How Words Capture Visual Experience: The Perspective From Cognitive Neuroscience

David Kemmerer
Purdue University


Would it be better to sit in silence?
To think everything, to feel everything, to say nothing?
This is the way of the orange gourd.
This is the habit of the rock in the river, over which the water pours all night and all day.
But the nature of man is not the nature of silence.
Words are the thunders of the mind.
Words are the refinements of the flesh.
Words are the responses to the thousand curvaceous moments—we just manage it—sweet and electric, words flow from the brain and out the gate of the mouth.
We make books of them, out of hesitations and grammar.
We are slow, and choosy.
This is the world.

Mary Oliver (2001)

1. INTRODUCTION

The other chapters in this volume adopt the perspectives of linguistics, psychology, and anthropology to approach the central theme of how words capture experience. Here I adopt yet another perspective—that of cognitive neuroscience—and attempt to demonstrate how recent discoveries in this rapidly growing field can help illuminate the nature of word meaning. So far the greatest progress has been made by investigations of the brain structures that underlie the linguistic encoding of the visual world, and for this reason I concentrate on this complex and intriguing set of findings. This area of inquiry is not without controversy (for reviews of several competing theoretical positions, see Caramazza & Mahon, 2006; Gainotti, 2006; Hart & Kraut, 2007b). However, there is widespread agreement about the importance of a variety of studies which suggest that different visual components of word meaning—for instance, the range of colors encoded by brown, the type of shape encoded by horse, the motion pattern encoded by gallop, and the network of spatial relations encoded by around—depend on cortical regions that either overlap with or lie adjacent to some of the same regions that are engaged during the visual perception of those properties. It is this body of data that is the main focus of the chapter. I argue that these new findings, which come from both functional neuroimaging studies with normal subjects and neuropsychological studies with brain-damaged patients, support an approach to linguistic meaning, and to conceptual knowledge more generally, that is often referred to as the Embodied Cognition Framework or, as I will call it, the Simulation Framework.
This approach treats semantic structures as being grounded in modality-specific sensorimotor systems, as opposed to being completely amodal in character.

The chapter is organized as follows. Section 2 presents the basic tenets of the Simulation Framework and summarizes three recent behavioral studies that provide evidence for it. Section 3 then reviews research in cognitive neuroscience that further elaborates and substantiates the Simulation Framework. The section begins by describing a neurobiologically based instantiation of the Simulation Framework called Convergence Zone Theory. It then surveys empirical findings from a variety of studies that address the neural correlates of the visual semantic components of words for objects, events, and spatial relations. Next, Section 4 covers some recent typological work on crosslinguistic variation in the lexicalization of the visual world, and points out several questions that this work poses for cognitive neuroscience. The major topics are as follows: shape classifiers and the count/mass distinction; language-specific semantic maps across the universal conceptual domain of manner-of-motion; and topological vs. projective spatial relations, the latter category being defined in terms of three different frames of reference—intrinsic, relative, and absolute. Finally, Section 5 concludes the chapter by highlighting some important general aspects of how the Simulation Framework illuminates the nature of word meaning, while also acknowledging some equally important limitations of this approach.

2. THE SIMULATION FRAMEWORK

Many traditional approaches to the human conceptual system assume that semantic knowledge is represented separately from modality-specific systems for perception, action, and introspection (e.g., Fodor, 1975; Pylyshyn, 1984; Barsalou & Hale, 1993; Kintsch, 1998). According to this classic view, sensorimotor representations are transduced into amodal structures such as feature lists, semantic networks, frames, etc., and cognitive processes operate on these structures, not on memories of the original sensorimotor states. Moreover, the content of all types of concepts, including those encoded by words, is believed to consist entirely of combinations of these abstract symbols.

A very different line of thinking is currently being pursued by a growing number of researchers in linguistics (e.g., Hampe, 2005; Evans & Green, 2006; Bergen, 2007), philosophy (e.g., Prinz, 2005; Gallagher, 2006; Johnson, 2007), psychology (e.g., Barsalou et al., 2003; Zwaan, 2004; Gibbs, 2006; Barsalou, 2008; Glenberg, in press), and neuroscience (e.g., Gallese & Lakoff, 2005; Jeannerod, 2006; Keysers & Gazzola, 2006; Thompson-Schill et al., 2006; Martin, 2007), all of whom endorse one form or another of what is often called the Simulation Framework. The central claim of this approach is that semantic knowledge is not purely amodal but is instead grounded in modality-specific input/output systems. Rather than being transduced into amodal symbols, complex unimodal (e.g., visual) feature patterns that recur across different presentations of the same category of stimuli are captured by conjunctive units in correspondingly unimodal memory systems, and correlations between feature patterns across different modalities (e.g., visual and auditory) are captured by higher-order conjunctive units in more integrative crossmodal memory systems. Conceptual tasks, such as processing word meanings, are assumed to involve partial re-enactments of the sensorimotor states that occurred
when the referents were directly experienced. According to the Simulation Framework, these recapitulations are modality-specific in format; however, because they are driven in top-down rather than bottom-up fashion, they are prone to errors and biases, are rarely if ever represented as complete images, and are not necessarily conscious.\(^1\) Evidence for this theory comes from an increasing number of behavioral and neuroscientific studies. Before delving into the neuroscientific data, I will attempt to convey a sense of what the behavioral research is like by summarizing three studies—all conducted by Rolf Zwaan and his colleagues—that provide evidence for linguistically induced, egocentrically anchored, perceptual simulations of shape, motion, and part-whole spatial relations (for reviews of other behavioral work see Barsalou, 2003, 2005, 2008; Pecher & Zwaan, 2005).

First, Zwaan, Stanfield, and Yaxley (2002) tested the prediction that during language comprehension people mentally represent the shapes of concrete objects denoted by count nouns. Subjects read simple sentences describing objects in certain locations, and after each item they indicated whether an object shown in a line drawing had been mentioned in the sentence. Reaction times were faster when the shape of the object in the line drawing matched the implied shape of the same object in the sentence. For example, after reading *The ranger saw the eagle in the sky*, subjects made faster responses to a picture of an eagle with outstretched wings than to a picture of an eagle with folded wings, but after reading *The ranger saw the eagle in the nest*, subjects’ reaction times to the two pictures were reversed. A follow-up experiment obtained similar results when subjects simply named the depicted objects after reading the sentences. These findings are consistent with the claim that when people activate the meanings of count nouns during online sentence processing, they use contextual information to guide a mental reconstruction of the most likely shape that the designated object would have in that situation—a reconstruction that may facilitate visual processing of a subsequently presented picture (for an extension of this research, see Yaxley & Zwaan, 2007).

Second, Zwaan, Madden, Yaxley, and Aveyard (2004; see also Kaschak et al., 2005) investigated whether dynamic perceptual simulations of motion are triggered during language comprehension. Subjects listened to sentences describing movement either toward or away from the observer—e.g., *The shortstop hurled the softball at you* or *You hurled the softball at the shortstop*. After each sentence, subjects were shown two pictures of objects, each presented for 500 ms and separated by 175 ms. On the critical trials, both pictures showed the kind of ball mentioned in the sentence, but the two images differed slightly in size, so that a smaller ball followed by a bigger one suggested movement toward the participant, and a bigger ball followed by a smaller one suggested movement away from the participant. The task was to indicate whether the two pictures showed the same object, and as predicted by the Simulation Framework, subjects were quicker when the motion trajectory suggested by the visual stimuli corresponded to the one implied by the preceding sentence. Given that the subjects were not told to relate the pictures to the sentences, and that the picture comparison task could strictly speaking be performed independently of the sentences, it is remarkable that the sentences nevertheless significantly influenced the subjects’ responses. The authors interpret their results as supporting the view that linguistic descriptions of motion rapidly and perhaps automatically engender dynamic perceptual representations. They also point out that their study dovetails nicely with a

\(^{1}\) For an introduction to this line of thinking in the context of a textbook, see Smith and Kosslyn (2007, chapter 4).
famous experiment by Loftus and Palmer (1974) in which subjects watched short traffic safety films of car accidents and then answered questions about how fast the cars were going when they “contacted,” “bumped into,” “collided with,” “smashed into,” or “hit” each other. Subjects’ judgments were affected by the speed expressed by the verb in question, and this finding has been used to argue that eyewitness testimony is highly susceptible to distortion by post-event questions. Zwaan et al. (2004, p. 618) offer the following account: “The verbs used to probe the participants’ memory for the target event were cues to start dynamic perceptual simulations. Verbs associated with greater speed will produce faster perceptual simulations (i.e., more perceptual change per time unit) than verbs associated with lower speeds. These simulations, rather than the initial memories, were then used to estimate the speed of the vehicle, suggesting that words can, indeed, move mental representations.”

Third, Zwaan and Yaxley (2003) examined whether spatial iconicity affects semantic relatedness judgments. Subjects read pairs of nouns that were presented vertically, one above the other, and indicated whether the members of each pair were semantically related. Some of the word pairs consisted of nouns denoting objects, or parts of objects, that typically have a vertical spatial arrangement—e.g., branch-root, attic-basement, nose-mouth, flame-candle. As expected, subjects’ reaction times were faster when the spatial layout of the words iconically reflected the spatial layout of the designated objects (e.g., branch above root) than when the lexical and referential spatial layouts did not match (e.g., root above branch). The authors suggest that performance was influenced by spatial iconicity in the vertical condition because the two words in each pair activated their corresponding perceptual representations, which included the larger referential unit together with attentional focus on the relative locations of the parts in question—e.g., a canonically oriented tree with focus on the branches in the upper region and the roots in the lower region. When the spatial layout of the words was consistent with the spatial layout of the perceptual simulation, semantic relatedness was easier to judge.

These three studies exemplify the kind of purely behavioral data that has been used to support the Simulation Framework. In the next section I turn to theoretical and empirical work in cognitive neuroscience that provides further elaboration and substantiation of this approach to semantic knowledge. As I will demonstrate, it is now possible to estimate with a fairly high degree of confidence some of the specific brain structures that are engaged when people perform the different types of linguistically triggered visual simulations that Zwaan and his colleagues have investigated.

3. CONTRIBUTIONS FROM COGNITIVE NEUROSCIENCE

3.1. Convergence Zone Theory and the Similarity in Topography Principle

As already mentioned, several competing theories are currently available regarding the organization, representation, and processing of semantic knowledge in the brain. One model that bears many important similarities to the Simulation Framework is Convergence Zone Theory (Damasio, 1989a,b,c; Damasio & Damasio, 1994; Tranel et al., 1997a,b, 2003; Damasio et al., 2004; see also Mesulam, 1998), an extended version of which is Conceptual Topography Theory (Simmons & Barsalou, 2003). This approach assumes that when the exemplars of a conceptual
category are experienced, they are initially represented at the cortical level as fluctuating patterns of activity across multiple modality-specific feature maps in early sensory areas. For example, watching a dog run across a field generates time-locked activation patterns in anatomically distributed regions of the visual system that contain feature maps dedicated to coding properties such as color, shape, and motion (Wandell et al., 2007). Likewise, hearing a dog bark elicits activity in auditory feature maps, and petting a dog triggers activity in somatosensory feature maps. When a pattern of activity arises in a feature map, it is captured by ensembles of “conjunctive neurons” that reside in higher-order microscopic cortical areas referred to as “convergence zones” (CZs). The nature of CZs is discussed in depth in the references cited above, and only a few key points need to be elaborated here to provide a theoretical foundation for the review of empirical findings presented in section 3.2.

First, CZs exist at many hierarchical levels (Simmons & Barsalou, 2003). Within the visual system, the initial level consists of property-specific CZs that receive input from certain low-level feature maps (e.g., those for shape) and that contain conjunctive neurons for detecting complex patterns of activation within those maps (e.g., the distinctive shape properties of dogs). This is the level that I focus on in section 3.2, where I describe in detail various visual semantic components of words for objects, events, and spatial relations. At the next level of coding, but still within the visual system, there are modality-specific CZs housing conjunctive neurons that register correlations among different visual properties of objects (e.g., the tendency for dogs to have both certain shapes and certain movements). Mounting evidence suggests that these CZs reside in the anterior temporal lobes (e.g., Quiroga et al., 2005; Rogers et al., 2006, 2007; Bright et al., 2007; Noppeney et al., 2007; see also Lin et al., 2007). Finally, at the apex of the processing hierarchy there are crossmodal CZs comprised of conjunctive neurons that create complete category “files” by detecting correspondences across different modalities (e.g., visual as well as auditory, tactile, and olfactory information about dogs). It is important to note, however, that, as pointed out by Simmons and Barsalou (2003, p. 465), “crossmodal CZs do not literally represent this conceptual content, as is typically the case in current theories that employ knowledge stores. Instead conjunctive neurons in these CZs point to a hierarchy of lower-order conjunctive neurons that eventually activate features in feature maps.” Anatomically, crossmodal CZs appear to depend on the anteromedial temporal lobes, especially the perirhinal cortex (e.g., Murray & Bussey, 1999; Murray & Richmond, 2001; Taylor et al., 2006).²

Second, as alluded to in the passage quoted above from Simmons and Barsalou (2003), CZs are reciprocally connected with each other and with feature maps via extensive feedforward and feedback pathways, thereby enabling both bottom-up recognition and top-down recall. This constitutes one of the key mechanisms that relates CZ theory to the Simulation Framework, as indicated by Simmons and Barsalou (2003, p. 461) in their detailed discussion of the concept of a wheel: “On perceiving a car, the edges of a wheel are represented in visual feature maps. If selective attention focuses on this region of the perceived car, conjunctive neurons in [a property-specific] CZ capture the features in this region. Later, on reactivating these conjunctive neurons [in a top-down manner], the visual representation of this particular wheel is partially re-enacted. As the perceived wheels of subsequent cars similarly receive attentional processing,

² A qualifier: Independently of long-term memory, during passive perception multisensory integration may occur at much earlier stages of processing, as suggested by Beauchamp (2005) and Ghazanfar and Schroeder (2006).
they activate overlapping conjunctive neurons in the CZ, thereby linking the visual features across different wheels to each other. The result is … a \textit{property simulator}, namely, a system that can simulate the various forms a property takes in different categories. As the simulator for wheel develops, it produces simulations of different wheels in different objects, such as cars, bicycles, and skates.\textsuperscript{3}

Third, for all types of CZs, conjunctive neurons are anatomically distributed according to the Similarity in Topography principle: “The spatial proximity of two neurons in a CZ reflects the similarity of the features they conjoin. As two sets of conjoined features become more similar, the conjunctive neurons that link them lie closer together in the CZ’s spatial topography” (Simmons & Barsalou, 2003, p. 457). Support for this principle comes not only from single-cell recording studies with non-human species (e.g., Tanaka, 1996; Graziano & Aflalo, 2007b; Kiani et al., 2007), but also from functional neuroimaging studies with humans, as discussed below.

Before reviewing a number of experimental studies that address the neural bases of different visual semantic components of words, it is worthwhile to quote at some length an example used by Damasio (1989c) to illustrate his seminal views about how CZ theory accounts for conceptual knowledge, including word meaning: “The presentation of a line drawing of a violin, or presentation of the word ‘violin’ (aurally or orthographically), generates a set of time-locked activations of sensory and motor representations. The activations are generally pertinent to manipulable man-made objects, more specifically pertinent to musical instruments of the string variety, and even more narrowly so to the class of violins. In the visual realm the perceiver is likely to evoke representations of shape, motion, color, and texture which will vary from individual to individual according to the personal experience with violins that each has enjoyed. For those who have held violins in their own hands, or even played a violin, numerous somatosensory representations will also be evoked related to tactile impressions of wood and strings, or relative to the pressure the instrument will have exerted in the perceiver’s body. But that is hardly all. Auditory representations of the range of sounds produced by the instrument may also be generated; motor programs according to which the appropriate posture and motions applicable to a violin can be organized may also be evoked and readied for appropriate display; finally, a range of somatic states appropriate to one’s experience with violins, e.g., like or dislike, pleasurable or painful sensation, and so on, will also be activated. In short, a wide array of representations will be generated that together define the meaning of the entity, momentarily…. The mechanism that permits co-activation of representations depends on devices I have called \textit{convergence zones}, which are ensembles of neurons that ‘know about’ the simultaneous occurrence of patterns of activity during the perceived or recalled experience of entities and events. The probability of simultaneous activation of representations prompted by a stimulus thus depends on the operation of convergence zones which, so to speak, embody a binding code for those representations….”

\textsuperscript{3} For useful figures that illustrate this theory, see Smith and Kosslyn (2007, pp. 166-8).
3.2. **Empirical Findings**

Having outlined the basic architecture of CZ theory, I now show how this model, which can be regarded as a variant of the Simulation Framework, provides a useful context for understanding a number of recent studies that have begun to address, from the perspective of cognitive neuroscience, the intriguing question of how words capture visual experience. The set of studies reviewed below focus on the neural underpinnings of several different visual semantic components of words for objects, events, and spatial relations. A major finding that is supported by these studies, and that I highlight throughout the review, is that, as predicted by the Simulation Framework, specific types of lexically encoded visual properties appear to depend on cortical areas that are very closely related to those that subserve the perception of the very same properties.

3.2.1. **Words for Objects**

Count nouns that designate different types of bounded concrete entities have complex meanings that include information about the appearance of objects in the given category. Here I focus on the neural correlates of three visual semantic components of such words: color properties, shape properties, and motion properties. During our ordinary observation of the world, these three attributes of objects are tightly bound together in unified conscious images (e.g., Edelman & Tononi, 2000; Koch, 2004). Thus, if one watches a black crow fly over a corn field, the color, shape, and motion of the bird are experienced as being perfectly fused so that, for instance, the black color does not “spill out” beyond the edges of the flapping wings. In the brain, however, these different perceptual properties are known to be represented in anatomically separate processing streams of the visual system, and, crucially, recent research suggests that the semantic representations of the corresponding properties parallel this organization. The following brief review of this research gives special emphasis to the influential work of Alex Martin and his colleagues (for a more detailed summary see Martin, 2007).

3.2.1.a. **Color Properties**

Many kinds of objects have typical or “canonical” colors. This applies to numerous categories of artifacts whose colors are determined by social convention (e.g., yellow bulldozers) and even more strongly to various categories of animals (e.g., white swans) and plant life (e.g., green lettuce) whose colors are genetically programmed. Such object-color associations constitute an important part of a person’s semantic knowledge of the relevant nouns.

Turning to the brain, substantial evidence suggests that color constancy—i.e., the ability to see an object as having a stable color under different lighting conditions—is mediated by area V4, which occupies a portion of the lingual gyrus on the ventral surface of the occipital lobe (Zeki & Bartels, 1999). This area is activated during tasks requiring color constancy (Corbetta et al., 1990; Zeki et al., 1991; Sakai et al., 1995), and damage to it causes achromatopsia, that is, acquired color blindness (Zeki, 1990; Bouvier & Engel, 2006; see also Sacks, 1995). Thus, it is generally assumed that V4 plays a crucial role in conscious, passive color sensation (for a splendid, albeit now somewhat dated, discussion, see Zeki, 1993).
To identify the cortical regions that underlie semantic knowledge about the canonical color properties of objects, Martin, Haxby, Lalonde, Wiggs, and Ungerleider (1995) conducted two experiments using positron emission tomography (PET). In both experiments the subjects’ task was to produce words denoting the typical colors of objects, but in the first experiment the stimuli were black-and-white line drawings of objects (e.g., a drawing of a child’s wagon, to elicit red), and in the second they were printed words for objects (e.g., wagon, again to elicit red). Relative to a baseline task in which the subjects simply named each object/word that was presented, both of the experimental tasks elicited bilateral activation in an area of the fusiform gyrus just anterior to V4. This area may function as a property-specific CZ that is reciprocally connected with V4 and that houses distinct populations of conjunctive neurons for distinct color categories. Further evidence for this view comes from several additional studies. For example, the fusiform color area is engaged when subjects name the colors of chromatic line drawings of objects (Chao & Martin, 1999), when subjects imagine named colors (Howard et al., 1998), and when subjects with synaesthesia report automatically experiencing certain colors upon hearing certain words (Paulesu et al., 1995). (See Figure 1). Most impressive of all, however, is a recent functional magnetic resonance imaging (fMRI) study by Simmons, Ramjee, Beauchamp, McRae, Martin, and Barsalou (2007) which predicted and confirmed partially overlapping patterns of activation in the fusiform color area during the following two conditions: first, a conceptual property verification task in which subjects judged whether the type of color denoted by an adjective (e.g., green) generally applies to the type of object denoted by a noun (e.g., lettuce); and second, the Farnsworth-Munsell 100 Hue Test, which is a perceptual task that requires not just color constancy, but also close attention to the distinctions between color categories (see also Beauchamp et al., 1999). According to Simmons et al., (2007, p. 2808), these findings indicate that, in keeping with the Simulation Framework, “retrieving property knowledge shares [part of] the neural substrate underlying property perception.”

On the other hand, an opponent of the Simulation Framework—that is, someone who is skeptical of the view that conceptual knowledge is grounded in the brain’s modality-specific systems—might argue that the fusiform activations that Simmons et al. (2007) observed during the property verification task could reflect non-semantic visual imagery rather than intrinsically semantic content. Simmons et al. (2007, pp. 2807-8) address this criticism as follows: “The neural bases of color perception and explicit color imagery share many commonalities, and were we to ask subjects to explicitly imagine colors, we would expect to see areas of activation similar to those reported here. This expectation follows naturally from a theoretical stance that perceptual information is used in many areas of cognition, including imagery and memory…. Importantly, at no point were subjects instructed to use imagery to perform the property verification task, nor from the amodal perspective should imagery even be necessary. In fact, most amodal accounts would posit that property information is stored propositionally with the relevant concept information. This being the case, it would seem extremely odd for a proponent of amodal accounts to argue that the task cannot be performed using the amodal representations central to amodal theories, but instead must be performed using additional, ancillary, effortful processes…. If property information is amodal, then why would one need to ‘imagine’ the property? Yet subjects do activate modality-specific cortex when accessing property knowledge.”
Moreover, returning to the topic of object-color associations, further evidence for the Simulation Framework comes from a recent behavioral study in which subjects were asked to adjust, via interactive manipulation of a graphic display, the color of natural fruits until they appeared gray. Remarkably, the stimuli were not perceived to be gray until their color was shifted away from the observers’ independently identified gray point in the direction opposite the canonical color of the fruit (e.g., slightly toward blue for a banana). This discovery suggests that “color sensations are not determined by the incoming sensory data alone, but are significantly modulated by high-level visual memory” (Hansen et al., 2006, p. 1367). Indeed, this phenomenon might be generated in part by top-down projections from the fusiform color area to V4.

Finally, it is noteworthy that several recent studies point to an intriguing left-hemisphere-mediated interaction between linguistic and perceptual representations of color. Gilbert, Regier, Kay, and Ivry (2006) report an experiment in which subjects focused their visual attention on a central fixation marker and were then shown a ring of 12 colored squares. One of the squares (the target) was a different hue than the others, and the task was to indicate as quickly as possible whether it was in the left or right visual field. Responses were significantly faster when the target and distractor colors had different names (e.g., a blue target among green distractors) compared to when they had the same name (e.g., a light green target among somewhat darker green distractors). However, this category effect only occurred when the targets were presented in the right visual field. Furthermore, a follow-up experiment showed that the effect disappeared when subjects performed a secondary task requiring verbal working memory, but not when they performed a secondary task requiring spatial working memory. Because information in the right visual field is routed first to the language-dominant left hemisphere, whereas information in the left visual field is routed first to the right hemisphere and must then cross the corpus callosum to access lexical codes, the results can be interpreted as providing evidence for a lexical influence on rapid color discrimination. Gilbert et al. (2006, p. 493) summarize the implications of their study as follows: “Previous studies addressing the possible influence of language on perception have tended to look for a simple yes or no answer to the question. Our findings suggest a more complex picture, based on the functional organization of the brain. The [left hemisphere] appears to sharpen visual distinctions between lexically defined categories and to blur visual distinctions within these categories, whereas the right hemisphere does so much less, if at all. To the degree that these results can be generalized to everyday perception, our representation of the visual world may be, at once and the same time, filtered and not filtered through the categories of language.” Similar findings have been obtained in two subsequent studies (Drivonikou et al., 2007; Roberson et al., in press), and future research will undoubtedly delineate the nature of this fascinating phenomenon in even greater detail, both behaviorally and neurobiologically.

3.2.1.b. Shape Properties

Perhaps the most critical visual semantic component of count nouns is shape. This point is expressed quite forcefully in the following passage from Landau and Jackendoff (1993, p. 218): “In the average [English-speaking] adult vocabulary, there are roughly 10,000 names for things—count nouns that label different kinds of objects. For a large proportion of object categories, shape is among the most important criteria for identification, and in particular for
judgments of what a thing should be called: Categories of things with the same shape, including natural kind objects and artifacts, often share the same name.”

Based on many neuropsychological and functional neuroimaging studies, it is now well-established that within the visual system the shape properties of objects are represented in the ventral occipito-temporal processing stream (e.g., Farah, 2004; Haxby et al., 2004; Milner & Goodale, 2006; Vandenbulcke et al., 2006). Currently, one of the most intensively investigated questions is the following: Across the large expanse of ventral temporal cortex, are CZs for the shape properties of different categories of objects uniformly distributed, or are they clustered together in patches (Reddy & Kanwisher, 2006)? While there is some evidence for distributed coding, there is even greater evidence—consistent with the Similarity in Topography principle—that certain areas are preferentially responsive to certain categories of objects, particularly faces (e.g., Kanwisher & Yovel, 2006; Tsao et al., 2006), non-facial body parts (e.g., Peelen & Downing, 2007; see also Kemmerer & Tranel, 2008), animals (e.g., Chao et al., 1999, 2002), tools (e.g., Chao et al., 1999, 2002), scenes (e.g., Epsten & Kanwisher, 1998; Epstein et al., 2001), and printed words (e.g., McCandliss et al., 2003; Gaillard et al., 2006).

Here I focus on several functional neuroimaging studies that point to segregated cortical representations of the shapes of animals and tools. In one set of fMRI experiments, Chao, Haxby, and Martin (1999) evaluated perceptual processing of animals and tools by using passive viewing tasks and delayed match-to-sample tasks; in addition, they evaluated conceptual processing of animals and tools by using silent picture naming tasks and property verification tasks (i.e., answering yes/no questions like “Forest animal?” and “Kitchen tool?” in response to printed words for animals and tools). Across all of the tasks, significantly greater bilateral activation for animals was consistently found in a lateral portion of the fusiform gyrus, whereas significantly greater bilateral activation for tools was consistently found in a medial portion of the fusiform gyrus. (See Figure 2, left panel). As the authors point out, it is especially interesting that these adjacent but nevertheless distinct regions of the fusiform gyrus were activated not only by pictures but also by words.

The same question arises here, however, as arose for Simmons et al.’s (2007) investigation of object-color associations. In particular, one might suppose that the activations evoked by words were merely a reflection of the subjects' deliberate efforts to conjure up explicit visual images of the shapes of the lexically encoded animals and tools. Evidence against this interpretation, and in favor of the hypothesis that the lexically-driven category-related fusiform activations index automatic semantic processing, comes from a recent fMRI study by Wheatley, Weisberg, Beauchamp, and Martin (2005). This study took advantage of the neurophysiological phenomenon known as “repetition suppression.” Basically, repetition suppression is a decrease in neural response associated with repeated presentation of identical, or conceptually related, stimuli. It reflects the well-established fact that prior exposure to a stimulus leads to greater processing efficiency (for a review see Grill-Spector et al., 2006). In

---

4 In the studies discussed below, tools are usually restricted to small man-made objects that are manipulated in specific ways to serve specific functions.
Wheatley et al.’s study, subjects read rapidly presented word pairs (each word presented for 150 msec with a 100-msec inter-stimulus interval) that were either unrelated (e.g., celery giraffe), related (e.g., horse goat), or identical (e.g., camel camel). The investigators found that as the degree of semantic relatedness between the two words progressively increased for a particular category—in this case, animals—the neural activation evoked by the second word progressively decreased in the lateral fusiform gyrus, this being the same area that Chao et al. (1999) linked with the animal category. This repetition suppression effect may be the neural manifestation of automatic priming of semantically related words (see also Gold et al., 2006; Wible et al., 2006). Moreover, given the processing time constraints imposed by the task, it is highly unlikely that the fMRI results reflect explicit visual images that the subjects deliberately generated after understanding the words. Rather, as Wheatley et al. (2005, p. 1877) put it, “it may be that the visual image of an object is automatically retrieved as an unconscious and obligatory by-product of normal word reading. In this case, automatic, implicit generation of an object image would be the mechanism by which we access an important property underlying the meanings of words denoting concrete entities. In this sense, implicit visual imagery would be an obligatory component of reading for meaning.”

The convergent results of the studies by Chao et al. (1999) and Wheatley et al. (2005) suggest that the shape features of the meanings of concrete count nouns are captured by conjunctive neurons in ventral temporal CZs that not only overlap partially with the CZs that subserve visual perception of the very same properties (consistent with the Simulation Framework), but are also clustered according to semantic category (consistent with the Similarity in Topography principle). Further evidence for these ideas comes from a number of other studies (Martin et al., 1996; Thompson-Schill et al., 1999; Okada et al., 2000; Chao et al., 2002; Whatmough et al., 2002; Kan et al., 2003; Martin & Weisberg, 2003; Devlin et al., 2005; Mechelli et al., 2006; Noppeney et al., 2006; for studies focusing specifically on hemispheric asymmetries see Gilbert et al., in press; Lincoln et al., 2007). Thus, returning to the purely behavioral study by Zwaan et al. (2002) that I mentioned in section 2, it may be that when people read sentences like The ranger saw the eagle in the sky and The ranger saw the eagle in the nest, the contextually appropriate eagle images that they mentally simulate depend on the lateral fusiform gyrus.

3.2.1.c. Motion Properties

Yet another visual semantic component of many count nouns involves the characteristic motion patterns of the designated objects. To take a few simple examples, part of our knowledge of the word rabbit is the typical hopping movement of this kind of animal, and part of the meaning of scissors is the distinctive cutting movement of this kind of tool. As with the attributes of color and shape, the motion information encoded by count nouns appears to be neurally represented in cortical areas that are closely related to those that subserve perception of the same properties.

It is well-established that an area called MT (a.k.a. MT+, hMT+, V5, and hOc5), which resides in the posterior lateral occipito-temporal cortex, is critically involved in the passive perception of moving visual stimuli (Malikovic et al., 2007). Damage to this area causes akinetopsia, that is, acquired motion blindness, a rare neurological syndrome that is just as
striking as achromatopsia for its specificity (Zeki, 1991; Zihl et al., 1991). Although MT does not itself appear to distinguish systematically between different types of object-associated motion, it projects to higher-level posterolateral temporal areas that clearly do. These areas consist of property-specific CZs that are organized in the following way according to the Similarity in Topography principle. (See Figure 2, right panel). One processing stream extends from MT into regions of the posterior superior temporal sulcus that respond preferentially to the sight of biological (e.g., human) motion patterns, with further segregation according to body parts (for reviews of a large literature, see Grossman, 2006, and Blake & Shiffrar, 2007; for a neurophysiologically plausible computational model, see Giese & Poggio, 2003, and Giese, 2006). Another processing stream extends from MT into areas of the posterior middle temporal gyrus that respond preferentially to the sight of nonbiological (e.g., tool) motion patterns (Beauchamp et al., 2002, 2003; Martin & Weisberg, 2003; Beauchamp & Martin, 2007; see also Martin et al., 1996; Tranel et al., 1997a; Kellenbach et al., 2003; Boronat et al., 2005; Noppeney et al., 2005; Weisberg et al., 2007).

During the past twenty years, many functional neuroimaging studies have provided evidence that these posterolateral temporal areas—especially those in the left hemisphere—contribute not only to the transient perception, but also to the long-term semantic representation, of object-associated motion properties encoded by concrete count nouns. Some of the original studies employed the so-called “verb generate” paradigm in which subjects are presented with the name of an object (e.g., knife) and must produce a semantically related motion verb (e.g., cut) (for a review of early literature see Martin et al., 2000; for further discussion of motion verbs see section 3.2.2). More recently, it has been found that semantic representations of biological (e.g., human) and nonbiological (e.g., tool) motion properties are supported by the two segregated perceptual processing streams described in the previous paragraph. For instance, in the study by Chao et al. (1999) that I discussed in the context of shape properties, another important finding was that the pathway for biological motion patterns was recruited more than the one for nonbiological motion patterns during tasks involving either animal pictures or animal words, whereas the opposite asymmetry occurred during tasks involving either tool pictures or tool words. (See Figure 2, right panel). As shown in Figure 2, tool pictures as well as tool words also elicited activation in two regions—the anterior intraparietal sulcus and the ventral premotor cortex—that subserve visuomotor knowledge about how such objects are grasped and manipulated; however, I will not discuss these results further, since my focus here is restricted to visual information (for an in-depth review see Lewis, 2006). In short, the available data suggest that the motion patterns associated with the types of objects encoded by many count nouns are captured by conjunctive neurons in category-related CZs in the posterolateral temporal cortex that reside very close to those that are engaged during perception of the very same motion patterns.

3.2.1.d. Summary

The research reviewed above supports the idea that, as proposed by CZ theory and the broader Simulation Framework, the typical color, shape, and motion properties of various categories of objects encoded by count nouns depend on anatomically segregated regions of the temporal lobe that are either overlapping with or anterior to some of the regions that underlie perception of those properties. Consider, for example, the visual components of the meaning of
the word *elephant*: the canonical gray color of elephants seems to be captured by conjunctive neurons located close to those for color perception; the idiosyncratic shape of elephants seems to be captured by conjunctive neurons located close to those for shape perception; and the characteristic movements of elephants seem to be captured by conjunctive neurons located close to those for motion perception. When the word *elephant* is heard or read, these cortically distinct representations of color, shape, and motion properties are automatically activated in a top-down fashion, perhaps via certain conjunctive neurons in an anterior temporal region that houses higher-order modality CZs for binding in long-term semantic memory the diverse visual attributes of objects (Bright et al., 2007). This spatially distributed but temporally synchronous pattern of activation may, but need not, co-occur with an explicit conscious image of an elephant. Either way, the two most important points to bear in mind are as follows: first, the relevant fragments of semantic knowledge appear to be represented in modality-specific and, even more narrowly, property-specific format; and second, their transient co-activation upon presentation of the word *elephant* can be interpreted as constituting a rough re-enactment or simulation of how the properties would normally be represented during perception.

### 3.2.2. Words for Events

Verbs are the preferred class of words for describing events, and many verbs encode types of events that involve particular visual motion patterns. To take an especially well-studied case, Slobin (2000, 2003, 2004) notes that in English the multidimensional psychological space of “manner of locomotion” is intricately partitioned into discrete categories by well over 100 verbs that fall into specialized subclasses like the following: rapid motion (e.g., *dash, hurry, rush, scramble, sprint*), leisurely motion (e.g., *amble, drift, mosey, saunter, stroll*), smooth motion (e.g., *glide, slide, slink, slip, slither*), awkward motion (e.g., *limp, lurch, stagger, stumble, trip*), furtive motion (e.g., *creep, sidle, skulk, sneak, tiptoe*), manners of walking (e.g., *march, plod, sashay, strut, trudge*), and manners of jumping (e.g., *bound, hop, jump, leap, spring*). Such highly specialized verbs are not just dictionary entries, but are actively employed by speakers in a variety of contexts, including oral narrative, spontaneous conversation, creative writing, naming videoclips of motion events, and speeded fluency, i.e., listing as many motion verbs as possible in one minute (Slobin, 2003).

As I indicated in the discussion of the motion properties of objects, several functional neuroimaging studies have shown that the “verb generate” task elicits activation in the posterolateral temporal cortex anterior to MT, predominantly in the left hemisphere. If this reflects retrieval of stored visual semantic information about motion, one would expect to find similar patterns of activation during other tasks that require access to the motion component of verb meanings. This prediction has been confirmed in two studies by Joseph Kable and colleagues. In the first study, Kable, Lease-Spellmeyer, and Chatterjee (2002) asked subjects to perform the same kind of semantic similarity judgment task in two conditions, one with pictures of events as stimuli and the other with the corresponding verbs as stimuli. In both conditions, subjects saw three items in a triangular array—one at the top of the screen and two at the bottom—and had to determine whether the top item was more similar to the left-hand bottom

---

5 Hart and Kraut (2007a) propose that anatomically segregated feature representations are “glued” together during online semantic processing by means of neural firing rates that are synchronized with a 30 Hz gamma rhythm. This synchronization process may be mediated in part by the pulvinar nucleus in the thalamus.
item or to the right-hand bottom item. For example, they had to recognize that “digging” is more similar to “shoveling” than to “listening.” The investigators also ran control experiments using pictures of objects and corresponding nouns as stimuli. When subjects performed the judgment task with pictures, there was greater activation for events, compared to objects, in MT and in nearby portions of the posterolateral temporal cortex, with stronger signals in the right than the left hemisphere. The engagement of MT is consistent with previous evidence that this region responds not only to veridical motion but also to static pictures that imply motion, such as a dolphin photographed in mid-leap out of the ocean (Kourtzi & Kanwisher, 2000; Senior et al., 2000, 2002; Urgesi et al., 2006; Assmus et al., 2007; but see also Alford et al., 2007, and Fawcett et al., 2007). When subjects performed the judgment task with words, there was greater activation for verbs, compared to nouns, in portions of the posterolateral temporal cortex anterior to MT, but not in MT itself; moreover, the signals were stronger in the left than the right hemisphere. In a follow-up study, Kable, Kan, Wilson, Thompson-Schill, and Chatterjee (2005) replicated these findings under conditions that employed the same kind of judgment task but with stimuli that required greater discrimination of subtle aspects of the visual motion patterns encoded by verbs. For instance, subjects had to determine that “skipping” is more similar to “bouncing” than to “rolling.”

The authors offer the following interpretation of their discovery that words for events activate regions anterior to MT, whereas pictures of events activate MT itself: “These anterior-posterior differences between words and pictures could be evidence of a gradient of motion information represented in the occipito-temporal cortex, with areas closer to [MT] representing more concrete visual information and areas closer to the middle temporal gyrus representing more abstract propositional information” (Kable et al., 2005, p. 1863). The notion of a concrete-to-abstract gradient is plausible, but it is not clear if motion information really becomes more “propositional” as it becomes more “abstract.” From the perspective of CZ theory and the Simulation Framework, the information is still modality-specific in format, and it may consist of visual motion properties at varying levels of detail, captured by hierarchies of conjunctive neurons in CZs extended along the posterolateral temporal cortex. While it is indisputable that the types of motion patterns implied by static pictures are more concrete than those encoded by verbs, it is also the case that, as emphasized by both Slobin (2000, 2003, 2004) and Kable et al. (2005), the types of motion patterns encoded by verbs can be remarkably fine-grained. In fact, it is precisely for this reason that some linguists, such as Jackendoff (2002, p. 350), have argued that the semantic nuances distinguishing between verbs like walk, jog, limp, strut, and shuffle are best left to modality-specific sensorimotor systems.

In a recent fMRI study, Kemmerer, Gonzalez Castillo, Talavage, Patterson, and Wiley (2008) employed a “triads” paradigm exactly like the one used by Kable et al. (2002, 2005) in order to investigate the neural substrates of the following five classes of verbs, based on Levin (1993): Running verbs (e.g., run, jog, walk), Speaking verbs (e.g., shout, mumble, whisper), Hitting verbs (e.g., hit, poke, jab), Cutting verbs (e.g., cut, slice, hack), and Change of State verbs (e.g., shatter, smash, crack). Relative to a baseline condition, the five verb classes evoked complex and widely distributed patterns of activation that differed from each other in many theoretically important ways. For present purposes, however, what is most relevant is that all five classes significantly engaged the left posterolateral temporal cortex anterior to MT, just like in Kable et al.’s (2002, 2005) studies. Moreover, there was some evidence for a rough body-
part-based organization, since unique portions of the posterolateral temporal cortex were found to be activated by Running verbs (encoding distinctive leg movements), Cutting verbs (encoding distinctive hand movements), and Speaking verbs (encoding distinctive mouth movements).⁶ (See Figure 3).

Interestingly, in three other studies Daniel Tranel and colleagues claim to have identified links between verb processing and area MT itself, suggesting that, contrary to the findings of Kable et al. (2002, 2005) and Kemmerer et al. (2008), this region may actually play a role in representing the motion component of verb meanings (for other pertinent functional neuroimaging studies see Wallentin et al., 2005; Rüschemeyer et al., submitted; Pirog et al., submitted; for a pertinent behavioral study see Meteyard et al., in press). Right off the bat, it is important to note that all three of these studies employed pictures as stimuli, which raises the question of whether the putative involvement of MT was due to the motion patterns encoded by the verbs or to the motion patterns implied by the stimuli; however, as shown below, in the first two studies the nature of the contrasts between experimental conditions rules out the possibility that the observed activation was due only to the pictorial stimuli. First, Damasio, Grabowski, Tranel, Ponto, Hichwa, and Damasio (2001) presented subjects with pictures of actions being performed with tools (e.g., stirring a cup of coffee with a spoon) and asked them to name the actions in one condition (e.g., stirring) and to name the tools in another condition (e.g., spoon). The subtraction of naming tools from naming actions revealed robust activation in an area that the authors refer to as MT, suggesting that the engagement of this region was driven more by the verb retrieval task than by the pictorial stimuli. Second, Tranel, Martin, Damasio, Grabowski, and Hichwa (2005) used the same kind of experimental design as Damasio et al. (2001), with subjects performing both action-naming and tool-naming tasks for the same stimuli, except in this study the target verbs and nouns were homophones (e.g., comb to name both the action and the tool). Once again, the subtraction of naming tools from naming actions revealed robust activation in an area that the authors refer to as MT. Third, in a large-scale lesion study with 90 brain-damaged patients, Tranel, Kemmerer, Adolphs, Damasio, and Damasio (2003) found that many patients who failed two standardized tests that probe conceptual knowledge of actions (and that require judgments about pictorial stimuli) had lesions that included either MT or the white matter underneath it. These three studies seem to provide leverage for the idea that, contrary to Kable et al.’s (2002, 2005) and Kemmerer et al.’s (2008) findings, area MT may in fact comprise part of the neural basis for the visual motion component of verb meanings. Future research is clearly needed to investigate these issues more carefully, but transcending the empirical minutiae is the overarching discovery that the visual motion patterns encoded by verbs appear to be captured by conjunctive neurons in posterolateral temporal CZs that are closely related to those that subserve perception of the same properties. So, to return briefly to the purely behavioral study by Zwaan et al. (2004) that I described in section 2, it seems likely that when dynamic, egocentrically anchored simulations of motion are triggered by sentences like The shortstop hurled the softball at you or You hurled the softball at the shortstop, those simulations are

---

⁶ The area uniquely associated with Speaking verbs may reflect not only the activation of lip/tongue-related visual manner-of-motion representations—e.g., the distinctive appearances of screaming vs. whispering (Allison et al., 2000; Pelphrey et al., 2005)—but also the activation of vocalization-related auditory representations—e.g., the distinctive sound patterns, especially loudness levels, of screaming vs. whispering (Schirmer & Kotz, 2006).
neurally instantiated in a sector of the posterolateral temporal cortex that resides near the region that underlies actual motion perception (see also Kaschak et al., 2005).

Although this chapter is concerned primarily with how words capture visual experience, at this juncture it would be remiss to not mention that one of the most fascinating recent developments in our understanding of the neural substrates of verb meaning involves motoric rather than visual information (for reviews see Pulvermüller, 2005; Kemmerer, 2006a; Fischer & Zwaan, in press; for pertinent behavioral studies, see Glenberg & Kaschak, 2002; Zwaan & Taylor, 2006; Boulenger et al., 2006; Scorolli & Borghi, 2007; Borreggine & Kaschak, in press). Within the frontal lobes, both the primary motor cortex and the adjacent premotor cortex are somatotopically organized, which is to say that they contain maps of the muscular layout of the body, with the lips/tongue represented in the ventrolateral sector, the arms/hands represented in the dorsolateral sector, and the legs/feet represented in the dorsomedial sector. Single-cell recording studies with macaque monkeys, as well as studies using a variety of methods with humans, have revealed that body-part-specific motor areas are activated not only when certain types of actions are executed, but also when they are imagined, observed, or heard (for reviews see Rizzolatti & Craighero, 2004; Wilson & Knoblich, 2005; Jeannerod, 2006; for more recent studies see, e.g., Calvo-Merino et al., 2006; Gazzola et al., 2006; Pobric & Hamilton, 2006; Assmuss et al., 2007; Caetano et al., 2007; Filimon et al., 2007; Saygin, 2007; Van Shie et al., 2007). Most importantly in the present context, there is increasing evidence that the motoric information encoded by action verbs is also represented by motor-related structures in the frontal lobes. In particular, several studies employing a wide range of brain mapping techniques have shown that verbs encoding face actions (e.g., bite), arm/hand actions (e.g., punch), and leg/foot actions (e.g., kick) differentially engage the corresponding ventrolateral, dorsolateral, and dorsomedial sectors of somatotopically mapped primary motor and premotor regions (Hauk et al., 2004; Hauk & Pulvermüller, 2004; Pulvermüller et al., 2005a,b; Tettamanti et al., 2005; Buccino et al., 2005; Aziz-Zadeh et al., 2006; Rüschemeyer et al., 2007; Kemmerer et al., 2008; Glenberg et al., in press). These findings support the provocative notion that the motoric aspects of the meanings of action verbs—that is, what one might call the semantics of kinematics—are not part of an amodal symbolic representation in the brain, but are instead subserved by frontal cortical structures that overlap at least partly with those underlying the execution, imagination, observation, and audition of actions. Further evidence for this view comes from studies indicating that damage in the left premotor/prefrontal region disrupts knowledge of the meanings of action verbs (e.g., Kemmerer & Tranel, 2003; Bak et al., 2006; Silveri & Ciccarelli, in press; see also Tanel et al., 2003). However, as yet no neuropsychological studies have directly and systematically tested the hypothesis that the meanings of verbs for lips/tongue, arm/hand, and leg/foot actions should be differentially impaired by lesions affecting the relevant

---

7 Actually, the somatotopy is more complex than this. For instance, the lateral precentral gyrus of the macaque brain contains at least three separate hand representations (one in primary motor cortex, a second in ventral premotor cortex, and a third in dorsal premotor cortex) whose topographic partitioning reflects multiple, conflicting mapping requirements, including the need to encode distinct repertoires of complex, ethologically relevant movements, such as climbing/leaping behaviors, reaching behaviors, hand-to-mouth behaviors, defensive behaviors, and central space/manipulation behaviors (for reviews see Graziano, 2006; Graziano & Aflalo, 2007a). Regarding the motor representation of hand actions in the lateral precentral gyrus of the human brain, yet another mapping requirement may involve the development of neurons tuned to the idiosyncratic types of hand actions that are expressed by verbs in languages.
3.2.3. Words for Spatial Relations

Finally, speakers often use locative prepositions to describe the static spatial arrangements of objects in the visual field. During the past couple decades, the literature in linguistics on the meanings of locative prepositions has expanded dramatically, leading from the classic work of Talmy (1983) and Herskovits (1986) to a variety of new approaches, such as those of Evans (2003) and Conventry and Garrod (2004) (for a broad survey of recent proposals, see Evans & Chilton, 2008). In striking contrast, very little research in cognitive neuroscience has sought to identify the brain structures that mediate this complex and intriguing conceptual domain. In what follows, I begin by summarizing several aspects of the meanings of locative prepositions, and then I briefly review a few neuropsychological and functional neuroimaging studies that have focused on these types of words. As I will show, these studies suggest that the cortical areas underlying the linguistic encoding of spatial relations are close to, but nevertheless distinct from, those underlying the representation of spatial relations for purely perceptual purposes.

The types of spatial scenarios designated by locative prepositions generally involve two objects—the "figure" (F), which is the thing to be located, and the "ground" (G), which is an object that serves as a point of reference. For example, in the sentence The beer is in the refrigerator, the noun-phrase the beer specifies F, the noun-phrase the refrigerator specifies G, and the preposition in specifies the nature of the spatial relationship between them. For the most part, locative prepositions express spatial relations in terms of very sketchy or schematic structural properties of the objects involved; metrical details are usually ignored, such as the exact sizes, shapes, and orientations of the objects, or the precise distances between them. For instance, in the example given above, the real-world situation that the sentence refers to might consist of a geometrically rich, idiosyncratic, three-dimensional spatial layout, perhaps involving a Corona longneck standing upright on the top shelf of a big refrigerator with many levels and compartments; yet the semantic structure of in is very austere and skeletal, since it abstracts away from these spatial particularities and instead treats the beer bottle as just a dimensionless point and the refrigerator as simply an idealized container. Hence, locative prepositions designate "categorical" as opposed to "coordinate" spatial relationships (Postma & Laeng, 2006).

Two major subclasses of locative prepositions are often distinguished—topological and projective. Topological prepositions refer to spatial relations that involve various types of “coincidence” of F and G. Besides in, which expresses containment, other topological prepositions include on for contact, around for encirclement, and through for penetration. Some topological prepositions actually encode a combination of geometrical and functional features—e.g., on expresses not only contiguity but also the force-dynamic notion of support (Coventry & Garrod, 2004). Projective prepositions, on the other hand, employ a different strategy of locational reference for other sorts of spatial arrays. They specify the location of F as being within a search-domain that is projected from one of the major facets or dimensional axes of G. Thus, in front of and in back of (or behind) designate relations of anteriority and posteriority with respect to the front/back axis of G, and above and below designate relations of superiority and
inferiority with respect to the vertical axis of G. Determining the principal axes of G is not always straightforward, however, because it depends on the frame of reference that is adopted. Focusing on just the front/back distinction, many objects have what might be called an intrinsic front which is based on factors like the canonical direction of the sense organs (for people and animals), the canonical direction of motion (for vehicles), or the canonical direction of encounter (for TVs, computers, etc.). But some objects, like trees, lack an intrinsic front, in which case an egocentric "orientation mirroring" frame of reference is adopted so that the front of the object is conceptualized as the side facing the observer.

Many locative prepositions appear to have a network of distinct but closely related meanings that are organized around a central or prototypical meaning. For instance, in ideally describes a spatial relation of containment in which G (1) is a three-dimensional object, (2) is hollow, and (3) completely encloses F. But each of these conditions can be violated, thereby yielding extended meanings as exemplified in the following situations: (1) a person standing inside a circle painted on the floor (G is two-dimensional), (2) a nail that has been pounded into a board (G is solid), and (3) an apple in a bowl even though it rests of top of other fruit so that it is technically above the horizontal upper edge of the bowl (F not enclosed by G). Accounting for this remarkable flexibility of locative prepositions has proven to be a difficult challenge for semantic analysis.

Turning to the brain, Landau and Jackendoff (1993) speculated that the meanings of locative prepositions might be neurally instantiated in the left inferior parietal lobule, a brain region known to be involved in the perceptual representation of schematic, categorical spatial relations between objects (for reviews see Jager & Postma, 2003; Laeng et al., 2003; Postma & Laeng, 2006). Several recent studies have not only corroborated this proposal but have made it more precise by suggesting that the critical cortical region may be the left supramarginal gyrus (SMG; for reviews see Kemmerer, 2006b, 2008). Damasio, Grabowski, Tranel, Ponto, Hichwa, and Damasio (2001) report a PET study in which English speakers viewed drawings of static spatial relations between objects (e.g., a cup on a table) and performed two tasks: naming F, and naming the spatial relation between F and G with an appropriate preposition. When the condition of naming objects was subtracted from that of naming spatial relations, the largest and strongest area of activation was in the left SMG. The authors did not indicate which prepositions were targeted for production, but it appears that a mixture of topological and projective prepositions were included, which suggests that the SMG activation reflects semantic processing of both types.

Additional evidence comes from a neuropsychological study conducted by Tranel and Kemmerer (2004; see also Kemmerer & Tranel, 2000a, 2003; Kemmerer, 2005). They administered a set of four standardized tests that collectively evaluate production, comprehension, and semantic analysis of 12 English prepositions (encoding topological relations as well as several kinds of projective relations) to 78 brain-damaged patients with lesions distributed throughout the left and right cerebral hemispheres, and then contrasted the lesion sites of the patients who were impaired on the tests with the lesion sites of those who were unimpaired. Poor performance was linked with damage in the left SMG and the left frontal operculum. (See Figure 4). The involvement of the left SMG strengthens the hypothesis that this region plays an essential role in representing the meanings of locative prepositions (for
convergent neuropsychological data see Wu et al., 2007). The investigators did not, however, conduct separate analyses to determine whether the different semantic classes of prepositions dissociated from each other behaviorally and neuroanatomically, nor did they investigate whether prototypicality influenced the results. As for the involvement of the left frontal operculum, it may reflect either or both of two functions: phonological encoding, possibly in Brodmann area 44 (Amunts et al., 2004), and semantic working memory, possibly in Brodmann areas 45 and/or 47 (Badre & Wagner, 2007).

Figure 4 about here

In a follow-up experiment with just those patients who failed all four preposition tests (N=6), Tranel and Kemmerer (2004) assessed nonlinguistic visuospatial processing by administering a large battery of standardized neuropsychological tests (Benton & Tranel, 1993). Although a few of the tests emphasize sensitivity to metrically precise coordinate spatial relations (e.g., subtle variations in line orientation), the majority of them require an appreciation of more schematic categorical spatial relations (e.g., the typical arrangements of the parts of complex objects). Overall, the patients performed extremely well on the various tests, revealing a strong dissociation between impaired linguistic and preserved perceptual processing of spatial relations. Moreover, Kemmerer and Tranel (2000a) describe a patient who manifested a dissociation that was the opposite of the kind manifested by the patients in Tranel and Kemmerer's (2004) study—namely, intact knowledge of the meanings of locative prepositions but impaired nonlinguistic visuospatial processing of both coordinate and categorical spatial relations.

This neuropsychological double dissociation constitutes evidence that the spatial image schemas expressed by locative prepositions are separate from those that are required in order to execute certain kinds of visuospatial tasks (for convergent behavioral data see Munnich et al., 2001; see also Kemmerer, 1999). In terms of CZ theory, these two types of spatial representations appear to be mediated by distinct populations of conjunctive neurons in anatomically close but, importantly, non-overlapping portions of the left inferior parietal lobule, especially the SMG. Does this pose a challenge to the Simulation Framework? Not necessarily. For one thing, even though the meanings of locative prepositions do not seem to be needed in order to perform certain types of perceptual tasks, it is still reasonable to assume that they are perceptually grounded. After all, they derive in large part from visual experience, and it is quite possible that they employ the relatively high-level representational resources of this particular modality to capture various austere categories of spatial relations such as containment (in), contact (on), and penetration (through). Pursuing this line of thinking a step further, it may even be the case that the transient activation of prepositional meanings during online language processing involves, at least to some extent, embodied simulations of skeletal spatial arrays, most likely modulated by contextual factors. This is clearly a strong prediction of the Simulation Framework (e.g., see many of the chapters in Hampe, 2005), but I am not aware of any experimental studies that have attempted to test it. If such simulations do occur, however, their content is probably very language-specific in content; that is, it probably reflects the construal of space for communicative purposes much more than for purely perceptual purposes. Gentner and Boroditsky (2001) make essentially the same point when they state that spatial relational concepts are among the least likely to be “given by the world” and hence tend to exhibit strong
language-specific influences. Support for this view comes not only from the neuropsychological double dissociation reported by Tranel and Kemmerer (2004), but also from the discovery of extensive crosslinguistic variation in the semantic domain of spatial relations, as discussed below in section 4.3.

3.3. Summary

In recent years, a rapidly growing body of research in cognitive neuroscience has been generating evidence for the view that word meanings are not represented in a single amodal brain region, like entries in a dictionary. Instead, they appear to be stored in multiple, anatomically distributed, modality-specific and, even more narrowly, property-specific cortical areas, and their ephemeral activation during online language processing seems to involve the automatic, implicit recapitulation of sensorimotor states. I have reviewed a number of studies that focus on the neural substrates of the visual semantic components of words for objects, events, and spatial relations. Overall, these studies suggest that different visual semantic components are housed in cortical areas that are closely related to those that underlie the perception of the relevant properties. In particular, for the color, shape, and motion components of word meanings, there is mounting evidence that the online activation of the pertinent information takes the form of covert, property-specific, perceptual simulations. Such simulations may also occur for the types of spatial image schemas that are encoded by locative prepositions, but this remains to be carefully investigated.

4. SOME QUESTIONS RAISED BY CROSSLINGUISTIC VARIATION

The material covered in the previous section demonstrates that cognitive neuroscience is starting to contribute, in its own unique way, to our understanding of how words capture visual experience. So far, however, this approach has been applied almost exclusively to English. In fact, as someone who was originally trained in linguistic typology before migrating to the field of cognitive neuroscience, I am acutely aware of the fact that crosslinguistic variation in lexical-semantic systems—variation of the sort that is discussed in many of the other chapters in this volume—has been almost entirely ignored by the vast majority of researchers who are currently investigating the organization, representation, and processing of conceptual knowledge in the brain (for an attempt at consciousness raising, see Kemmerer, 2006b). In this section I present thumbnail sketches of just a few forms of variation involving the lexical encoding of three types of visual information—shape, motion, and spatial relations—and for each one I consider some questions that arise concerning the underlying neural substrates.  

4.1. Shape

As noted earlier, Landau and Jackendoff (1993) maintain that the vocabulary of the average adult English speaker contains roughly 10,000 count nouns that designate types of objects that are bounded and individuated primarily according to inherent shape. At the same time, however, English speakers also make regular use of a large number of mass nouns that

---

8 Due to space limitations, I will not discuss a fourth type of visual information, namely color.
designate types of substances with no inherent shape. In order to describe a particular instance of a certain type of substance, speakers often combine the appropriate mass noun with one of several possible terms that provide rough geometric information about the form that the substance happens to take on that occasion—e.g., a sheet of paper, a stack of paper, a pile of paper, a wad of paper, etc.

Some languages in the world have predominantly mass nouns together with rich systems of nominal classification that are much more elaborate than what is found in English and other Indo-European languages (e.g., Senft, 2000; Aikhenvald, 2003). For example, Lucy (1992) has argued that in Yucatec Maya (Mayan, Mexico) the noun lo’bal could be glossed “banana stuff” because, depending on which shape-discriminating classifier it co-occurs with, it can be used to refer to a single banana, a bunch of bananas, a bit of the fruit, the leaf of the plant/tree, or the whole plant/tree:

\[
\begin{align*}
\text{'un-tz'ii}t & \; \text{lo'bal} & \text{“one 1-dimensional banana (i.e., the fruit)”} \\
\text{'un-kuch} & \; \text{lo'bal} & \text{“one load banana (i.e., the bunch)”} \\
\text{'un-p'ii}t & \; \text{lo'bal} & \text{“one bit banana (i.e., a bit of the fruit)”} \\
\text{'un-waal} & \; \text{lo'bal} & \text{“one 2-dimensional banana (i.e., the leaf)”} \\
\text{'un-kuul} & \; \text{lo'bal} & \text{“one planted banana (i.e., the plant/tree)”}
\end{align*}
\]

Likewise, a single classifier in Yukatec Maya can impose the same kind of individuation on a variety of different materials, “in much the way that a cookie-cutter cuts up undifferentiated dough,” as Levinson (1996, p. 185) observed in a discussion of Lucy’s (1992) study:

\[
\begin{align*}
\text{'un-tz'ii}t & \; \text{lo'bal} & \text{“one 1-dimensional banana (i.e., the fruit)”} \\
\text{'un-tz'ii}t & \; \text{kib'} & \text{“one 1-dimensional wax (i.e., a candle)”} \\
\text{'un-tz'ii}t & \; \text{che'} & \text{“one 1-dimensional wood (i.e., a stick)”} \\
\text{'un-tz'ii}t & \; \text{nal} & \text{“one 1-dimensional corn (i.e., an ear)”}
\end{align*}
\]

Crosslinguistically, the semantic features of shape that are most commonly expressed by classifiers are the one-dimensional (1D) long shape, the two-dimensional (2D) flat shape, and the three-dimensional (3D) round shape. In some languages, however, more fine-grained features are encoded. A good illustration is Japanese, which, according to Sanches (1977), has the following system for classifying entities according to predominant dimension: 1D (-hon); 2D with length predominating (-suji); 2D with length and breadth equally important (-mai); 2D with height and breadth equally important (-men); 3D with length and breadth predominating (-hen); 3D cubic (-cho); 3D irregularly shaped (-kai); and 3D spherical (-ko). Based in part on an analysis of how Japanese children acquire this complex system, Inoue (2000) proposes that classifier choice is highly dependent on what she calls “visualizing ability”—a notion that accords nicely with the Simulation Framework. As she puts it, “speakers have to imagine just how round or tall an object-noun can be in order for the referent to still be associated with the numeral classifier hon (‘cylindrical object’). If a pen is a prototypical instance of a referent for the numeral classifier hon, how about a 10-meter telephone pole? Or a coffee mug?” (Inoue, 2000, p. 219).
A small but growing body of literature in cognitive neuroscience has begun to address both semantic and syntactic aspects of the count/mass distinction in English and in a few other Indo-European languages, especially Italian (for a short summary, see Semenza, 2005; for studies investigating specifically the semantic aspects of the distinction, see Bisiacchi et al., 2005; Taler et al., 2005). But as yet no research has explored the neural substrates of the kinds of shape classifiers mentioned above. Are the highly schematic meanings of such classifiers captured by conjunctive neurons in the ventral temporal cortex, and if so, where exactly do they reside? Could future research in cognitive neuroscience help explain why the crosslinguistically most prevalent shape classifiers are those that specify 1D, 2D, and 3D forms? How does the cortical representation of lo’bal in Yucatec Maya differ from that of banana in English, especially with regard to the regions of the fusiform gyrus that store the shape properties of objects? For speakers of Yucatec Maya, does lo’bal automatically trigger perceptual simulations of not only single bananas, but also bunches of bananas, the leaf of the plant/tree, or even the whole plant/tree? These are only a handful of the many intriguing questions that classifier languages pose for cognitive neuroscience in general, and for the Simulation Framework and CZ theory in particular.

4.2. Motion

As already indicated, English has a well-developed inventory of “manner of locomotion” verbs that encode subtle semantic distinctions along multiple dimensions such as visual pattern, motor pattern, rate, and social-emotional evaluation. But this inventory differs significantly from the inventories in other languages. Even English and German, which are closely related languages, differ somewhat in this domain—e.g., German has no exact equivalents to English scamper, scurry, scuttle, and scramble, and English has no exact equivalents to German stapfen, stiefen, trampeln, and stampfen, which designate different kinds of firm, heavy walking (Snell-Hornby, 1983). These differences are quite minor, however, in comparison to the much more substantial variation that has been documented worldwide (Levinson & Wilkins, 2006; but see also Malt et al., in press). Building on a foundation of previous work by a number of typologists, Slobin (2004) reports extensive crosslinguistic diversity in how motion events are described, and he attempts to account for this diversity by positing a “cline of manner salience.” In high-manner-salient languages (e.g., the Germanic and Slavic languages, Hungarian, and Mandarin), there is an easily accessible morphosyntactic slot for expressing manner (such as the main verb position in English), and this encourages speakers to attend to, and eventually lexicalize, increasingly fine-grained manner details, which in turn causes a rich lexicon of manner morphemes to arise diachronically and to even influence co-speech gesture (Kita & Özyürek, 2003; see also Kemmerer et al., 2007). At the other end of the continuum, in low-manner-salient languages (e.g., the Romance languages, Turkish, and Hebrew), manner information is grammatically subordinated to path information, so the former is provided only when it must be foregrounded for some reason, and there is less motivation to create a large inventory of manner morphemes. The tremendous difference between the expressive resources of languages at the two extremes can be seen by contrasting English with both Spanish and French. For example,
creep, glide, slide, slip, and slither⁹ are all translated into Spanish as escabullirse, and bound, hop, jump, leap, and spring are all translated into French as bondir.

Shifting to the brain, it is interesting to consider that CZ theory leads to the following hypothesis regarding the neural substrates of the visual component of locomotion verbs. Perhaps the crosslinguistic diversity in the lexicalization of manner of locomotion is reflected, at least in part, in corresponding neurobiological diversity in the spatial arrangements and “tuning curves” of conjunctive neurons in CZs within the mosaic of cortical areas extending anteriorly from area MT into the posterior superior temporal sulcus, including the region that responded preferentially to Running verbs in Kemmerer et al.’s (2008) fMRI study (see Figure 3). According to this hypothesis—more precisely, according to the Similarity in Topography principle—the layout of the relevant conjunctive neurons is systematically different for English speakers compared to, say, Spanish speakers. For English speakers there are separate but tightly clustered conjunctive neurons for the closely related visual motion patterns encoded by creep, glide, slide, slip, and slither; however, for Spanish speakers such conjunctive neurons do not exist because (1) the Spanish manner verb lexicon does not make any of those subtle semantic distinctions (the whole spectrum is covered by just one verb, escabullirse), and (2) there is no independent reason to expect those particular distinctions to be "natural" in the sense of being universally employed in the nonverbal categorization of motion events (see Slobin, 2000, 2003, for relevant data and discussion from the perspective of language acquisition). As the spatial resolution of brain mapping techniques continues to improve, it may eventually become feasible to test hypotheses of this nature, thereby shedding further light on the biological bases of the meanings of motion verbs. For present purposes, the essential point is this: It may not be a coincidence that prominent theorists in both linguistic typology (e.g., Croft, 2001; Haspelmath, 2003) and cognitive neuroscience (e.g., Simmons & Barsalou, 2003; Crutch & Warrington, 2003) increasingly use the mapping metaphor (i.e., analyzing semantic domains in terms of geometric spaces) in their characterizations of the organization of conceptual knowledge. Perhaps the metaphor is more appropriate than we have hitherto realized (for reviews of pertinent neurocomputational modeling, see Kohonen & Hari, 1999; Graziano & Aflalo, 2007b).

4.3. Spatial relations

Even greater crosslinguistic variation has been documented in the domain of spatial relations than in the domains of physical shape and manner of locomotion, perhaps because, as noted at the end of section 3.2.3, spatial relational concepts are less likely to be “given by the world” (Gentner & Boroditsky, 2001). Several recent books provide in-depth descriptions of the remarkable diversity, as well as the many overarching commonalities, of the categorical spatial coding systems manifested in carefully selected samples of the 6000+ languages of the world

---

⁹ Another member of this family is slink, which is how Dr. Seuss describes the movement of the Grinch in his well-known children’s story How the Grinch Stole Christmas: “Then he slithered and slunk, with a smile most unpleasant, around the whole room, and he took every present!” It is noteworthy that the word-initial consonant cluster sl- seems to have phonesthetic qualities, and that many languages actually have large inventories of ideophones that capture different manners of motion. Slobin (2004, p. 251) points out, for instance, that Westermann’s (1930) grammar of Ewe (Niger-Congo, West Africa) gives examples of 37 ideophones that can be used with the verb [for] “walk,” with the additional information that these forms can be reduplicated and can occur with high tone for diminutives and low tone to describe motions of large entities.”
(Levinson, 2003b; Levinson & Wilkins, 2006). Here I mention just a few forms of variation involving topological and projective relations, and I briefly introduce some questions and hypotheses about how these types of variation might be implemented in the brain (for a deeper discussion of these issues, see Kemmerer, 2006b, 2008).

In the topological realm, a great deal of crosslinguistic diversity has been documented, but several underlying patterns have also emerged. In a recent study reported by Levinson and Meira (2003), nine unrelated languages were investigated by comparing native speaker responses to a standardized set of 71 pictures showing a wide range of topological relations (for more recent developments see Levinson & Wilkins, 2006). The results indicated that crosslinguistically the labels for pictures were not randomly distributed but instead tended to cluster, suggesting that the topological domain forms a coherent similarity space with a number of strong “attractors,” i.e., taxonomically basic-level categories that are statistically likely to be recognized by languages—in particular, notions such as containment, attachment, superadjacency, subadjacency, and proximity. Several generalizations about the organization of this abstract similarity space were uncovered. First, each core concept has a prototype structure. For example, at the center of the cluster of containment pictures were scenes in which F is enclosed within G (e.g., a dog in a cage); scenes involving partial two-dimensional containment on a planar surface (e.g., a dog in a yard) were more peripheral, implying that English is somewhat unusual in using in for such topological relations. Second, the core concepts are arranged as neighbors along gradients in the similarity space, making some conflations of categories more natural than others. For instance, English on embraces both superadjacency (e.g., a cup on a table) and attachment (e.g., a picture on a wall), Berber di embraces both attachment (e.g., a picture on a wall) and containment (e.g., an apple in a bowl), and Spanish en embraces all three categories; however, there should not be, and do not as yet appear to be, any languages with a spatial morpheme that applies to superadjacency and containment while excluding attachment, since the latter concept is intermediate between the other two along the relevant gradient of the abstract similarity space. Third, each core concept can be further fractionated, leading to more fine-grained categories of topological relations. For example, the cluster of pictures for superadjacency included scenes both with and without contact (e.g., a cup on a table, and a lamp above a table), suggesting that languages are likely to use the same morpheme for these kinds of relations—a tendency that seems somewhat surprising from the perspective of English, since on and above/over divide the superadjacency category into separate subcategories distinguished by the presence or absence of contact between F and G. Levinson and Meira also report many intriguing cases of category fractionation in other languages, such as the exotic Tiriyó morpheme asee, glossed "astraddle," which applies to the subset of attachment pictures in which F is suspended from a point on G and hangs down on either side of it (e.g., a coat on a hook, an earring dangling from a person's ear, a pendant on a chain, clothes drying on a line, a balloon on a stick, and a tablecloth on a table).

The neuropsychological and functional neuroimaging studies that I summarized in section 3.2.3 suggest that the left inferior parietal lobule—in particular, the SMG—is a key cortical region for representing the meanings of English locative prepositions, including those

---

10 Basque (Isolate, Europe), Dutch (Indo-European, Europe), Ewe (Niger-Congo, West Africa), Lao (Tai-Kadai, Southeast Asia), Lavukaleve (Isolate, Solomon Islands), Tiriyó (Cariban, South America), Trumai (Isolate, South America), Yélî Dnye (Isolate, Papua New Guinea), Yukatec Maya (Mayan, Mexico).
that encode topological spatial relations. Further research on the neural correlates of linguistically encoded topological relations could benefit greatly by utilizing carefully designed stimuli that take into account theoretically important semantic dimensions, like the standardized set of 71 pictures that Levinson and Meira (2003) employed in their crosslinguistic comparison (see also Levinson & Wilkins, 2006). By conducting high-resolution brain mapping studies with such materials, it might be possible to test the hypothesis that the conceptual similarity space discovered by Levinson and Meira (2003)—a similarity space organized in terms of notions such as containment, attachment, superadjacency, subadjacency, and proximity—is neuroanatomically implemented in the form of a topographically structured cortical map in the left inferior parietal lobule, most likely the SMG. Within this map, the representational dimensions of the conceptual space might be captured, albeit in a warped manner, by the physical distribution of neuronal ensembles, in accord with the Similarity in Topography Principle.

Turning to the realm of projective spatial relations, I indicated earlier that when English speakers need to say that F (e.g., a ball) is located within a region that is projected from a G without an intrinsic front or back (e.g., a tree), they use an egocentric "orientation mirroring" frame of reference, so that the front of G is construed as the side facing the observer. Thus, a statement like *The ball is in front of the tree* means that the ball is between the tree and the observer. However, in some languages, such as Hausa (West Chadic, Niger), an egocentric “orientation preserving” frame of reference is employed instead, so that the observer’s front-back bodily axis is mapped onto G without any rotation. In this type of system, a statement like *The ball is in front of the tree* means that the ball is on the opposite side of the tree as the observer. Yet another strategy is to avoid the egocentric frame of reference entirely by rejecting the assumption that trees lack intrinsic fronts and backs. This approach is taken by Chamus (Nilo-Saharan, Kenya), which treats the front of a tree as the side it leans toward, or, if it is vertical, the side with the biggest branch or the most branches. An even more radical departure from English can be found in languages that specify all projective relations on the horizontal plane in terms of an absolute frame of reference that provides a set of fixed bearings or cardinal directions. In systems of this sort, one might describe the spatial array involving the ball and tree by saying *The ball is north of the tree*, a statement that is completely independent of both the observer’s perspective and the intrinsic facets of the tree. Absolute systems are fundamentally geocentric, and languages often base terms for cardinal directions on stable environmental features like mountain slopes, river drainages, and prevailing wind patterns. For example, Tzeltal (Mayan, Mexico) has an absolute system that is anchored in the mountain incline of the local landscape, giving rise to three directional terms: *alan* "downhill" (roughly north), *ajkol" uphill" (roughly south), and *jejch" across" (either east or west) (Brown & Levinson, 1993; Brown, 2006). Although the terminology of absolute systems typically derives from environmental landmarks, such systems are fully abstracted (at least in some languages\textsuperscript{11}), and in order to use them spontaneously and accurately, speakers must constantly monitor their spatial orientation by running a kind of mental compass. This is a truly remarkable ability, as demonstrated by Levinson (2003b). Here is just one of many eye-opening examples and analyses that he presents throughout his book, this particular one focusing on Guugu Yimithirr (Pama-Nyungan,\textsuperscript{11})

\textsuperscript{11} Determining whether the absolute system in a given language is fully abstracted from its geocentric anchor(s) requires careful fieldwork, as shown, for instance, by Schultz-Berndt’s (2006) investigation of Jaminjung (Jaminjungan, Australia). This language has an absolute system based on water flow and verticality, but it “breaks down for reference beyond the drainage system which includes the territory the speakers are familiar with” (p. 105).
Australia), whose speakers use exclusively the absolute frame of reference for characterizing horizontal projective relations (p. 114): “In GY, in order to describe someone as standing in front of the tree, one says something equivalent (as approximate) to 'George is just north of the tree', or, to tell someone to take the next left turn, 'go north', or, to ask someone to move over a bit, 'move a bit east', or, to instruct a carpenter to make a door jamb vertical, 'move it a little north', or, to tell someone where you left your tobacco, 'I left it on the southern edge of the western table in your house,' or, to ask someone to turn off the camping gas stove, 'turn the knob west,' and so on. So thoroughgoing is the use of cardinal directions in GY that just as we think of a picture as containing virtual space, so that we describe an elephant as behind a tree in a children's book (based on apparent occlusion), so GY speakers think about it as an oriented virtual space: if I am looking at the book facing north, then the elephant is north of the tree, and if I want you to skip ahead in the book I will ask you to go further east (because the pages would then be flipped from east to west).”

Languages like this constitute an excellent example of how semantic typology can inspire future research on the neural representation of categorical spatial relations. As noted above, the speakers of such languages must constantly compute their orientation within a conventional framework of fixed bearings. Many non-human species have evolutionarily specialized sensory devices that enable them to use absolute coordinates for navigation—e.g., some species of migratory birds have light-absorbing molecules in their retinas that are sensitive to the magnetic field of the earth and that may enable the birds to see this information as patterns of color or light intensity (Ritz et al., 2004); sea turtles have the biological equivalent of a magnetically based global positioning system that allows them to pinpoint their location relative to geographically large target areas (Luschi et al., 2007); and locusts perceive polarization patterns in the blue sky and use them as cues for spatial orientation (Heinze & Homberg, 2007). But for people in "absolute" communities, the mental compass that generates their superb sense of direction—a sense that may be comparable in accuracy to that of homing pigeons (Levinson, 2003b, p. 232)—is presumably not genetically programmed but may instead be a "knock-on" effect of the intensive training in orientation tracking that comes with speaking a language that regularly employs cardinal direction terms to describe spatial arrays at every level of scale (Levinson, 2003b, p. 278; see also Haun et al., 2006). It is reasonable to suppose that the relevant brain areas include parietal and hippocampal structures that have been implicated in both constructing landmark-based cognitive maps of the environment and monitoring one's movement through them (e.g., Ekstrom et al., 2003; Hartley et al., 2003; Janzen & van Turennout, 2004; Hafting et al., 2005; Spiers & Mcguire, 2006; Leutgeb et al., 2007). However, because the use of the mental compass does not necessarily require input from visually perceived landmarks, other neural systems must also be recruited, presumably to carry out the computations that underlie dead-reckoning—that is, keeping track of distances traveled along each angular heading (Sargolini et al., 2006). To be sure, I would hardly expect my fellow brain scientists to attempt to set up an fMRI scanner in Hopevale, North Queensland, Australia in order to gather functional neuroimaging data about the biological bases of spatial description in Guugu Yimithirr. But I do hope that in the coming years an increasing number of these researchers will develop a genuine

12 For intriguing proposals about similarities between navigating through physical space and navigating through episodic and semantic space, see Buzsaki (2005).
appreciation of the neuroscientific questions raised by languages like this, and that ultimately such questions will somehow be addressed.

4.4. Summary

Recent research in the nascent field of semantic typology has begun to reveal considerable crosslinguistic variation in numerous domains of word meaning. Studies along these lines point to what are sometimes called language-specific semantic maps across universal conceptual spaces. I have briefly surveyed some of the major forms of diversity in the visual semantic fields of shape, motion, and spatial relations. Crosslinguistic variation of this nature has already had a significant influence on developmental psychologists who investigate the acquisition of language during childhood (e.g., Bowerman & Levinson, 2001; Guo et al., 2008) as well as on cognitive scientists who investigate the relation between language and thought (e.g., Gentner & Goldin-Meadow, 2003; Gleitman & Papafragou, 2005); however, it has largely been ignored by cognitive neuroscientists who investigate the organization, representation, and processing of conceptual knowledge in the brain. This is unfortunate, because typological data raise many interesting questions about the neural substrates of word meaning, questions that go beyond the idiosyncrasies of English to embrace much richer patterns of similarities and differences among the 6000+ languages of the world.

5. CONCLUSION

In this chapter I have presented the perspective from cognitive neuroscience on the issue of how words capture experience. I have deliberately concentrated on the lexical encoding of the visual world, because that is where the most progress has been made with regard to identifying the underlying brain structures. Although the neuropsychological and functional neuroimaging studies that I have reviewed are restricted entirely to English and are therefore quite parochial from the point of view of semantic typology, they are nevertheless broadly consistent with the Simulation Framework insofar as they suggest that different kinds of visual semantic components of words depend on cortical regions that are closely related to those that process the same properties during perception. I have not discussed any studies involving semantic representations in non-visual sensory modalities, but it is worth noting that they provide further evidence for the Simulation Framework (e.g., Kellenbach et al., 2001; Simmons et al., 2005; James & Gauthier, 2003; Goldberg et al., 2006a,b). For instance, Goldberg, Perfetti, and Schneider (2006a) found that making semantic decisions about visual, auditory, tactile, and gustatory aspects of word meaning activated the corresponding sensory brain regions.

In a recent summary of the Simulation Framework, Zwaan and Madden (2005, p. 242) point out that this theory will seem trivial to some people yet counterintuitive to others: “[It] will seem trivial to the lay person, or even to people with great expertise in the use of language, such as novelists and poets. Of course words can be used to conjure up images in the reader’s [or listener’s] mind! However, these same claims will seem counterintuitive to researchers trained in traditional cognitive science. To them, the claim that meaning can be captured by experiential representations does not make sense.” I submit that this is because the core assumption of the traditional view is that semantic structures are by definition entirely amodal in character. It is
precisely this assumption—one might even call it an axiom—that advocates of the Simulation Framework are attempting to challenge through careful theoretical argumentation and an increasingly sophisticated and compelling progression of psychological and neuroscientific investigations.

Many questions remain unresolved, however, especially regarding the true nature and function of lexically triggered simulations of sensorimotor states. In my own opinion, one of the top priorities for future research should be to clarify, both theoretically and empirically, the distinctions between the following three types of modality-specific information processing. First, there are conscious perceptual experiences, as when one opens one’s eyes and sees, for example, a dog. These experiences are subserved by a vast array of neural mechanisms, beginning in the retina and proceeding all the way up to, and beyond, the highest levels of the visual system. Second, there are explicit perceptual simulations, as when one closes one’s eyes and voluntarily, effortfully imagines a dog. Although the topic of mental imagery has a long and controversial history, the weight of evidence strongly favors the view that the top-down generation of explicit visual images shares a great deal, but certainly not all, of the cortical territory underlying bottom-up perception (for a comprehensive review see Kosslyn et al., 2006). Third, there are implicit perceptual simulations, as when one hears or reads the word dog, and the word automatically, reflexively triggers an unconscious visual representation of a dog. According to the Simulation Framework, these types of simulations constitute substantial portions of the meanings of words. However, relatively little attention has been devoted to delineating the precise ways in which they are similar to and different from the types of representations that are employed in veridical perception and explicit imagery. As I emphasized in the review of experimental studies in section 3.2, cognitive neuroscientists have frequently found that when subjects perform semantic tasks involving, say, the shape properties of objects encoded by concrete count nouns, there is activation in some of the same cortical areas that are engaged during both the perception and the imagination of those properties. And yet it is patently obvious that understanding a word like dog is not at all the same thing as passively seeing a dog out in the world, nor is it the same thing as voluntarily evoking an explicit image of a dog in the mind’s eye. A few cognitive neuroscientists, such as Alex Martin and his colleagues, have recently started to use advanced methodologies, such as repetition suppression and priming paradigms, in order to tease apart these three kinds of modality-specific information processing, with the ultimate goal of isolating and properly characterizing the neural substrates as well as the cognitive content of the visual semantic components of words. This line of research is still in its infancy, however, and most of the work still lies ahead.

The usual goal of communication is, of course, to set up “the same thought” in the receiver’s brain as is currently taking place in the sender’s brain. In this context, words can be regarded as instructions for running embodied simulations. As noted above, however, these simulations are generally implicit, which is to say, covert, unconscious, and automatic. Although we can, if we wish, bring these sensorimotor re-enactments into the light of awareness, our brains usually hide them from us, presumably so we can consciously attend to other things, such as how the message fits into our belief system, whether the speaker is being sincere, how we intend to respond, and so on (Frith, 2007). The implicit nature of lexically triggered simulations is actually quite consistent with other recent discoveries about the fundamental role that embodied simulations play in our mental lives, especially when it comes to reflexively
understanding each other’s actions, emotions, and sensations. The basic idea is simple: By virtue of having common brain circuits and common sensorimotor and affective experiences, people can, so to speak, automatically translate the sights and sounds of what other individuals do and feel into the language of their own actions and feelings (for reviews see Gallese et al., 2004; Wilson & Knoblich, 2005; Decety & Lamm, 2006; Jeannerod, 2006; Keysers & Gazzola, 2006; Haggard et al., 2007). Among the many forms of evidence for this view are the following findings. First, as I pointed out at the end of section 3.2.2, the observation of an action engages some of the same somatotopically mapped neural networks that are active during its execution (e.g., Buccino et al., 2001; Sakreida et al., 2005). This kind of motor resonance is strongest for goal-directed actions that the observer is skilled at performing (e.g., Calvo-Merino et al., 2006; Cross et al. 2006), but it can also be triggered by (a) “degraded” actions that are perceived only as point-light displays (e.g., Saygin et al., 2004; Saygin, 2007), (b) actions that are merely implied by static pictures (e.g., Longcamp et al., 2006; Urgesi et al., 2006), and (c) actions that are heard but not seen (e.g., Gazzola et al., 2006; Caetano et al., 2007). Second, deciphering the intentions of observed actions—e.g., whether a person who is grasping a cup intends to drink from it or clean it—also evokes motor simulations (e.g., Iacoboni et al., 2004; Hamilton & Grafton, 2007; Kilner et al., 2007). Third, recognizing and empathizing with other people’s emotions involves covertly recapitulating the types of body states that generate them (e.g., Wicker et al., 2003; Avenanti et al., 2005; Harrison et al., 2006). Fourth, the observation of someone being touched on a particular part of their body induces activation in the somatosensory cortices of the viewer, as if the viewer’s own body were the subject of tactile stimulation (e.g., Keysers et al., 2004; Hasson et al., 2004; Blakemore et al., 2005). Taken together, these findings, together with a large and rapidly growing assortment of additional discoveries involving both normal and pathological cognition (e.g., Iacoboni & Dapretto, 2007), provide strong support for the Simulation Framework. The point I wish to highlight, however, is that all of these types of sensorimotor re-enactments are usually implicit in the same way that those underlying word comprehension are usually implicit. Overall, the available data suggest that the same neurocognitive mechanism—simulation—lies behind the entire range of phenomena.

At the same time, though, it is clear that words are special, since they reflect historically shaped, culturally shared conventions for conceptual coordination. Words are not merely coded instructions for covertly re-enacting certain kinds of sensorimotor states; they are language-specific instructions for running such re-enactments. Section 4 barely scratched the surface of the tremendous diversity that linguists specializing in semantic typology have already found in the multifaceted domain of the lexicalization of visual experience. The fact that so much cross-linguistic diversity exists cannot be over-emphasized because, as Levinson (2003a, p. 29) observes, “that's the fundamentally interesting thing about language from a comparative point of view. We are the only known species whose communication system is profoundly variable in both form and content…. So we can’t have the same kind of theory for human communication that we have for bee or even monkey communication….” Tomasello (2003, p. 1) makes essentially the same point when he states that one of the most bizarre traits of Homo sapiens is that “whereas the individuals of all nonhuman species can communicate effectively with all of their conspecifics, human beings can communicate effectively only with other persons who have grown up in the same linguistic community—typically, in the same geographic region.” The significance of this basic fact has yet to be fully grasped by and absorbed into the branch of cognitive neuroscience that concentrates on the representation, organization, and processing of
conceptual knowledge. Hence many intriguing questions await investigation. What is the relation between cross-linguistic variation in body part terms (Brown, 2005a,b; Enfield et al., 2006) and the mapping of the extrastriate body area (Peelen & Downing, 2007; Kemmerer & Tranel, 2008)? What is the relation between cross-linguistic variation in landscape terms (Burenhult & Levinson, in press) and the mapping of the parahippocampal place area (Epstein & Kanwisher, 1998; Aziz-Zadeh et al., in press)? What is the relation between cross-linguistic variation in verbs of “cutting and breaking” (Majid et al., 2007) and the mapping of the posterolateral temporal cortex (Blake & Shiffrar, 2007; Kemmerer et al., 2008)? My hope is that this chapter will help inspire more scientists to explore these and countless other questions about how words capture visual experience.

REFERENCES


Pirog, K.A., Aslin, R.N., & Tannenhaus, M.T. (submitted). Gradient activation of human V5 by spoken words and phonological competitors that have newly acquired semantic properties.


Figure 1. Summary of findings suggesting that color information is stored in the ventral temporal lobes, anterior to the regions that mediate passive color perception. White circles show the location of regions active during passive color perception (Corbetta et al., 1990; Sakai et al., 1995; Zeki et al., 1991); black circles show the location of regions in the ventral temporal lobes active when subjects generated color words (Martin et al., 1995, two studies; Wiggs et al., 1999; Chao & Martin, 1999); the gray circle on the left hemisphere shows the location of the region active when color-word synesthetes experienced color imagery (Paulesu et al., 1995); the gray circle on the right hemisphere shows the location of the region active in normal subjects during a color imagery task (Howard et al., 1998). (This figure is adapted from Martin et al., 2000.)
Figure 2. Distinct activation patterns for animal (including human) and tool concepts as revealed by functional neuroimaging studies employing both pictures and printed words as stimuli. The left image shows parallel but segregated lateral (red) and medial (blue) sectors of the fusiform gyrus that may contain assemblies of conjunctive neurons for capturing the complex shape patterns of the two categories of objects. The right image shows parallel but segregated superior temporal sulcus (red) and middle temporal gyrus (blue) regions that may contain assemblies of conjunctive neurons for capturing the complex motion patterns of the two categories of objects. The right image also shows ventral premotor and intraparietal sulcus regions that may contribute to visuomotor knowledge of tools. (This figure was kindly provided by Alex Martin.)
Figure 3. Activation patterns for five classes of verbs—Running, Speaking, Hitting, Cutting, and Change of State—in the posterolateral temporal cortex (PLTC). A: Activations rendered on the left hemisphere of an inflated brain, with the PLTC enclosed in a red box. Yellow patches signify areas of overlapping activation for two or more verb classes. Other colored patches signify areas of activation unique to particular verb classes, according to the color key in C. B: Enlargement of the red box in A. C: Activations for each separate verb class in the territory of the PLTC indicated by the red box in A and B.
**Figure 4.** Results from a lesion study in which 78 brain-damaged subjects performed the Matching Test, which assesses knowledge of English locative prepositions. For each test item (n=50), the subject is shown three pictures of objects in various spatial relationships, and is asked which picture best represents the meaning of a particular preposition. For example, in one item the preposition is *in* and the three pictures show (1) one window above another window on the outside of a house, (2) eggs in a carton, and (3) a boy on a swing. The figure shows the subtraction of lesion overlaps for 15 unimpaired subjects from the lesion overlaps for 15 impaired subjects. The color bar indicates the number of lesions in the overlap difference (the difference reached as high as 7 in the red-coded zone for “more impaired”). The top panel shows a lateral view of the left hemisphere and reveals that, relative to unimpaired subjects, impaired subjects more often had lesions in the left frontal operculum extending posteriorly into the left inferior parietal lobule, specifically the supramarginal gyrus. The vertical white lines in the top panel denote the planes of the coronal sections depicted in the bottom three panels. The leftmost line, and corresponding leftmost bottom panel, indicates the plane of the frontal operculum, and the rightmost line, and corresponding rightmost bottom panel, indicates the plane of the supramarginal gyrus. (Adapted from Tranel & Kemmerer, 2004.)