values from brain activation patterns. *Human Brain Mapping, 34*(10), 2624–2634. doi: 10.1002/hbm.22087

Damasio, A. R. (1989). Time-locked multiregional retroactivation: A systems-level proposal for the neural substrates of recall and recognition. *Cognition, 33*(1–2), 25–62.

Davis, T., LaRocque, K. F., Mumford, J. A., Norman, K. A., Wagner, A. D., & Poldrack, R. A. (2014). What do differences between multi-voxel and univariate analysis mean? How subject-, voxel-, and trial-level variance impact fMRI analysis. *NeuroImage, 97*, 271–283. doi: 10.1016/j.neuroimage.2014.04.037

de Zubicaray, G., Arciuli, J., & McMahon, K. (2013). Putting an “end” to the motor cortex representations of action words. *Journal of Cognitive Neuroscience*, *25*(11), 1957–1974.

Eger, E., Michel, V., Thirion, B., Amadon, A., Dehaene, S., & Kleinschmidt, A. (2009). Deciphering cortical number coding from human brain activity patterns. *Current Biology, 19*(19), 1608–1615. doi: 10.1016/j.cub.2009.08.047

Epstein, R., & Kanwisher, N. (1998). A cortical representation of the local visual environment. *Nature, 392*(6676), 598–601. doi: 10.1038/33402

Friederici, A. D., Opitz, B., & von Cramon, D. Y. (2000). Segregating semantic and syntactic aspects of processing in the human brain: An fMRI investigation of different word types. *Cerebral Cortex*, *10*(7), 698–705.

Friston, K. J., Holmes, A. P., Worsley, K. J., Poline, J.-P., Frith, C. D., & Frackowiak, R. S. J. (1994). Statistical parametric maps in functional imaging: A general linear approach. *Human Brain Mapping, 2*(4), 189–210.

Gobet, F., & Simon, H. A. (1996). Templates in chess memory: A mechanism for recalling several boards. *Cognitive Psychology, 31*(1), 1–40.

Goldstone, R. L., & Kersten, A. (2003). Concepts and categorization. In A. F. Healy & R. W. Proctor (Eds.), *Comprehensive handbook of psychology*, Vol. 4: *Experimental psychology* (pp. 591–621). New York: John Wiley & Sons.

Hassabis, D., Spreng, R. N., Rusu, A. A., Robbins, C. A., Mar, R. A., & Schacter, D. L. (2013). Imagine all the people: How the brain creates and uses personality models to predict behavior. *Cerebral Cortex, 24*, 1979–1987.

Hauk, O., Johnsrude, I., & Pulvermüller, F. (2004). Somatotopic representation of action words in human motor and premotor cortex. *Neuron, 41*, 301–307.

Haxby, J. V. (2012). Multivariate pattern analysis of fMRI: The early beginnings. *NeuroImage*, *62*(2), 852–855. doi: 10.1016/j.neuroimage.2012.03.016

Haxby, J. V, Connolly, A. C., & Guntupalli, J. S. (2014). Decoding neural representational spaces using multivariate pattern analysis. *Annual Review of Neuroscience, 37*, 435–456. doi: 10.1146/annurev-neuro-062012-170325

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.

Haxby, J. V, Guntupalli, J. S., Connolly, A. C., Halchenko, Y. O., Conroy, B. R., Gobbini, M. I., Hanke, M., & Ramadge, P. J. (2011). A common, high-dimensional model of the (p. 544) representational space in human ventral temporal cortex. *Neuron, 72*(2), 404–416. doi: 10.1016/j.neuron.2011.08.026

Haynes, J.-D., Sakai, K., Rees, G., Gilbert, S., Frith, C., & Passingham, R. E. (2007). Reading hidden intentions in the human brain. *Current Biology, 17*(4), 323–328. doi: 10.1016/j.cub.2006.11.072

Hsu, N. S., Schlichting, M. L., & Thompson-Schill, S. L. (2014). Feature diagnosticity affects representations of novel and familiar objects. *Journal of Cognitive Neuroscience*, *26*, 2735–2749. doi: 10.1162/jocn_a_00661

Huth, A. G., Nishimoto, S., Vu, A. T., and Gallant, J. L. (2012). A continuous semantic space describes the representation of thousands of object and action categories across the human brain. *Neuron, 76*, 1210–1224.

James, T. W., & Gauthier, I. (2003). Auditory and action semantic features activate sensory-specific perceptual brain regions. *Current Biology*, *13*, 1792–1796.

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*(1), e8622. doi: 10.1371/journal.pone.0008622

Just, M. A., Cherkassky, V. L., Buchweitz, A., Keller, T. A., & Mitchell, T. M. (2014). Identifying autism from neural representations of social interactions: Neurocognitive markers of autism. *PLOS One*, *9*(12), e113879. doi: 10.1371/journal.pone.0113879

Karuza, E. A, Emberson, L. L., & Aslin, R. N. (2014). Combining fMRI and behavioral measures to examine the process of human learning. *Neurobiology of Learning and Memory, 109*, 193–206. doi: 10.1016/j.nlm.2013.09.012

Kassam, K. S., Markey, A. R., Cherkassky, V. L., Loewenstein, G., & Just, M. A. (2013). Identifying emotions on the basis of neural activation. *PLoS ONE*, *8*(6), e66032.

Kay, K. N., Naselaris, T., Prenger, R. J., & Gallant, J. L. (2008). Identifying natural images from human brain activity. *Nature, 452*(7185), 352–355. doi: 10.1038/nature06713

Kiani, R., Esteky, H., Mirpour, K., & Tanaka, K. (2007). Object category structure in response patterns of neuronal population in monkey inferior temporal cortex. *Journal of Neurophysiology, 97*(6), 4296–4309. doi: 10.1152/jn.00024.2007

Kiefer, M., & Pulvermüller, F. (2012). Conceptual representations in mind and brain: Theoretical developments, current evidence and future directions. *Cortex, 48*(7), 805–825. doi: 10.1016/j.cortex.2011.04.006

Kiefer, M., Sim, E.-J., Herrnberger, B., Grothe, J., & Hoenig, K. (2008). The sound of concepts: Four markers for a link between auditory and conceptual brain systems. *The Journal of Neuroscience, 28*(47), 12224–12230. doi: 10.1523/JNEUROSCI.3579-08.2008

Kriegeskorte, N. (2009). Relating population-code representations between man, monkey, and computational models. *Frontiers in Neuroscience, 3*(3), 363–373. doi: 10.3389/neuro.01.035.2009

Kriegeskorte, N., Mur, M., & Bandettini, P. (2008). Representational similarity analysis: Connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience, 2*(4). doi: 10.3389/neuro.06.004.200

Kriegeskorte, N., Mur, M., Ruff, D. A., Kiani, R., Bodurka, J., Esteky, H., Tanaka, K., & Bandettini, P. A. (2008). Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron, 60*(6), 1126–1141. doi: 10.1016/j.neuron.2008.10.043

Lewis, J. W. (2006). Cortical networks related to human use of tools. *The Neuroscientist, 12*(3), 211–231. doi: 10.1177/1073858406288327

(p. 545) Mahon, B. Z., & Caramazza, A. (2003). Constraining questions about the organisation and representation of conceptual knowledge. *Cognitive Neuropsychology, 20*(3), 433–450. doi: 10.1080/02643290342000014

Mahon, B. Z., & Caramazza, A. (2008). A critical look at the embodied cognition hypothesis and a new proposal for grounding conceptual content. *Journal of Physiology, Paris, 102*(1–3), 59–70. doi: 10.1016/j.jphysparis.2008.03.004

Martin, A. (2007). The representation of object concepts in the brain. *Annual Review of Psychology, 58*, 25–45. doi: 10.1146/annurev.psych.57.102904.190143

Mason, R. A., & Just, M. A. (2015). Physics instruction induces changes in neural knowledge representation during successive stages of learning. *NeuroImage, 111*, 36–48. doi: 10.1016/j.neuroimage.2014.12.086

Mason, R. A., & Just, M. A. (2016). Neural representations of physics concepts. *Psychological Science, 27*(6), 904–913. doi: 10.1177/0956797616641941

Meteyard, L., Cuadrado, S. R., Bahrami, B., & Vigliocco, G. (2010). Coming of age: A review of embodiment and the neuroscience of semantics. *Cortex, 48*(7), 788–804. doi: 10.1016/j.cortex.2010.11.002

Mitchell, T. M., Hutchinson, R., Just, M. A, Niculescu, R. S., Pereira, F., & Wang, X. (2003). Classifying instantaneous cognitive states from fMRI data. *AMIA Annual Symposium Proceedings,* 465–469.

Mitchell, T., Hutchinson, R., Niculescu, R. S., Pereira, F., Wang, X., Just, M. A., & Newman, S. D. (2004). Learning to decode cognitive states from brain images. *Machine Learning, 57*, 145–175.

Mitchell, T. M., Shinkareva, S. V, Carlson, A., Chang, K.-M., Malave, V. L., Mason, R. A, & Just, M. A. (2008). Predicting human brain activity associated with the meanings of nouns. *Science, 320*(5880), 1191–1195. doi: 10.1126/science.1152876

Moseley, R. L., & Pulvermüller, F. (2014). Nouns, verbs, objects, actions, and abstractions: Local fMRI activity indexes semantics, not lexical categories. *Brain and Language, 132,* 28–42. doi:10.1016/j.bandl.2014.03.001

Mur, M., Bandettini, P. A, & Kriegeskorte, N. (2009). Revealing representational content with pattern-information fMRI: An introductory guide. *Social Cognitive and Affective Neuroscience, 4*(1), 101–109. doi:10.1093/scan/nsn044

Naselaris, T., Kay, K. N., Nishimoto, S., & Gallant, J. L. (2011). Encoding and decoding in fMRI. *NeuroImage, 56*(2), 400–410. doi: 10.1016/j.neuroimage.2010.07.073

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (1998). The University of South Florida word association, rhyme, and word fragment norms. **http://w3.usf.edu/FreeAssociation**

Noppeney, U., & Price, C. J. (2004). Retrieval of abstract semantics. *NeuroImage, 22*(1), 164–170. doi: 10.1016/j.neuroimage.2003.12.010

Norman, K. A, Polyn, S. M., Detre, G. J., & Haxby, J. V. (2006). Beyond mind-reading: Multi-voxel pattern analysis of fMRI data. *Trends in Cognitive Sciences, 10*(9), 424–430. doi: 10.1016/j.tics.2006.07.005

O’Toole, A. J., Jiang, F., Abdi, H., Pénard, N., Dunlop, J. P., & Parent, M. A. (2007). Theoretical, statistical, and practical perspectives on pattern-based classification approaches to the analysis of functional neuroimaging data. *Journal of Cognitive Neuroscience, 19*(11), 1735–1752. doi: 10.1162/jocn.2007.19.11.1735

Paivio, A. (1986). *Mental representations: A dual-coding approach*. New York: Oxford University Press.

(p. 546) 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*(12), 976–987. doi: 10.1038/nrn2277

Pereira, F., Botvinick, M., & Detre, G. (2013). Using Wikipedia to learn semantic feature representations of concrete concepts in neuroimaging experiments. *Artificial Intelligence, 194*, 240–252. doi:10.1016/j.artint.2012.06.005

Pereira, F., Mitchell, T., & Botvinick, M. (2009). Machine learning classifiers and fMRI: A tutorial overview. *NeuroImage, 45*(1 Suppl), S199–S209. doi: 10.1016/j.neuroimage.2008.11.007

Princeton University. (2010). “About WordNet.” **http://wordnet.princeton.edu**.

Pobric, G., Jefferies, E., & Ralph, M. A. L. (2007). Anterior temporal lobes mediate semantic representation: Mimicking semantic dementia by using rTMS in normal participants. *PNAS, 104*(50), 20137–20141. doi: 10.1073/pnas.0707383104

Poeppel, D. (2012). The maps problem and the mapping problem: Two challenges for a cognitive neuroscience of speech and language. *Cognitive Neuropsychology, 29*(1–2), 34–55. doi: 10.1080/02643294.2012.710600

Poldrack, R. A. (2007). Region of interest analysis for fMRI. *Social Cognitive and Affective Neuroscience, 2*(1), 67–70. doi: 10.1093/scan/nsm006

Raizada, R. D. S., & Connolly, A. C. (2012). What makes different people’s representations alike: Neural similarity space solves the problem of across-subject fMRI decoding. *Journal of Cognitive Neuroscience, 24*(4), 868–877. doi: 10.1162/jocn_a_00189

Sadtler, P. T., Quick, K. M., Golub, M. D., Chase, S. M., Ryu, S. I., Tyler-Kabara, E. C., Yu, B. M., & Batista, A. P. (2014). Neural constraints on learning. *Nature, 512*(7515), 423–426. doi: 10.1038/nature13665

Schwanenflugel, P. J., & Harnishfeger, K. K. (1988). Context availability and lexical decisions for abstract and concrete words. *Journal of Memory and Language, 27*(5), 499–520.

Seymour, K., Clifford, C. W. G., Logothetis, N. K., & Bartels, A. (2009). The coding of color, motion, and their conjunction in the human visual cortex. *Current Biology, 19*(3), 177–183. doi: 10.1016/j.cub.2008.12.050

Shapiro, K. A., Mottaghy, F. M., Schiller, N. O., Poeppel, T. D., Flüß, M. O., Müller, H.-W., . . . Caramazza, A., & Krause, B. J. (2005). Dissociating neural correlates for nouns and verbs. *NeuroImage, 24*(4), 1058–1067. doi: 10.1016/j.neuroimage.2004.10.015

Shinkareva, S. V, Malave, V. L., Mason, R., Mitchell, T. M., & Just, M. A. (2011). Commonality of neural representations of words and pictures. *NeuroImage, 54*(3), 2418–2425. doi: 10.1016/j.neuroimage.2010.10.042

Simmons, W. K., Hamann, S. B., Harenski, C. L., Hu, X. P., & Barsalou, L. W. (2008). fMRI evidence for word association and situated simulation in conceptual processing. *Journal of Physiology, Paris, 102*(1–3), 106–119. doi: 10.1016/j.jphysparis.2008.03.014

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*(5), 1565–1572. doi: 10.1523/JNEUROSCI.4657-08.2009

Sudre, G., Pomerleau, D., Palatucci, M., Wehbe, L., Fyshe, A., Salmelin, R., & Mitchell, T. (2012). Tracking neural coding of perceptual and semantic features of concrete nouns. *NeuroImage, 62*(1), 451–463. doi: 10.1016/j.neuroimage.2012.04.048

Takerkart, S., Auzias, G., Thirion, B., & Ralaivola, L. (2014). Graph-based inter-subject pattern analysis of fMRI data. *PLOS One*, *9*(8), e104586. doi: 10.1371/journal.pone.0104586

Thompson-Schill, S. L., D’Esposito, M., & Kan, I. P. (1999). Effects of repetition and competition on activity in left prefrontal cortex during word generation. *Neuron, 23*(3), 513–522.

(p. 547) Vann, S. D., Aggleton, J. P., & Maguire, E. A. (2009). What does the retrosplenial cortex do? *Nature Reviews Neuroscience*, *10*(11), 792–802. doi: 10.1038/nrn2733

Vigliocco, G., Vinson, D. P., Druks, J., Barber, H., & Cappa, S. F. (2011). Nouns and verbs in the brain: A review of behavioural, electrophysiological, neuropsychological and imaging studies. *Neuroscience and Biobehavioral Reviews*, *35*(3), 407–426. doi: 10.1016/j.neubiorev.2010.04.007

Walther, A., Nili, H., Ejaz, N., Alink, A., Kriegeskorte, N., & Diedrichsen, J. (2016). Reliability of dissimilarity measures for multi-voxel pattern analysis. *NeuroImage, 137*, 188–200. doi: 10.1016/j.neuroimage.2015.12.012

Wang, J., Baucom, L. B., & Shinkareva, S. V. (2013). Decoding abstract and concrete concept representations based on single-trial fMRI data. *Human Brain Mapping*, *34*(5), 1133–1147. doi: 10.1002/hbm.21498

Wehbe, L., Murphy, B., Talukdar, P., Fyshe, A., Ramdas, A., & Mitchell, T. (2014). Simultaneously uncovering the patterns of brain regions involved in different story reading subprocesses. *PLOS One, 9*(11), e112575. doi: 10.1371/journal.pone.0112575

Weisberg, J., van Turennout, M., & Martin, A. (2007). A neural system for learning about object function. *Cerebral Cortex, 17*(3), 513–521. doi: 10.1093/cercor/bhj176

Yoo, J. J., Hinds, O., Ofen, N., Thompson, T. W., Whitfield-Gabrieli, S., Triantafyllou, C., & Gabrieli, J. D. E. (2012). When the brain is prepared to learn: Enhancing human learning using real-time fMRI. *NeuroImage, 59*(1), 846–852. doi: 10.1016/j.neuroimage.2011.07.063

Andrew J. Bauer

Andrew J. Bauer received his PhD at Carnegie Mellon University and is currently a Postdoctoral Fellow at the University of Toronto. His research uses machine learning techniques applied to fMRI data to understand where and how knowledge is neurally represented in the brain, and how the brain changes with learning new concepts.

Marcel A. Just

Marcel A. Just, D. O. Hebb Professor of Cognitive Neuroscience at Carnegie Mellon and Director of its Center for Cognitive Brain Imaging, uses fMRI to study language-related neural processing. The research uses machine learning and other techniques to identify the semantic components of the neural signature of individual concepts, such as concrete objects (e.g., hammer), emotions (e.g., sadness), and quantities (e.g., three). The projects examine normal concept representations in college students, as well as disordered concepts in special populations, such as patients with autism or suicidal ideation.