Perception, 2012, volume 41, pages 1128–1152

doi:10.1068/p7271

Language as a source of evidence for theories of spatial representation

Ray Jackendoff
Center for Cognitive Studies, Tufts University, 115 Miner Hall, Medford, MA 02155, USA;
e-mail: ray.jackendoff@tufts.edu
Received 21 April 2012, in revised form 17 July 2012

Abstract. David Marr's metatheory emphasized the importance of what he called the computational level of description—an analysis of the task the visual system performs. In the present article I argue that this task should be conceived of not just as object recognition but as *spatial understanding*, and that the mental representations responsible for spatial understanding are not exclusively visual in nature. In particular, a theory of the visual system must interact with a theory of the language faculty to explain how we talk about what we see—and how we see all the things we talk about as though they are part of the perceived world. An examination of spatial language both raises the bar for theories of vision and provides important hints for how spatial understanding is structured.

Keywords: spatial representation, language–vision interface

1 Metatheory and theory

I am grateful for this opportunity to acknowledge my debt to David Marr's *Vision*, which has had a deep influence on my work in linguistics. Important evidence for vision research has emerged from the connections I have been able to make between Marr's work and my own (Jackendoff 1987a, 1991, 1996a, 1996b; Landau and Jackendoff 1993), as well as from many further investigations by linguists and psycholinguists. The present article reviews some of these considerations from the perspective of a linguist concerned with the character of cognition.

One thing that distinguishes Marr's approach to vision from nearly everything else in cognitive science is his attention to metatheory: what a theory of vision (or any other cognitive capacity) should be like. As hardly needs recounting, Marr argues that a complete theory must include three types of description. The *computational* theory is a formal account of the mental representations computed by the mental faculty under examination, so to speak a task analysis of the problem the faculty solves. The *algorithmic* theory is a formal account of how these representations are computed in real time by the brain; the *implementation* theory is an account of how such computations are actually carried out by the neurons.

Marr makes it clear that he considers an adequate computational theory to be a prerequisite for developing algorithmic and implementational theories. Again and again, *Vision* critiques algorithms in the literature that fail because they are based on the wrong assumptions of what is computed—that is, they are designed for the wrong task. At the same time, he acknowledges that the influence between levels goes both ways: algorithmic and implementational considerations can obviously bear on the choice of computational theory as well (eg Marr 1982: pp 211, 214–215).

Within this metatheoretical framework, Marr develops a partial theory of the visual faculty. Here I will concentrate on his computational theory, whose four major levels of mental representation are the retinal image, the primal sketch, the 2½D sketch, and the 3D model. Marr stresses the importance of determining exactly what information is encoded at each level, as well as what formal relations obtain among these levels: how distinctions on one level correlate with or help determine distinctions on another. For instance, Marr treats the computation of stereopsis as a relation between, on one hand, disparities in the primal sketches delivered by the two eyes and, on the other hand, the depth dimension in the 2½D sketch. I will call these correlations "interface relations" or "interface computations".

Theories of the individual levels and theories of the interfaces are interwoven. For instance, one important motivation for the 2½D sketch-representation is that several distinct interface computations, based on different characteristics of the primal sketch, converge on the same coding of visible surfaces and depth. Thus, one role of any particular level of representation is to bring together information derived by disparate interface computations into a coherent whole. This, in turn, can serve as input to multiple interface computations relating this level to further levels of representation, without regard to which source(s) it came from. For instance, as far as reaching and navigation are concerned, depth is depth is depth, no matter how it is computed.

If I may editorialize for a moment: From the vantage point of a linguist, I have the sense that Marr's work, though widely cited, has not led to a robust tradition of research in the vision community (and my friends in vision confirm this sense). There appear to be at least three reasons for this. One is that there were errors in the theory. The 2½D sketch and 3D model as Marr proposed them could not be shown to do the things that they are supposed to do; it is very difficult to compute a 3D model from a 2½D sketch; and so on. These issues could have been addressed by offering equally detailed computational theories and showing that they better met Marr's (or someone else's) general criteria of adequacy. But this seems not to have occurred.

A second reason for the virtual abandonment of Marr's quest for a computational theory—and therefore for the abandonment of his metatheory—is that very soon after the publication of his book, there was widespread rejection of the notion of symbolic mental representation altogether, coming especially from the hugely influential connectionist tradition and subsequent statistical approaches to cognition and learning. Moreover, shortly after the onset of connectionism came the explosion in brain imaging, which is concerned with locating computation in the brain, but not with what exactly the computation is *doing* (see eg DiCarlo et al 2012). This is the antithesis of Marr's approach. For instance, it is indeed an important discovery that there is a particular brain area involved in recognizing faces. But this discovery does not enlighten us as to *how* the brain recognizes faces: how it codes them in memory, how the visual system arrives at this coding, and how a viewed face is compared to a remembered one. Similarly, it is an important discovery that there are such things as mirror neurons. But their existence and location—and even what stimuli they are tuned to—tell us nothing about how the brain codes viewed actions and performed actions, and how these are compared in working memory.

Which leads to a third, more general reason, why Marr's approach fell so rapidly into neglect. Marr's call for an explicit account of what the brain does—a sufficiently general computational theory—is simply not part of the received tradition in psychology, neuroscience, and artificial intelligence. More commonly, a rough intuitive description of a limited domain is followed immediately by details of proposed implementation, a practice that Marr deplored. For instance, in the two decades after Marr's death there developed a substantial literature on the question of whether object recognition is "viewpoint-independent", like the 3D model, or "viewpoint-dependent", like the 2½D sketch (eg Edelman and Weinshall 1989, 1991; Poggio and Edelman 1990; Edelman and Poggio 1991; Tarr 1995; Tarr and Bülthoff 1998; Ullman 1998; Riesenhuber and Poggio 1999, 2000; Rolls 2011). But none of these makes a detailed proposal about the way objects *in general* are coded in long-term and/or working memory, one that can be compared with Marr's proposed levels (whatever their faults). Even when explicit computational models are proposed, they are typically tailored to limited classes of objects, with not too much concern toward showing how they scale up to the general case.

It is important to see here that the metatheory and the computational theory are separable issues. The computational theory may be wrong in many respects. But the metatheory—the codification of what the theory is trying to do—strikes me as unquestionable. In particular, I believe, with Marr, that a satisfactory computational theory is an essential factor in developing suitably general algorithmic and implementation theories, and deserves serious attention.

I will be concerned here with an informal computational theory—a specification of what sorts of problems the visual system must ultimately solve. In particular, I want to revisit and amplify the sorts of arguments Marr made for the existence of something like the 3D model level. Section 2 presents considerations from nonlinguistic spatial understanding. After an interlude on linguistic theory in section 3, sections 4–6 present considerations from the connection of language with the visual system, such that we can talk about what we see—and see things that we talk about.

2 Nonlinguistic constraints on spatial structure

A substantial segment of the post-Marr vision literature (eg references above) is concerned primarily with *object recognition*. The task is to match a visual input with a known object, and the issues are how the known object is encoded and how the input is matched to it. As mentioned above, one alternative is typically a "viewpoint-dependent" representation, in which objects are coded in terms of a number of representative 2D views. Recognition is taken to be achieved by developing a 2D representation of the object being viewed and interpolating it among the stored 2D views. In many cases the 2D representation is extremely sparse: a coding of vertices or of skeletal line segments representing luminance boundaries.

The other alternative typically cited is a "viewpoint-independent" representation, which codes objects volumetrically, as in Marr and Nishihara 1978; Biederman 1987; and Pentland 1986. In these approaches, recognition is taken to be achieved either (a) by developing a three-dimensional representation of the object being viewed and comparing it to the stored 3D representation, or (b) by developing a 2D representation of the object being viewed, computing a 2D projection of the 3D stored item, and comparing the two 2D representations. Most accounts argue, on experimental and/or computational grounds, that the viewpoint-dependent theory is correct; a few (eg Tarr 1995) advocate a hybrid theory, in which a viewpoint-independent representation of some minimal sort supplements the work of the viewpoint-dependent representations.

At the risk of sounding a bit romantic, I want to suggest that this goal for the visual system is far too limited: the goal should be not just *object recognition* but ultimately *spatial understanding*. For my taste, "object recognition" couched in terms of 2D matching to stored 2D templates is rather like "sentence understanding" where what is achieved is merely a syntactic parse. This is not to say syntactic parses are unimportant or uncomplicated. But they are far from what is needed to determine the message a sentence is intended to convey, which is after all the point of using language.

Similarly for object recognition. To illustrate the point, let us begin with the phenomenology. On the basis of seeing the world, we *understand* it, not as a collection of 2D images pasted together, but as a three-dimensional array of two- and three-dimensional objects, some of which may be in motion. As I write this, I see my hands manipulating the keyboard of my laptop, and I see text appearing on the screen. Behind the laptop I see the top of a pile of journals—and, even though I can't see the bottom of the pile, I know the top journals aren't just floating in space. Beyond the journals is a bulletin board; to the right are a printer and a mess of tax documents; to the left is a clock, a tape dispenser, a picture of my grandchildren, and a lamp made out of an old clarinet. And so on. As for the couch behind me with a skylight above it and a little workbench next to it, I don't *see* them, but I *know* they're there.

I understand all this volumetrically; the objects all have backs that I can't see. Somehow the visual system is responsible for coming up with all the stuff I see. Moreover, ultimately, by building up what we might call spatial memory, it is responsible for coming up with what I *don't* see at the moment but *know* is there anyway.

Another source of representations suitable for object recognition—and more generally for spatial understanding—is the haptic modality. The sense of touch can be used to determine the shapes and arrangements of objects, turning them over in the hands or running the hands over them. One can even use the tongue to tell the shapes of things like nuts and pills, by rolling them around in the mouth (Topolinski and Türk-Pereira 2012). People can correlate haptic shape identification with visual shape identification, though they are not perfect, especially when shapes get complex and subtle (several chapters in Eilan et al 1993; Ernst and Banks 2004; Norman et al 2004). Moreover, Landau and Gleitman (1985) show that blind children develop a notion of spatial layout altogether comparable to that of sighted individuals, despite the limitations of the haptic faculty.

The mental representations responsible for haptic perception are based on touch, pressure, and proprioception. It is evidently possible to perform object recognition on the basis of some small number of such "haptic views". I know of no discussion of such representations in the literature, but it seems unlikely that they are 2D images, fully compatible with visual images. As Rock 1983 (pp 69–71) observes, there is no way to correlate or compare these with 2D visual images unless visual and haptic processing converge on some common representation, one that characterizes the world in terms of *objects* and the spatial relations among them. A volumetric representation of some sort seems the obvious candidate for such a common representation.

In addition, the environment includes one object whose shape is absolutely crucial for guiding action: one's own body. It is crucial to integrate information about one's own body configuration and motion with one's understanding of the spatial environment, and indeed something like Marr's 3D model of the human body would be suitable for the task. The inputs in this case are the various systems of proprioception, including at least touch and pressure sensors (again), muscle stretch receptors, and the vestibular system (Lackner and Dizio 2000).

What does a volumetric representation have to encode? Marr's 3D model is a good starting point. Here are some of its important virtues.

- As a volumetric representation, it captures our understanding of objects as three-dimensional and unitary (as opposed to a collection of 2D views)
- It directly captures object constancy regardless of viewer's perspective
- As a volumetric representation, the 3D model is a suitable output representation for the haptic faculty and for proprioception as well as for vision
- It represents objects in terms of decomposition into parts. The decomposition is hierarchical, so that a part may itself be composed of parts, and so on
- The hierarchical structure provides a range of analyses from coarse to fine, depending how much decomposition is taken into account. For instance, to determine the distance or motion of an object as a whole, only its coarsest analysis need be considered. But the position of the fingers is best characterized at a finer layer of structure, namely in terms of their relation to the hand
- Differences among individual objects of the same category (eg different dogs) will tend to be localized at finer layers of description; differences among categories of objects (eg dogs vs giraffes) will tend to be localized at relatively coarser layers of description
- An object's affordances for changes of shape (eg the possible configurations of fingers in a hand) can be characterized in terms of degrees of freedom at joints between parts [1]

[1] Note though that this treatment doesn't extend to the shape affordances of things like snakes and ocean waves.

- An object's internal motion (eg walking) can be characterized in terms of change in angle of attachment of parts. The difference between different modes of internal motion (eg running vs walking) can be characterized in terms of differences in their motion schemas—how the limbs move with respect to each other over time (Marr and Vaina 1982)
- Each layer of description is characterized in terms of a principal axis that constitutes the "skeleton" of its principal part. "Pipe-cleaner animals" that replace generalized cones by just their axes can often be recognized as depictions of those animals
- Marr also allows for secondary axes. For example, a human figure has not only the principal vertical axis, but also a front-to-back axis that can be used to align body parts such as the face and the navel with each other

On the other hand, the 3D model, as Marr conceived it, also has substantial drawbacks. Some of them concern the interface between the 3D model and lower-level representations.

- As emphasized in the post-Marr vision literature, a 3D model is very difficult to compute from 2D or 2½D representations as currently understood[(2)]
- The organization of the 3D model is such that it is difficult to encode a partly underspecified representation, such as what one would be able to derive on the basis of only a single view of a novel object

These are the sorts of drawbacks most noted by those who object to volumetric representations. But other drawbacks concern the descriptive adequacy of the 3D model itself.

- Generalized cones, the main primitives of the 3D model, are patently not sufficient to characterize all the sorts of objects we can understand, such as tables and chairs, sheets of paper, laptops, ponds and puddles. More primitives are necessary (Some of these are available in Pentland's (1986) treatment of parts in terms of superquadrics.)
- The 3D model does not capture the understanding of the interior structure of objects, such as the hollowness of a balloon, a bubble, or a box—nor interior parts of an object such as the yolk of a hard-boiled egg. It also does not capture the understanding of "negative objects" such as holes and caves, which also have three-dimensional shape and potentially even axes. However, these gaps could be remedied by adding features such as hollowness and "negativity" to generalized cones or to other primitive shapes (Landau and Jackendoff 1993); some of them might be captured by Pentland's (1986) Boolean NOT-function (See section 4.)
- The 3D model as such offers no account of the overall spatial structure of the visual field as a whole, such as my understanding of the objects on my desk and elsewhere in my study. In particular, it does not characterize the spatial relations between pairs and *n*-tuples of objects. Now it may well be that spatial relations among objects are the domain of a different representation, say a representation generated by the dorsal stream. But in that case, (a) these other representations must be accessible to haptic and proprioceptive perception as well, and (b) there has to be an interface between the representations of individual objects and the representation of spatial arrays (Landau and Jackendoff 1993) (See sections 5 and 6.)
- Ambitiously, one might conceive of spatial understanding going so far as to incorporate the forces exerted by and on objects, ie notions of naive physics. For example, one can make an approximate visual judgment of whether a tower of blocks will stand up or fall down. The need to understand such forces is particularly striking in the case of proprioception, where one both experiences external forces and exerts forces oneself.

[(2)] One might wonder if some of the difficulty human subjects have in identifying the "paper-clip objects" and Shepard–Metzler figures, used in so many experimental paradigms on object recognition, is that these objects do not lend themselves to hierarchical decomposition, nor do they have clear axes. The difficulty might parallel that of learning strings of nonsense syllables as opposed to sentences.

> It is not clear where such judgments belong in the ecology of cognition, but the visual system provides crucial input. Such information is not part of the 3D model

These problems cannot be addressed by eliminating object-centered volumetric representations.

The issue, at least as I see it, is what a psychologically adequate representation of spatial understanding should be like. In particular, I am conceiving of spatial understanding not as a strictly *visual* sort of mental representation, but rather as amodal or multimodal, coding shape information that originates from multiple modalities. In addition, each modality contributes its own particular sorts of information: vision contributes color but hapsis does not; hapsis contributes temperature but vision does not; proprioception conveys force rather directly, but vision does so only indirectly at best. In addition, auditory localization conveys object location but not object shape (except by inference); presumably bats' sonar is yet another route to object recognition and localization.

A pause to negotiate terminology: The computations subsumed under Marr's three levels correspond more or less to what is generally called (eg by Cavanagh 2011) "low-level", "mid-level", and "high-level" vision, as in figure 1. (The bidirectional arrow represents the possibility of top–down influence.)

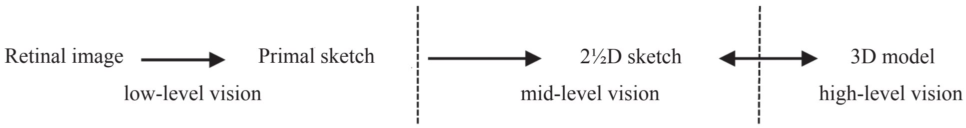

Figure 1. The architecture of Marr's theory.

The whole system is usually called "visual perception"; sometimes the retinal image is called "sensation" and sometimes the 3D model or its counterpart is called "visual cognition" (as Cavanagh does). The connection of this system to "general cognition" or "higher cognition" is left an open question. I'm now suggesting, though, that the 3D model or its suitably augmented volumetric replacement is an amodal or multimodal representation, and I will rename it "spatial structure", as it is no longer purely visual. The visual system proper then ends with the interface mapping between mid-level vision and spatial understanding. Whether one wants to call spatial structure "perceptual" or "cognitive" is only a question of policy, but I will call it cognitive, in part for reasons to become clearer in section 4. Terminology aside, the real issues for a computational theory of this level are: what does it encode, how are its contents mapped to and from the various perceptual modalities that interface with it, and how does it interact with other forms of cognition?

3 Connections of Marr's metatheory and theory to linguistic theory

In preparation for laying out the connections between vision and language, it is useful to look at the connections between Marr's enterprise and linguistic theory. I will begin with the metatheory, then move on to theoretical parallels and connections.

Marr recognizes the kinship between his enterprise and generative linguistics. Generative linguistics too is an attempt to characterize structures computed by the brain, in this case the structures involved in producing and comprehending sentences. What Chomsky (1965) calls a theory of "competence" or "knowledge of language" corresponds to Marr's computational theory. Like Marr, Chomsky insists that it is crucial to characterize these structures, so that theories of language processing—"performance" in Chomsky's sense and an algorithmic theory in Marr's sense—have a firm basis in what representations the processor has produced. Similarly, theories of language acquisition must be built on a theory of what is acquired.

Unfortunately, the models of linguistic knowledge that have emerged from mainstream generative grammar (eg Chomsky 1965, 1981, 1995) do not lend themselves to being easily adapted to theories of language processing; in Marr's terms, there is not a smooth transition from the computational to the algorithmic theory. There are at least three reasons for this. One is that these models have no theory of meaning to which the language faculty is connected. For instance, in recent formulations (eg Chomsky 1995; Hauser et al 2002[3]), meaning has simply been characterized, without elaboration, as the "conceptual–intentional interface". Moreover, the most influential tradition in linguistic semantics, so-called formal semantics (eg Heim and Kratzer 1998), grows out of the tradition of philosophy of language, which takes language to refer directly to "the world"; this tradition makes no contact whatsoever with psychological issues. As a result, it turns out to be of questionable use in developing theories of language in the brain and of language processing.

A second reason that mainstream models of language do not translate well into processing theories is that they are stated in terms of algorithmic generation of sentences. They start with an initial symbol *S* (in earlier formulations) or with a numeration of items to be combined (in more recent "minimalist" formulations), and they build from this a core of syntactic structure. From the syntactic structure, the grammar derives the sound and meaning of sentences.

This procedure bears little resemblance to what has to go on in language processing. Language production has to start with an intended meaning, and from that develop a syntactic and ultimately a phonological expression that can be uttered. Language comprehension has to go in the other direction, starting with a heard string of sounds and from that developing a meaning. In neither case is syntax the starting point. Linguists often respond to this disconnect by saying their algorithmic derivations are "metaphorical" in some sense, and that one should not expect any connection between theories of competence and performance. Not surprisingly, psychologists have often responded by rejecting generative grammar altogether (Ferreira 2005; Poeppel and Embick 2005; Jackendoff 2007a).

A third problem with mainstream models of language is that they do not distinguish what has to be stored in memory from what is built online. Algorithmic derivations are ascribed indiscriminately to material that is indisputably stored (such as idioms and morphologically complex but only partially regular words) as well as to material that is clearly computed online, such as the present sentence (Halle and Marantz 1993; Marantz 1997; Hale and Keyser 2002).

The upshot of these failings is that contemporary mainstream generative grammar has in many respects become irrelevant to studies of language processing and language acquisition. As a result, attempts to draw parallels between Marr's metatheory and that of generative linguistics have not been very useful. Mainstream linguists, focusing almost exclusively on the computational theory, find themselves largely out of touch with psychologists, who, as mentioned above, on the whole place little stock in developing an explicit theory of the mental representations involved in cognitive activities, and instead stress algorithmic and implementational-level accounts.

These three failings of mainstream generative theory are ameliorated in the central tenets of my own approach to the structure of language, the Parallel Architecture (Jackendoff 1987b, 1997, 2002, 2010; Culicover and Jackendoff 2005). This approach posits three independent representations, each with its own generative system: phonological (sound) structure, syntactic (grammatical) structure, and conceptual (meaning) structure. These structures are linked by interface components, which correlate representations at the three levels. In addition, phonological structure is linked via an interface to low-level auditory representations; this interface provides input to language comprehension. Another interface links phonological structure to instructions to the vocal tract, which makes overt speech possible. Figure 2 sketches the overall configuration of the theory. It should be noted that the arrangement of components in the Parallel Architecture resembles the layout of Marr's theory in figure 1: a collection of levels of representation linked by interface components.

[3] For a critique of Hauser et al 2002, see Jackendoff and Pinker 2005; Pinker and Jackendoff 2005.

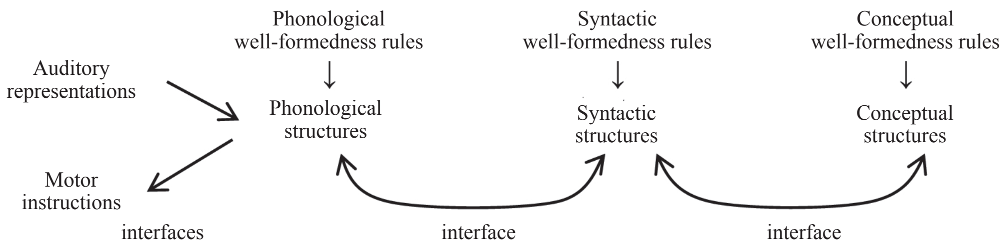

Figure 2. The Parallel Architecture.

The structure of a sentence is a triple of structures, with corresponding parts coindexed (or bound, in the neuroscience sense). Here is the structure of *The cat died*.

<u>Phonology:</u> $[\text{ðə}_1\ [\text{kæt}_2]]\ [\text{day}_3 + \text{d}_4]$
<u>Syntax:</u> $[_{NP}\ \text{Det}_1\ \text{N}_2]\ [_{VP}\ \text{V}_3 + \text{past}_4]$ (1)
<u>Semantics:</u> $[\text{PAST}_4\ [\text{DIE}_3\ ([\text{CAT}_2;\ \text{DEF}_1])]]$

Individual words are treated as small interface rules, allowing linking between parts of the three structures. Here are the structures of *cat* and *die*, for instance.

<u>Phonology:</u> $[\text{kæt}_2]$ <u>Phonology:</u> $[\text{day}_3]$
<u>Syntax:</u> N_2 <u>Syntax:</u> V_3 (2)
<u>Semantics:</u> $[\text{CAT}_2]$ <u>Semantics:</u> $[\text{DIE}\ (X)]_3$

An important component of the architecture is a detailed theory of meaning, Conceptual Semantics, developed in cognizance of what is known about conceptualization in humans and nonlinguistic organisms (Jackendoff 1983, 1990, 2002, 2007b; Pinker 2007). It is also deeply concerned with the connection of linguistic meaning to visual understanding, to which I return below.

As in several other non-mainstream frameworks, such as Head-Driven Phrase Structure Grammar (Pollard and Sag 1994), Lexical-Functional Grammar (Bresnan 2001), Construction Grammar (Goldberg 1995), Autolexical Syntax (Sadock 1991), and others, linguistic structures are characterized not in terms of algorithms that generate sentences, but rather in terms of constraints on the individual levels of representation and on the interfaces among them. This enables the competence theory to be plugged directly into a theory of processing: the competence theory says what structures are available, while the processing theory says how the constraints are used to develop these structures over time (Jackendoff 2002, chapter 7, 2007c; Sag and Wasow 2011). From this standpoint, experimental considerations can be brought to bear on the choice among competing accounts of competence (Piñango et al 1999; Kuperberg et al 2010; Wittenberg et al, submitted); the theory is not entirely driven from the computational level.

Finally, the theory of what is stored in memory (the "lexicon") says that "rules of grammar" are stored as schematic pieces of linguistic structure that can be used to assemble words into more complex structures. As a consequence, there is no principled formal distinction between words, rules, and larger stored assemblages such as idioms. This view of the lexicon is shared with many of the other constraint-based approaches.

Marr's counterpart of the linguistic lexicon is the "library" of 3D models. An alternative is that, in parallel to the linguistic lexicon, the "visual vocabulary" might consist of precomputed linkings of known 3D models with known 2(½)D images. This would in part mitigate the difficulty of constantly computing all these linkings online, one of the important drawbacks of Marr's theory. It would also comport with the experimental evidence that one does store 2D representations (see references above), suggesting a hybrid theory of the sort proposed by Tarr (1995).

4 The relation between spatial understanding and language

A basic question in the theory of language was posed bluntly by John Macnamara (1978) [and Miller and Johnson-Laird (1976) were on a similar mission].

How do we talk about what we see?

Macnamara's answer was that there must be a "translation" between mental representations proprietary to the visual faculty and those proprietary to the language faculty. In fact, the translation has to be two-way: we not only talk about what we see, we also use others' talk to direct our attention to the visual field, with utterances like those in expressions (3).

Look at that eagle! (3a)

[*Experimenter to subject*:] Fixate on the cross in the center of the screen. (3b)

[*"Visual world" paradigm* (Tanenhaus et al 1995):] Put the apple on the towel in the cup. (in which unconscious eye movements are directed by language) (3c)

How do we flesh out this "translation" between the linguistic and visual faculties? My hypothesis is that there is an interface between conceptual structure and what I've outlined as spatial structure in section 2. Note how this connection provides a natural link between figures 1 and 2.(4)

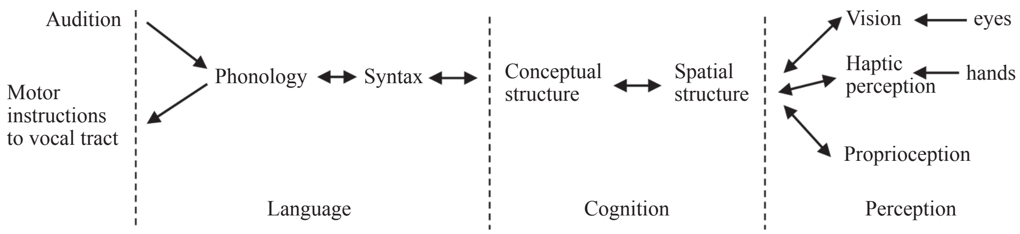

Figure 3. The connection of language, cognition, and perception.

The new interface permits information derived through vision and encoded in spatial structure to be reformatted in terms of conceptual structure, which in turn is a suitable format for encoding into language. Conversely, understanding a heard sentence consists in developing a conceptual structure derived via phonology and syntax; this conceptual structure can then be reformatted as a spatial structure and thereby direct attention to relevant parts of the visual field.

So far I have characterized this conceptual–spatial interface as a "translation" between two formats of representation, which "reformats" one into the other. Indeed, the two formats do share certain constructs such as "physical object". But just as each perceptual modality encodes only partly overlapping dimensions (eg color in vision and temperature in hapsis), so do conceptual structure and spatial structure each encode types of information not accessible to the other. For instance, the concepts associated with words such as *friendship*, *truth*, *justice*, *value*, *purpose*, *cousin*, *income*, *despite*, and *however* have nothing at all to do with visual/spatial cognition. At the same time, spatial structure is responsible for encoding precise details of shape, such as the precise difference between a duck's neck and a goose's neck, or between my face and yours—things that are not naturally coded in conceptual structure. This is why it is often very hard to convey such details in language—why a picture is worth ten thousand words.

The relation between the two representations is more than a link between language and vision: it also is necessary for visual memory independently of language. Consider the type–token distinction. The geometric/topological format of spatial structure has no provision for

(4) And notice how such a connection is far less natural in the mainstream architecture for language, in which everything is derived from syntax. I take this as important evidence in favor of the Parallel Architecture.

encoding the distinction between a memory of, for instance, what this particular cat looks like (a token) and what cats look like in general (a type). The temptation is to say that the difference is in the specificity of the description—my cat Peanut's representation in memory is precise but my representation of cats in general is vague or underspecified. But this cannot be right. On the one hand, some categories are geometrically very precise, say all the forks in a matched set. And on the other, one can perceive token objects whose visual description is very imprecise, say a moving animal in the bushes viewed only briefly or in dim light. Hence difference of precision in spatial structure cannot be the factor that decides between a token and a type.

On the other hand, the distinction between tokens and types is easily encoded in the algebraic feature–structure format of conceptual structure: it is a simple binary feature. Using this feature, we can treat the *spatial* representation of a particular fork to be identical with that for the category of identical forks in the set; however, in *conceptual* structure the two representations differ in the algebraic diacritic TOKEN vs TYPE. Thus object *identification* involves matching a percept to a spatial representation in memory that is linked to a TOKEN feature; object *categorization* matches a percept to a spatial representation linked to a TYPE feature. The percept, of course, is assigned the feature TOKEN: one is seeing a particular thing, not a category.[5]

In order for the distinction between object identification and object categorization to be possible, then, visual memory must include (or be linked to) elements of conceptual structure, in addition to spatial structure. Visual memory therefore begins to look more like a linguistic lexicon, which likewise links structures from disparate levels.

Not only does vision rely in part on conceptual structure, but the linguistic lexicon relies in part on spatial structure. It is often suggested that the meaning of a word for a physical object includes an imagistic component. For instance, the meaning of *dog* is taken to include an "image of a stereotypical dog" (eg Putnam 1975). However, as recognized as long ago as Bishop Berkeley, a fixed 2D visual image cannot serve such a function computationally: it is too specific. For instance, no single dog-image can serve as the representation for all dogs in all positions, seen from an arbitrary point of view.[6] On the other hand, a viewpoint-independent schema in spatial structure, which allows for a range of body proportions and a range of limb dispositions, can serve this function more adequately.

Moreover, as observed above, the distinction between the way ducks look and the way geese look is not naturally characterized in the algebraic terms of conceptual structure. For instance, a feature like [+long-neck] is obviously makeshift (and furthermore does not distinguish between a goose's neck and a giraffe's neck). However, this distinction *is* natural in the geometric format of spatial structure. The conclusion is that words for visually perceivable objects link not just stored phonology, syntax, and conceptual structure, as in expression (2) above, but also a stored spatial structure that can be used for visual—or haptic—identification. The link to this spatial structure is necessary for being able to see a dog and label it with the word *dog*.

For a different sort of case, consider the understanding of the verb *sit*, such that we can visually identify someone as "sitting" or not. The visual encoding of sitting has to contain a schema of a person (but no particular person) situated in a particular position on a support (but no particular support). However, taken alone, such a visual schema might equally be taken to designate the person or the support, rather than the person's posture and spatial relation to the support. The proper designation can be encoded by linking the visual schema

[5] This TOKEN feature in turn might be linked with the "index" or "FINST" (in the sense of Pylyshyn 2007) that individuates the percept as a distinct entity in lower-level vision.

[6] Multiple views of multiple exemplars, plus interpolation, as in the viewer-dependent theories of object recognition, might help, of course.

to a conceptual structure, as in figure 4, where the numbers label corresponding parts in the two structures. So the meaning of the word *sit* is distributed between conceptual structure and spatial structure.

Conceptual structure: $[_{\text{Situation}}\ \text{SIT}_3\ (\text{X}_1, \text{Y}_2)]$ (X_1 = sitter; Y_2 = seat; SIT_3 = sitting)

Spatial structure:

Figure 4. Conceptual structure and spatial structure of sitting.

The correlation of spatial and conceptual structures also has advantages for the visual identification of actions. Even nonlinguistic organisms can categorize classes of actions such as sitting, reaching, running, grasping, eating, and so on. How are these encoded? An encoding in terms of viewpoint-dependent templates would have at least two drawbacks. First, it could not distinguish the representation of an action from a visual representation of the character performing the action—is this a representation of *sitting*, or of a *sitter*? Second, many actions must be encoded in such a way as to establish an equivalence between an action performed by oneself (for which there is proprioceptive input) and one performed by someone else (for which there is not); furthermore, self-performed actions look different from actions by others. Nevertheless, such an equivalence is evidently observed in the tuning of mirror neurons, so we know monkeys can establish it. Moreover, the ability to imitate someone else's actions relies on establishing such an equivalence between visual observation and body control.

Both these drawbacks are addressed by a theory of action representation along the lines sketched in figure 4: a volumetric viewpoint-independent spatial structure linked with a conceptual structure. By using a highly schematic representation of the actor, the spatial structure can be applied to any actor. Because spatial structure can be informed by either vision or proprioception, one's own actions can be equated with observed actions by others. And because the schema is linked to conceptual structure, we know we are attending to an action rather than to the characters involved in it.

A final, very important case, concerns the notion of animacy. An object may be judged animate if it changes shape and/or position spontaneously and unpredictably, without any apparent external force acting on it. But animacy and the accompanying attributions of desire, knowledge, and volition are not a matter of an object's shape and position—they are only *cued* by shape and change of shape and position. Indeed, the cues can be rather sparse if they are of the right type, as in the well-known demonstrations of Heider and Simmel 1944 and Johansson 1975. The notion of animacy itself, as well as all its associated inferences, are a matter of conceptual structure encoding, where *animate/inanimate* is an algebraic feature distinction. So again, an understanding of the character of the perceived world is shared between spatial structure and conceptual structure.

An important aspect of this approach to cognition, then, is that it divides the work of cognition among two or more amodal representations, each of which is suitable for particular sorts of computations. This contrasts with the commonly held view (eg Fodor 1975; Baars 1988; Dehaene and Naccache 2001) that cognition involves a single format that is shared among faculties through a common "workspace". Here the sharing is done in part through interfaces that link various perceptual and cognitive representations to each other.(7)

5 Evidence from language for enriching spatial structure

As already hinted in the preceding section, we talk about physical objects and actions all the time, and we can identify them in the visual environment. This leads to the converse of Macnamara's question:

How do we see all the things we talk about?

Our use of language presumes a rich representational structure that can be linked to what we see. The rest of this article reviews some of this structure, following and elaborating on Miller and Johnson-Laird (1976) and Landau and Jackendoff (1993). The goal is to illustrate how language can serve as a source of evidence for the extraordinary richness of visual cognition, far beyond what is generally envisioned by vision researchers. The data to be discussed here, though fairly extensive and intricate, represent only the tip of an increasingly well-explored iceberg (eg Herskovits 1986; Vandeloise 1991; Brown 1994; Haviland 1994; Bierwisch 1996; Bowerman 1996; Landau 1996; Levelt 1996; Tversky 1996; Carlson-Radvansky and Logan 1997; Levinson 2003; Van der Zee and Slack 2003; Coventry and Garrod 2004; Evans and Chilton 2010). The challenge for vision research then is: how does the visual system come up with all this?

5.1 *Deictic anaphora*

First consider the use of deictic anaphora in a sentence like expression (4), whose interpretation depends on a mixture of linguistic and visual information.

Would you pick THAT up, please? [*pointing*] (4)

The hearer determines the reference of the deictic pronoun (or demonstrative) *THAT* by following the speaker's point, thereby discovering some relevantly salient object in the visual environment. Understanding the sentence (and therefore picking up the right thing) requires integrating information derived from language and information derived from vision, so that an appropriate action can be formulated.

In this case we have some idea of how the visual information is picked out: through the very complex and as yet not fully understood process of object perception. In particular, we can think of the deictic pronoun *that* as being linked to an object file in the sense of Kahneman et al 1992.

However, this is not the only kind of deictic anaphora in English. Expression (5) shows another.

Would you put your hat THERE, please? [*pointing*] (5)

In this case, the deictic anaphor invites the hearer to pick out a *location* in the visual environment. Locations are partly determined by objects, but they are not the *same* as objects.

(7) Another difference between the present account and the "workspace" theory concerns the level of representation that supports consciousness. Baars and Dehaene/Naccache, along with many others, claim that consciousness arises from the cognitive representations. Jackendoff (1987b, 2007b, 2012) makes the case that the form of consciousness most closely mirrors *perceptual* representations. For example, conscious experience, like perceptual representations, is modality-specific, while cognitive representations are amodal or multi-modal. In visual experience, one *experiences* the facing surface of objects, a function of perceptual representations; while one *infers* or *understands* that they have backs, a function of cognition. In viewing an illusion, one *experiences* it one way (perceptual) while *understanding* that it is some other way (cognitive).

We can see this from sentences where we replace the deictic anaphor by more fully fleshed out expressions:

Would you put your hat on the table/under the table/next to the table? (6)

The table itself is not the location—rather, the location is a region of space defined *in terms of* the table. Moreover, we can point to locations where there is no object at all:

There was a fly buzzing around right HERE. [*pointing to empty space*] (7a)

The chandelier should hang down to about HERE. [*pointing to empty space*] (7b)

For another such case, the deictic anaphor in expression (8) invites the hearer to pick out not a location, but a *path* or *trajectory* that the fly traversed. The fly is no longer in any of the locations along the trajectory.

The fly went THATAWAY! [*pointing*] (8)

Yet another type of deictic anaphora appears in expression (9).

Can you do THIS? [*demonstrating*] (9)

The grammatical structure *do this* directs the hearer to pick out an *action* in the visual environment. As discussed in the previous section, this action must be independent of who is performing it: the hearer is invited to evaluate his or her own ability to perform the same action as the speaker is currently performing. You can't perform someone else's action!

A related case is expression (10) below:

THAT had better never happen in MY house! [*pointing to, say, some kids smoking pot*] (10)

The hearer is again invited to pick out an action. What the speaker is condemning is not just the action by these particular individuals who are carrying it out, but any action of this kind by anyone at all, including the hearer.

Another action-related deictic anaphor appears in expression (11).

Can you walk like THIS? [*demonstrating, say, a Groucho Marx walk*] (11)

This sentence invites the hearer to pick out not just the action, but some distinctive character or *manner* of action. Manner of action might be encoded in the fine structure of Marr and Vaina's (1982) motion descriptions; another possible realization is Cavanagh's "sprites" (Cavanagh et al 2001).

A quite different domain is invoked by expression (12).

The fish that got away was THIS long. [*demonstrating*] (12)

Here, there need be no fish in the visual environment. Rather, the hearer is invited to pick out the *distance* between speaker's hands and apply that to the imagined fish. Finally, consider expression (13).

There were about THIS many people at the party last night. [*gesturing around room*] (13)

The individuals currently in the room may or may not have been at the party; their attendance is irrelevant. Rather, the hearer is invited to pick out the *numerosity* of the individuals in the room, considered as an *aggregate*, and apply that to the aggregate of remembered or imagined individuals at the remembered or imagined party. The sense of numerosity is presumably related to Dehaene's (1997) analog number sense.

Summing up: Deictic anaphora shows us that language talks about objects, locations, paths, actions, manners of action, distances, and numerosities, as though they are entities in the real world detectable through visual cognition. It therefore raises these challenges for a theory of visual understanding:

- How are these entities encoded in spatial structure?
- And how does the visual system derive them from the retinal image?

5.2 *Unusual objects*

Section 2 mentioned the need for spatial structure to encode certain aspects of objects that go beyond the 3D model's articulation of object shape in terms of generalized cones. It is worth enumerating some of the words whose meanings invoke these aspects of objects and object-like entities, just to see that these are far from exceptional cases.

The 3D model does take account of some sorts of objects whose identity comes from being a proper part of another object—things like *branches*, *legs*, *fingers*, *noses*, *handles*, *lids*, *bumps*, and *ridges*. The shape of such objects is determined in terms of the way they protrude from the surface of the object of which they form a part.

Another class of entities we talk about might be characterized as "negative parts". Such entities, instead of consisting of material tacked onto the surface of an object, might be thought of as the space created by "scooping out" material from the surface of an object. Examples are *holes*, *notches*, *slots*, *dents*, *grooves*, and, notably, *mouths* and *nostrils*, which are clearly treated as body parts. Although negative parts are not made of material, they have shape, size, and orientation, just like objects and object parts. [As mentioned earlier, Pentland's (1986) variant on the 3D model does allow for negative shapes carved out of larger shapes by means of the Boolean NOT-function.] Notice that we can compare the shape of a hole with the shape of an object, as when we notice that a square peg won't fit in a round hole. In fact, the notion of one object *fitting* into or against another often depends on a match between negative and positive shapes.

Doors and *windows* are interesting, because they can be construed either as an opening in a wall—a negative part—or as the "positive" part that closes this opening off. The first sense appears in a sentence like *He came in through the front door*; the second in a sentence like *He broke the door down*. Again, the notions of size and shape for these negative parts are related to those for ordinary objects; think of trying to maneuver a refrigerator through a doorway.

Related to "negative parts" are "negative objects": shaped spaces that are defined by material on the *outside* of their boundary rather than on the *inside*, as with normal objects. These objects include *caves*, *rooms*, *valleys*, *ditches*, and *wells*.

A huge number of objects we can identify are *hollow*—that is, they are conceptualized as a surface of a specified thickness that bounds an empty space. Such objects include *bubbles*, *balloons*, *drums*, *tubes*, *pipes*, *houses*, *violins*, and *clarinets*. A very important subclass of hollow objects is *containers*: *bags*, *boxes*, *pots*, *pouches*, *bottles*, *jars*, *suitcases*, *backpacks*, *trucks*, and so on. Hollow objects have an external shape, but once one passes through their surface, there is empty space—the interior of the object. The interior may or may not be visible, but it is definitely part of one's understanding of the object and how to use it.

Inside a hollow object there may be further unseen objects, whose presence must be encoded in spatial understanding. That's why we're surprised if we put something in our pocket and later it's gone. There are also interior parts of objects, which we know are there but can't see: *bones*, *muscles*, *brains*, *eggyolks*, *carburetors*, *motherboards*, and so on.

A final class of unusual entities are *aggregates*, made of multiple objects. Some aggregates have an inherent shape, for instance *piles*, *heaps*, *stacks*, and *rows* of things. Others have no inherent shape: *groups*, *herds*, *flocks*, and *swarms* of things. Visual understanding has to accommodate both these types of aggregates and distinguish how their shape is encoded.[8]

[8] A case that I find particularly puzzling is the category of *star*: how is this is to be encoded in such a way that the exact number of points is immaterial, as long as it is four or five or more?

5.3 *Manners of motion*

Section 4 alluded to the necessity of identifying manners of motion. English has numerous verbs that describe manner of motion, falling into a number of subcategories:

- Types of motion that can be attributed to any sort of object: *roll*, *slide*, *skid*, *float*, *fly*, *bounce*, *glide*
- Types of locomotion: *walk*, *run*, *swim*, *fly* (again), *slither*
- Further differentiated types of bipedal locomotion: *strut*, *waddle*, *mince*, *dance*, *swagger*, *stagger*, *shuffle*, *limp*, *lumber*, *jog*, *sprint*
- Types of motion in place (motion of object without changing overall position): *wiggle*, *shake*, *shudder*, *wave*, *flutter*, *twirl*, *rotate*
- Types of change of shape: *grow*, *shrink*, *elongate*, *widen*, *bend*, *straighten*, *twist*, *crumple*

Spatial structure must contain provision for encoding and identifying each of these types of motion.

5.4 *Parts and properties of objects defined by axes*

As mentioned in section 2, Marr makes provisions for additional axes of objects beyond the principal axis that defines a generalized cone. An interesting range of words depend for their meaning on axes of an object.

A first group denote what might be called "axial parts" of an object. For instance, the *top* of an object X is defined in terms of X's vertical axis: the top is the region of X's surface that is intersected by the upward end of X's vertical axis. The *bottom* is the region of X's surface intersected by the lower end of the vertical axis. If the object is of a familiar type, its vertical axis is often defined in terms of the object's canonical orientation, so that, for instance, the top of a hat is still the top if the hat is turned upside down.

If an object X has an asymmetric front-to-back horizontal axis (as does a human body, for instance), the *front* of X is the region of X's surface intersected by this axis in the frontward direction; the *back* is opposite. The front-to-back axis can be determined in terms of a canonical viewing position (eg the front of a house), or in terms of the object's canonical direction of normal motion (eg the front of a person or a car). (As will be seen below, *in front of* and *in back of* present other possibilities as well.) If an object does not have a front–back axis, for example a sphere or a flagpole, it also lacks a front and a back.

An interesting case is the *end* of an object. Rectangles and ellipses have ends, but squares and circles do not; a pole lying on the ground has ends, but a standing flagpole does not. From these and other examples we conclude that the end of an object X is a region of X's surface that is intersected by its longest horizontal axis. Since squares and circles lack a longest horizontal axis, and since a standing flagpole lacks a horizontal axis altogether, they also lack ends.

A second class of terms is involved in *measuring* objects along dimensions determined by their axes (Bierwisch and Lang 1989). The *height* of an object X is the distance from boundary to boundary along X's vertical axis. X's *length* is the distance from boundary to boundary along a linear axis that is significantly longer than other axes and that is preferably horizontal (standing flagpoles have height, not length). X's *width* is the distance from boundary to boundary along a secondary horizontal linear axis.

X's *thickness* is measured along a minor axis of X. There are two cases. If X is basically a line, say a string or a stripe, thickness is measured along X's second dimension. If X is basically a surface, say a board, thickness is measured along X's third dimension.

X's *depth* also has two cases. The first is the distance from the object's visible upper surface down to the bottom, as in the depth of a lake. The second is the distance from the entrance to X's interior to the other end of its interior, as in the depth of a cave. The two definitions coincide in the case of a vertically oriented hole such as a well.

All these measure terms have related pairs of adjectives: corresponding to *height* there are *high* and *low*; corresponding to *length* are *long* and *short*; corresponding to *width* are *wide* and *narrow*; corresponding to *depth* are *deep* and *shallow.*

The point of going through these examples is that, if asked, we can look at a novel object and point out its top, front, height, width, and so on. This implies that, as part of understanding the shapes of objects, the visual system constructs axis systems and distinguishes them by orientation and relative length, so that this information can be used to select the appropriate words. The challenge for visual theory, then, is how axis systems are encoded and how they are derived from the retinal image. In particular, it is not obvious how to encode axes in terms of impoverished 2D views of objects. For instance, as Jackendoff 1991 asks, what makes the *end* of a rope, the *end* of a road, the *end* of a car, and the *end* of a table alike? One cannot store a collection of 2D exemplars of ends and by interpolation understand ends of new objects (or at least it would seem). What makes ends alike is not how they look on the surface, but how their surfaces are related to a schematization of the objects in question, of the sort envisioned by Marr in the 3D model.[9]

6 Spatial relations among objects

Let us now return to the notion of location, briefly touched on in section 5.1. Consider expression (14).

The cat is on the table/under the table/next to the table/etc (14)

Each of these locates the cat, which I'll call the *figural object*, in a particular region of space determined by a relationship to the table, which I'll call the *reference object*. The preposition signals the particular spatial relationship of the figural object to the reference object.

Spatial relations can also be expressed by verbs:

There is a fly in my soup. = My soup <u>contains</u> a fly.
There is a wall around the city. = A wall <u>surrounds</u> the city.
The path goes across the road right here. = The path <u>crosses</u> the road right here. (15)
The road runs along the river. = The road <u>parallels</u> the river.

So the issue here is not so much grammatical form as how spatial relations are encoded in visual understanding, however they may be expressed in language.

Two questions raised early on by Leonard Talmy (1978, 1983), as well as by Miller and Johnson-Laird (1976), have continued to preoccupy linguists who study spatial language and psychologists studying language-directed navigation and scene perception (see references at the beginning of section 5):

- What is the range of spatial relations signaled by prepositions and other linguistic items?
- What does each spatial relation presume about the shape of its figural object and its reference object?

These issues should also be of interest to vision researchers, in that they provide some clues about how spatial relations are organized nonlinguistically as well. Again, since we talk about these relations as though they are out there in the world, our brains must be able to derive them from (or at least relate them to) the visual field and its partitioning into distinct objects. In particular, novel objects can serve as reference objects of spatial prepositions (*This strange thing is a fendle. What's under the fendle?*). Hence the regions of space defined by the combination of the spatial relation and the reference object cannot always be memorized: it must be possible to compute them online in response to novel visual inputs.

In order to work out the demands on the visual system, it is necessary to hack through a great deal of linguistic brush, because prepositions tend to be highly polysemous. For instance,

[9] Consider also the end of a speech, whose axis is in the temporal dimension: it needn't *look* like *any*thing.

into probably designates different spatial relations in *run into the room* and *run into the wall*; *over* probably designates different relations in *somewhere over the rainbow* and *turn the box over*. *Up* and *down* are usually antonyms, but, in certain circumstances, *up the street* and *down the street* can be synonymous. Furthermore, prepositions take part in numerous conventionalized collocations. For instance, whether one is *waiting in line* or *waiting on line* is a matter of dialect, not of one's spatial relation to the people ahead and behind. Despite this sort of complexity in the uses of spatial expressions, some overall trends are clear.

6.1 *Some simple relations*

The preposition *in* picks out a region within the boundary of its reference object. The reference object can have a two-dimensional interior, as in *the dots in the circle*; or it can have a three-dimensional interior, as in *the beetle in the box*. The use of *in* also extends to some cases in which the figural object is not entirely surrounded by the reference object, eg *the knife in the cheese*, where only the end of the blade needs to penetrate the interior (Herskovits 1986; Coventry and Garrod 2004).

In English, the preposition *on* picks out a region in contact with the surface of its reference object. In the stereotypical case, the figural object is also vertically above the reference object and supported by it, as in expression (16a). However, we can also use expression (16b), where there is support but not vertical alignment, and expression (16c), where there is neither support nor the proper vertical alignment, and only contact is invoked.

The book is on the table. (16a)

The picture is on the wall. (16b)

The fly is on the ceiling. (16c)

In German, these uses are distinguished: *auf* is used for the counterpart of (16a) and *an* for (16b and 16c).

Just to complicate matters a little, compare expressions (17a) and (17b).

There's a fly <u>on</u> the picture (17a)

There's a fly <u>in</u> the picture. (17b)

In expression (17a), the picture is understood as a 2D surface, and a 3D fly is in contact with it. But in expression (17b), the picture is understood as *depicting* a 3D space whose interior contains a fly. That is, spatial understanding has to encompass the understanding of pictures as depictions, a function that is most naturally assigned to conceptual structure (Jackendoff 2012).

Going beyond interiority and contact, *Y is <u>near</u> X* places the figural object in a region not in contact with the reference object, but proximal to it. This contrasts with *Y is <u>far from</u> X*, which places the figural object in a region distal to its reference object. As stressed by Talmy 1983, proximal and distal here are relative notions that depend on the objects in question and the spatial distribution of comparable objects: compare *a fly near the sink*, *a town near Boston*, and *a star near the sun*.

Many spatial relation terms invoke position and/or orientation with respect to the earth: either verticality or horizontality. These are so-called "absolute" or "allocentric" reference frames. For instance, *above X* and *over X* define regions projected vertically upward from X. *Under X* and *beneath X* are regions projected vertically downward from X.[10]

[10] Vandeloise (1991) observes that *over* and *under* have another sense: we can speak of *the paint over the wallpaper* or *the wallpaper under the paint*. Here *over X* defines a region such that objects in it occlude X; *under X* defines a region such that objects in it are occluded by X, regardless of the orientation of the surfaces in question.

Like *near*, the prepositions *beside* and *next to* also define regions proximal to their reference objects, but they add a further condition: the figural object must be in a *horizontal* direction from the reference object. Horizontality also appears—without proximity this time—in the spatial relations *Y is north/south/east/west of X*.

Li and Gleitman 2002 point out the curious spatial relations invoked in Manhattan: *Y is uptown/downtown/crosstown from X*. These depend on the axes of the street system on the island. However, Manhattan is not alone: Levinson 2003 observes that in the Mayan language Tzeltal, which is spoken in villages on a mountainside, the relative locations of objects are customarily expressed in terms of the axes of the terrain: essentially *uphill*, *downhill*, and *transverse*. Of course, English *uphill* and *downhill* have the same effect. However, the difference is that in Tzeltal their use is totally pervasive: if two objects are standing on a table, one says "this object is uphill from that one" rather than "this object is to the left of that one". In fact, Tzeltal has no words that translate as *left* and *right*.

6.2 *Relations that involve axis systems of objects ("intrinsic" reference frame)*

Section 5.4 discussed parts of objects that are determined by the objects' axes. A reference object's axes can also determine spatial relations, often using almost the same words. *On top of X* denotes a region projected vertically from the top of X; *in front of X* and *in back of X* denote regions projected horizontally outward from X's front and back respectively. *Behind X* is essentially the same as *in back of X*. *To the left of X* is a region projected horizontally outward from the left side of X, where the left side is determined by an axis perpendicular to X's front–back axis. These uses all fall into what is often called the "intrinsic" or "object-centered" reference frame.

Along X is an interesting case: Y can be *along the road* or *along the river*, but it cannot be *along the dot* or *along the circle*. The reason is that the reference object for *along* requires a horizontal principal linear axis, which roads and rivers have but dots and circles lack. *Along* also places a constraint on the figural object: if it has a principal linear axis, this axis must parallel the principal axis of the reference object. For instance, a road that is *along a river* must approximately parallel the river; it cannot be perpendicular to the river, no matter how close. In contrast, a house, which does not have a principal linear axis, need not have a particular orientation in order to be *along the river*.

Notice again that assigning intrinsic axes to objects in order to construct an intrinsic reference frame requires an object representation that goes beyond 2D exemplars. It requires the sort of schematization that Marr envisioned.

6.3 *Relations involving a "relative" reference frame*

Another way to define axes of objects is usually called a "relative" or "egocentric" reference frame, but it might also be termed a "viewpoint-dependent" reference frame. Here an object's axes are defined in terms of the position of an observer. The front–back axis is determined by a line from the observer to the object; the side visible to the observer counts as the (relative) front, and the occluded side counts as the (relative) back (but the West African language Hausa reverses this terminology). The left and right sides are determined by an axis perpendicular to the line of sight: the (relative) left side is the side to the left as seen by the observer, and similarly for the right side.

In this frame, *(relative) in front of X* is the region proximal to the relative front of X, and *(relative) in back of/behind X* is the region proximal to the relative back of X. *Beyond X* is the region defined by the axis projecting from the relative back of X, but with no stipulation of distance. *(Relative) to the left/right of X* is the region proximal to the relative left or right side of X, respectively.

The relative frame can be applied to objects without intrinsic horizontal axes, such as trees: if something is *in front of the tree*, it can only be construed in terms of the speaker's and/or hearer's viewpoint. However, when these terms are used with a reference object that has its own intrinsic axes, they are ambiguous if the relative and intrinsic frames do not align. For instance, in the configuration in figure 5, if the speaker adopts the intrinsic frame, he/she will say *X is behind the house and Y is to the right of the house*; in the relative frame, he/she will say *Y is behind the house and X is to the right of the house*.

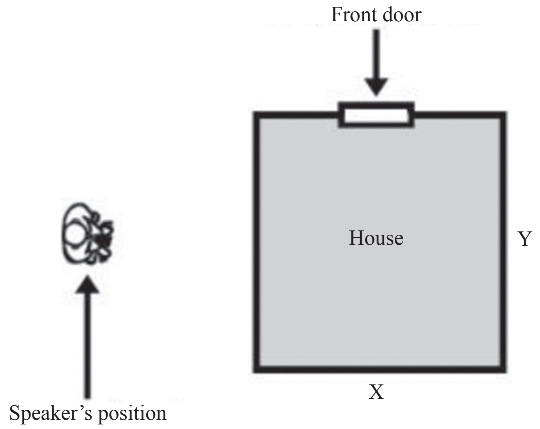

Figure 5. The positions of X and Y are characterized differently, depending on reference frame.

Different languages can favor different reference frames. In particular, Tzeltal and the Australian language Guggu Yimithirr strongly prefer absolute (earth- or terrain-based) reference frames in contexts where English speakers typically use relative reference frames (Levinson 2003).

It seems that these reference frames are not restricted to vision: Struiksma et al (2011) report reference frame variation in blind subjects exploring object configurations haptically. This shows that reference frames belong in the multi-modal (or amodal) domain of spatial understanding.

6.4 *Distances*

Section 5.1 observed that we can refer to demonstrated distances in the environment, for instance with sentences like *The fish was* <u>*that*</u> *long* [holding one's hands apart]. The grammatical system also allows us to use distances in specifying spatial relations. As mentioned above, *near* and *far* differ very coarsely in distance. But we can also make a distance explicit, as in expressions (18a) and (18b), and we can combine a distance with a direction, as in expressions (18c) and (18d).

The goat is five feet from the coat. (18a)

The fork should be <u>this</u> far away from the spoon [*demonstrating*]. (18b)

The tree is 150 feet behind the house. (18c)

New York is 90 miles northeast of Philadelphia. (18d)

6.5 *Paths*

Section 5.1 also observed that we can refer to trajectories or paths that objects traverse, as in *The fly went thataway* [pointing]. Again the language allows more elaborate path-descriptions to be constructed.

One type of path description uses earth-based coordinates: *upward* and *downward* in the vertical direction, *north(ward)*, *south(ward)*, *east(ward)*, and *west(ward)* in the horizontal direction.

The other type of path description is constructed in terms of a reference object or reference location. Some examples are listed in expressions (19), where the form of the path is specified by a preposition, and in expressions (20), where it is specified by a verb.

<u>to</u> X = 'path terminating at X'
<u>from</u> X = 'path originating at X'
<u>toward</u> X = 'path oriented in the direction of X, but not necessarily reaching X'
<u>into</u> X = 'path terminating at the interior of X' (19)
<u>up</u> X = 'path going upward along surface of X' (as in *up a tree*, *up the side of a building*)
or 'path going against direction of movement of X' (as in *up the river*)
<u>through</u> X = 'path traversing the interior of X (and reaching the other side)'
<u>past</u> X = 'path traversing a location near X'

The sun <u>rose</u>.
The train <u>approached</u> the station.
The train <u>passed</u> the station.
The bear <u>climbed</u> the tree.
The train <u>wound</u> through the canyon. (20)
The hikers <u>crossed</u> the stream.
The runner <u>zigzagged</u> through the trees.
The parade <u>circled</u> the city.
The crowd <u>wandered</u> across the fields.

It is important to recognize that paths have a cognitive role independent of the motion of objects traversing them. To be sure, the typical role of a path in the meaning of a sentence is to specify an object's path of motion, as in *The fly flew in the window*. This has often tempted analysts to think of paths simply in terms of motion, not as a separate spatial category of their own.

However, there are three other uses of paths in sentences, describing different spatial situations (Jackendoff 1983). First, a path can describe the extent of a stationary linear object [expressions (21a), (21b), and (21c)], as well as the extent of a gaze [expression (21d)].

The branch extends across the path. (21a)

This road goes from Boston to New York. (21b)

The path climbs up the mountain. (21c)

Bill looked/saw into the room. (21d)

Paths can also describe the orientation of a stationary object. In expression (22a), the principal axis of the sign is oriented on a path that extends from the sign to the theater; in expression (22b), the front-to-back axis of the house is oriented on a path that extends from the house toward the river.

The sign points to the theater. (22a)

The house faces toward the river. (22b)

Finally, a path may be used to define a location. For instance, in expression (23a), the location of my house is specified as the terminus of a path whose origin is *here* and which extends across the street. In expression (23b), Bill's location is specified as the terminus of a path whose origin is the store and which extends down the street 100 feet.

My house is across the street from here.[11] (23a)

Bill is 100 feet down the road from the store. (23b)

[11] An interesting ambiguity arises when the figural object is something that can be schematized as linear. For instance, *The rope is across the street* can denote a rope *stretched* across the street, parallel to expressions (21), or a rope *located* on the other side of the street, parallel to expressions (23).

These three uses of paths all involve static configurations, so they show that a path or trajectory is not necessarily associated with motion. Rather, motion along a path is only one of the uses of paths, albeit the most prominent one.[12]

6.6 *General discussion*

Looking at this menagerie of spatial relations as a whole, it is hard to see how they could be encoded in terms of a collection of 2D views of exemplars. For instance, a car next to a tree does not look like a coat next to a goat, or like a fork next to a spoon. Rather, these spatial relations must (at least on the face of it) be encoded in a representation in which the figural and reference objects are extremely coarse object descriptions: points, lumps, lines, or sets of axes. Such coarse representations might be regarded as variables into which detailed descriptions of particular objects are inserted—much as in language, an expression of spatial relation such as *X is on Y* can be elaborated as *the cat is on the mat*.

It might be suggested that spatial relations are encoded in the dorsal stream, and are irrelevant to object recognition. Still, they are part of visual cognition. Moreover, as Landau and Jackendoff 1993 point out, a coarse dorsal-stream object representation has to be correlated with a detailed ventral-stream representation, so that one can say not just (a) that there's a coat and a goat, and (b) that there's something in front of something, but rather that there's a coat in front of a goat—that is, to bind the identified objects to their roles in spatial configuration. Such a binding again, on the face of it, requires both ventral and dorsal object representations to have enough structure to identify the front—that is, to represent the axes.

The issue becomes more complex when we examine spatial verbs in Tzeltal (Brown 1994). Rather than simply saying "X is at/on/in Y", or "X stands at/on/in Y", Tzeltal speakers draw on a sizable vocabulary of verbs (over 250) that specify different shapes and orientations of the figural object: a standing person or another creature standing erect on its hind legs, an animate object (such as a cat) lying on its side, a vertically erect sticklike object, a vertically leaning sticklike object (where English might say *lean*), a bowl-shaped object, a tall oblong-shaped object (such as a bottle), a blob (of say, dough) with a flat surface lying down, a wide flat object (such as a frying pan) lying flat, a full bulging bag supported underneath, an object immersed in liquid in a container, and so on. These rely on coarse-grained shapes of the objects in question, as well as on animacy. Similar configurations appear in the related language Tzotzil (Haviland 1994). It is an interesting question whether this level of detail could be present in dorsal-stream representations.

7 The point

Let us sum up the spatial distinctions expressed by language:

- Basic categories such as object, action, location, path, distance
- Unusual objects such as negative parts, negative objects, hollow objects, and aggregates
- Axes of objects
- Manners of motion/change
- Locations defined in absolute, intrinsic, and relative reference frames, independent of the objects occupying them
- Paths, independent of the objects traversing, occupying, or oriented along them

These distinctions are made in everyday speech, crosslinguistically. There is nothing exotic about them.

(12) These other uses of paths have sometimes been described (eg by Talmy 1996) as "fictive motion", as though one imagines oneself traveling along the path. However, the branch, the road, and the path in expressions (21), the sign and the house in expressions (22), and my house and Bill in expressions (23) don't travel at all. Moreover, a fictive motion would take an interval of time; all these sentences can describe the situation at a moment in time.

In composing expressions such as (4)–(23) to describe one's spatial environment, one must structure the visual configuration in a way that corresponds to the individual words. In particular, in order to select the verb and/or the preposition, one must abstract the spatial relations away from the objects that instantiate these relations. Moreover, we can (and do) study the psychophysics of these relations, examining what configurations in the visual field lead to use of what linguistic expressions (eg Bowerman 1996; Carlson-Radvansky and Logan 1997; Li and Gleitman 2002; Levinson 2003; Coventry and Garrod 2004; many papers in Evans and Chilton 2010).

We conclude that the visual system must deliver information on the basis of which all of these distinctions (and many more, crosslinguistically) can be made. The challenge for the theory of vision is how the visual system does this. Specifically:

- What are the representations? How can they be characterized formally? (part of the computational theory)
- What is stored in memory that permits these judgments? (part of the computational theory)
- What representations are computed online so that people can judge novel visual configurations and describe them? (the computational and algorithmic theories)
- How are these representations and online computations encoded neurally? (the implementation theory)

Although the considerations brought to bear here are not formalized in the manner that Marr advocated for a computational theory, they constitute boundary conditions for a successful computational theory of the visual system. Such a theory cannot simply account for the recognition of single objects: it must account for spatial understanding in all its glory, in particular for all the things in the visual world we understand and talk about as if they are really there. It is legitimate to argue that one should start small and concentrate on the recognition of single objects. Nevertheless, such a theory should be formulated with attention to how it will scale up naturally to a full account of spatial understanding. I'm proposing here that evidence from language can play an important role in developing such an account.

References

Baars B J, 1988 *A Cognitive Theory of Consciousness* (Oxford: Oxford University Press)

Biederman I, 1987 "Recognition-by-components: A theory of human image understanding" *Psychological Review* **94** 115–147

Bierwisch M, 1996 "How much space gets into language?", in Bloom et al 1996, pp 31–76

Bierwisch M, Lang E, 1989 *Dimensional Adjectives* (Berlin: Springer-Verlag)

Bloom P, Peterson M, Nadel L, Garrett M (Eds), 1996 *Language and Space* (Cambridge, MA: MIT Press)

Bowerman M, 1996 "Learning how to structure space for language: A crosslinguistic perspective", in Bloom et al 1996, pp 385–436

Bresnan J, 2001 *Lexical Functional Syntax* (Oxford: Blackwell)

Brown P, 1994 "The INs and ONs of Tzeltal locative expressions: the semantics of static descriptions of location" *Linguistics* **32** 743–790

Carlson-Radvansky L A, Logan G D, 1997 "The influence of reference frame selection on spatial template construction" *Journal of Memory and Language* **37** 411–437

Cavanagh P, 2011 "Visual cognition" *Vision Research* **51** 1538–1551

Cavanagh P, Labianca A T, Thornton I M, 2001 "Attention-based visual routines: Sprites" *Cognition* **80** 47–60

Chomsky N, 1965 *Aspects of the Theory of Syntax* (Cambridge, MA: MIT Press)

Chomsky N, 1981 *Lectures on Government and Binding* (Dordrecht: Foris)

Chomsky N, 1995 *The Minimalist Program* (Cambridge, MA: MIT Press)

Coventry K, Garrod S C, 2004 *Saying, Seeing, and Acting* (New York: Psychology Press)

Culicover P W, Jackendoff R, 2005 *Simpler Syntax* (Oxford: Oxford University Press)

Dehaene S, 1997 *The Number Sense: How the Mind Creates Mathematics* (Oxford: Oxford University Press)

Dehaene S, Naccache L, 2001 "Towards a cognitive neuroscience of consciousness: Basic evidence and a workspace framework" *Cognition* **44** 1–37

DiCarlo J J, Zoccolan D, Rust N C, 2012 "How does the brain solve visual object recognition?" *Neuron* **73** 415–434

Edelman S, Poggio T, 1991 "Models of object recognition" *Current Opinion in Neurobiology* **1** 270–273

Edelman S, Weinshall D, 1989 "Computational vision: a critical review", MIT AI Memo 1158, October 1989

Edelman S, Weinshall D, 1991 "A self-organizing multiple-view representation of 3D objects" *Biological Cybernetics* **64** 209–219

Eilan N, McCarthy R, Brewer B (Eds), 1993 *Spatial Representation: Problems in Philosophy and Psychology* (Oxford: Oxford University Press)

Ernst M, Banks M, 2004 "Humans integrate visual and haptic information in a statistically optimal fashion" *Nature* **415** 429–433

Evans V, Chilton P (Eds), 2010 *Language, Cognition, and Space: The State of the Art and New Directions* (London: Equinox)

Ferreira F, 2005 "Psycholinguistics, formal grammars, and cognitive science" *The Linguistic Review* **22** 365–380

Fodor J A, 1975 *The Language of Thought* (Cambridge, MA: Harvard University Press)

Goldberg A E, 1995 *Constructions: A Construction Grammar Approach to Argument Structure* (Chicago, IL: University of Chicago Press)

Hale K, Keyser S J, 2002 *Prolegomenon to a Theory of Argument Structure* (Cambridge, MA: MIT Press)

Halle M, Marantz A,1993 "Distributed morphology and the pieces of inflection", in *The View from Building 20* Eds K Hale, S J Keyser (Cambridge, MA: MIT Press) pp 111–176

Hauser M D, Chomsky N, Fitch W T, 2002 "The faculty of language: What is it, who has it, and how did it evolve?" *Science* **298** 1569–1579

Haviland J, 1994 "'*Te xa setel xulem*' [The buzzards were circling]: categories of verbal roots in (Zinacantec) Tzotzil" *Linguistics* **32** 691–742

Heider F, Simmel M, 1944 "An experimental study of apparent behavior" *American Journal of Psychology* **57** 243–249

Heim I, Kratzer A, 1998 *Semantics in Generative Grammar* (Oxford: Blackwell)

Herskovits A, 1986 *Language and Spatial Cognition* (Cambridge: Cambridge University Press)

Jackendoff R, 1983 *Semantics and Cognition* (Cambridge, MA: MIT Press)

Jackendoff R, 1987a "On beyond zebra: The relation of linguistic and visual information" *Cognition* **26** 89–114; reprinted in Jackendoff 2010

Jackendoff R, 1987b *Consciousness and the Computational Mind* (Cambridge, MA: MIT Press)

Jackendoff R, 1990 *Semantic Structures* (Cambridge, MA: MIT Press)

Jackendoff R, 1991 "Parts and boundaries" *Cognition* **41** 9–45; reprinted in Jackendoff 2010

Jackendoff R, 1996a "The proper treatment of measuring out, telicity, and perhaps even quantification in English" *Natural Language and Linguistic Theory* **14** 305–354; reprinted in Jackendoff 2010

Jackendoff R, 1996b "The architecture of the linguistic–spatial interface", in Bloom et al 1996, pp 1–30; reprinted in Jackendoff 2010

Jackendoff R, 1997 *The Architecture of the Language Faculty* (Cambridge, MA: MIT Press)

Jackendoff R, 2002 *Foundations of Language* (Oxford: Oxford University Press)

Jackendoff R, 2007a "Linguistics in cognitive science: The state of the art" *The Linguistic Review* **24** 347–401

Jackendoff R, 2007b *Language, Consciousness, Culture* (Cambridge, MA: MIT Press)

Jackendoff R, 2007c "A parallel architecture perspective on language processing" *Brain Research* **1146** 2–22

Jackendoff R, 2010 *Meaning and the Lexicon* (Oxford: Oxford University Press)

Jackendoff R, 2012 *A User's Guide to Thought and Meaning* (Oxford: Oxford University Press)

Jackendoff R, Pinker S, 2005 "The nature of the language faculty and its implications for the evolution of language (reply to Fitch, Hauser, and Chomsky)" *Cognition* **97** 211–225

Johansson G, 1975 "Visual motion perception" *Scientific American* **232** 76–88

Kahneman D, Treisman A, Gibbs D J, 1992 "The reviewing of object files: Object-specific integration of information" *Cognitive Psychology* **24** 175–219

Kuperberg G, Choi A, Cohn N, Paczynski M, Jackendoff R, 2010 "Electrophysiological correlates of complement coercion" *Journal of Cognitive Neuroscience* **22** 2685–2701

Lackner J, Dizio P, 2000 "Aspects of body self-calibration" *Trends in Cognitive Sciences* **4** 279–288

Landau B, 1996 "Multiple geometric representations of objects in languages and language learners", in Bloom et al 1996, pp 317–364

Landau B, Gleitman L, 1985 *Language and Experience* (Cambridge, MA: Harvard University Press)

Landau B, Jackendoff R, 1993 " 'What' and 'where' in spatial language and spatial cognition" *Behavioral and Brain Sciences* **16** 217–238

Levelt W J M, 1996 "Perspective taking and ellipsis in spatial descriptions", in Bloom et al 1996, pp 77–108

Levinson S, 1996 "Frames of reference and Molyneux's question: Crosslinguistic evidence", in Bloom et al 1996, pp 109–170

Levinson S, 2003 *Space in Language and Cognition* (Cambridge: Cambridge University Press)

Li P, Gleitman L, 2002 "Turning the tables: language and spatial reasoning" *Cognition* **83** 265–294

Macnamara J, 1978 "How can we talk about what we see?", unpublished paper, Department of Psychology, McGill University

Marantz A, 1997 "No escape from syntax: Don't try morphological analysis in the privacy of your own lexicon" *University of Pennsylvania Working Papers in Linguistics* **4/2** 201–225

Marr D, 1982 *Vision* (San Francisco, CA: Freeman)

Marr D, Nishihara H K, 1978 "Representation and recognition of the spatial organization of three dimensional structure" *Proceedings of the Royal Society of London, B* **200** 269–294

Marr D, Vaina L 1982 "Representation and recognition of the movements of shapes" *Proceedings of the Royal Society of London, B* **214** 501–524

Miller G A, Johnson-Laird P N, 1976 *Language and Perception* (Cambridge, MA: Harvard University Press)

Norman J F, Norman H F, Clayton A M, Lianekhammy J, Zielke G, 2004 "The visual and haptic perception of natural object shape" *Perception & Psychophysics* **66** 342–351

Pentland A P, 1986 "Perceptual organization and the representation of natural form" *Artificial Intelligence* **28** 293–331

Piñango M M, Zurif E, Jackendoff R, 1999 "Real-time processing implications of enriched composition at the syntax–semantics interface" *Journal of Psycholinguistic Research* **28** 395–414

Pinker S, 2007 *The Stuff of Thought* (New York: Viking)

Pinker S, Jackendoff R, 2005 "The faculty of language: What's special about it?" *Cognition* **95** 201–236

Poeppel D, Embick D, 2005 "Defining the relation between linguistics and neuroscience", in *Twenty-first Century Psycholinguistics* Ed. A Cutler (Mahwah, NJ: Lawrence Erlbaum) pp 103–120

Poggio T, 2010 "Afterword: Marr's vision and computational neuroscience" (afterword to 2010 edition of David Marr's *Vision*) (Cambridge, MA: MIT Press)

Poggio T, Edelman S, 1990 "A network that learns to recognize three-dimensional objects" *Nature* **343** 263–266

Pollard C, Sag I, 1994 *Head-driven Phrase Structure Grammar* (Stanford, CA: Center for the Study of Language and Information)

Putnam H, 1975 "The meaning of 'meaning' ", in *Language, Mind and Knowledge* Ed. K Gunderson (Minneapolis, MN: University of Minnesota Press) pp 131–193

Pylyshyn Z, 2007 *Things and Places: How the Mind Connects with the World* (Cambridge, MA: MIT Press)

Riesenhuber M, Poggio T, 1999 "Hierarchical models of object recognition in cortex" *Nature Neuroscience* **2** 1019–1025

Riesenhuber M, Poggio T, 2000 "Models of object recognition" *Nature Neuroscience Supplement* **3** 1199–1204

Rock I, 1983 *The Logic of Perception* (Cambridge, MA: MIT Press)
Rolls E T, 2011 "David Marr's Vision: *floreat* computational neuroscience" (review of 2010 edition of David Marr's *Vision*) *Brain* **134** 913–916
Sadock J, 1991 *Autolexical Syntax* (Chicago, IL: University of Chicago Press)
Sag I, Wasow T, 2011 "Performance-compatible competence grammar", in *Non-transformational Syntax: Formal and Explicit Models of Grammar* Eds R D Borsley, K Börjars (Malden, MA: Wiley-Blackwell) pp 359–377
Struiksma M E, Noordzij M L, Postma A, 2011, "Reference frame preferences in haptics differ for the blind and sighted in the horizontal but not in the vertical plane" *Perception* **40** 725–738
Talmy L, 1978 "The relation of grammar to cognition: A synopsis", in *Theoretical Issues in Natural Language Processing* volume 2, Ed. D Waltz (New York: Association for Computing Machinery); revised version in Talmy L, 2000 *Toward a Cognitive Semantics* (Cambridge, MA: MIT Press)
Talmy L, 1983 "How language structures space", in *Spatial Orientation: Theory, Research, and Application* Eds H Pick, L Acredolo (New York: Plenum); revised version in Talmy L, 2000 *Toward a Cognitive Semantics* (Cambridge, MA: MIT Press)
Talmy L, 1996 "Fictive motion in language and 'ception' ", in Bloom et al 1996, pp 211–276
Tanenhaus M, Spivey-Knowlton M J, Eberhard H H, Sedivy J C, 1995 "Integration of visual and linguistic information in spoken language comprehension" *Science* **268** 1632–1634
Tarr M J, 1995 "Rotating objects to recognize them: A case study on the role of viewpoint dependency in the recognition of three-dimensional objects" *Psychonomic Bulletin & Review* **2** 55–82
Tarr M J, Bülthoff H H, 1998 "Image-based object recognition in man, monkey and machine" *Cognition* **67** 1–20
Topolinski S, Türk-Pereira P, 2012 "Mapping the tip of the tongue: deprivation, sensory sensitisation, and oral haptics" *Perception* **41** 71–92
Tversky B, 1996 "Spatial perspective in descriptions", in Bloom et al 1996, pp 463–492
Ullman S, 1998 "Three-dimensional object recognition based on the combination of views" *Cognition* **67** 21–44
Vandeloise C, 1991 *Spatial Prepositions: A Case Study from French* (Chicago, IL: University of Chicago Press)
Van der Zee E, Slack J (Eds), 2003 *Representing Direction in Language and Space* (Oxford: Oxford University Press)
Wittenberg E, Paczynski M, Wiese H, Jackendoff R, Kuperberg G (submitted) "Light verbs don't make light work"