from R. Jackendoff
from Consciousness and the
Computational Mind.
MIT Press, 1987, pp. 193-212

Chapter 10

The Connection between Language and Vision

#24

10.1 Preliminary Correspondences

As mentioned in section 7.2, one of the fundamental problems for a theory of natural language appears as the title of John Macnamara's (1978) paper: "How Do We Talk about What We See?" In order to approach this problem, a point of connection must be found between the theory of language and the theory of vision. Following Marr, we have so far spoken of the 3D model primarily as the culmination of visual processing and the locus of visual form recognition, only hinting at its connection to conceptual structure and thence to language. This chapter develops the relationship between language and vision more explicitly and shows that their connection not only helps answer Macnamara's question but also solves certain outstanding problems in both areas.

Within an information-processing theory of mind it is not enough to say that we talk about what we see by transferring information (or representations) from visual memory into linguistic memory. The heart of the problem, as Macnamara recognized, is one of *translation*: in order for us to talk about what we see, information provided by the visual system must be translated into a form compatible with the information used by the language system. The essential questions are these: (1) What form(s) of information does the visual system derive? (2) What form of information serves as input to speech? (3) How can the former be translated into the latter?

The importance of establishing a formal translation between visual and linguistic information cannot be overestimated: it makes possible for the first time a theory whose levels of representation extend all the way from the retinal image through the core of thought out to the vocal tract. The levels of linguistic representation therefore need not back up indefinitely into murkier and murkier central representations, nor does linguistic theory have to appeal to a mysterious direct connection of "intentionality" with the external world. Rather, the computational theory will itself provide all the necessary links for describing how we talk about what we see. It will thus be able to relate mental representations to an

important part of the "world of non-symbols," in the sense discussed in chapter 7.

Within the present framework the translation between language and vision should be specified by a set of correspondence rules between one or more visual levels and one or more linguistic levels. Ideally, the two faculties should interact via those representations whose units come into closest correspondence and whose functions are most closely related.

Under this basic criterion the most appropriate levels to link appear to be conceptual structure and the 3D model. Of all the levels we have examined, it is only these two that encode the notion of *physical object*—though in quite different fashions, to be sure. In linking these levels, our first principle of correspondence can therefore relate 3D constituents that encode objects to conceptual constituents that encode objects; there is no other pair of levels over which this essential relation could be stated.

Another important correspondence that can be established immediately is in the encoding of the conceptual relation of *part-whole* or *inalienable possession*. This relation has always been one of the staples of linguistic semantics; Marr's hierarchical theory of the 3D model allows it to be spelled out explicitly in spatial terms (recall the discussion of the human body shown in figure 9.6). Basically, for physical objects X and Y, X IS A PART OF Y obtains in conceptual structure just in case the 3D model corresponding to X is an elaboration within the hierarchical 3D model corresponding to Y.

A third basic correspondence arises in the functions that Marr considers central to the 3D model: object identification (that thing before me is a known individual, say, Rover) and object categorization (that thing before me is of a known type, say, a dog). As discussed in section 8.1, the linguistic level that performs similar functions is conceptual structure. Again a good first hypothesis is that, in order to capture the relationship of visual and conceptual categorization, the 3D model and conceptual structure are the right levels to link. We return to these functions in section 10.3.

10.2 The 3D Model as a Central Representation

Before exploring the connection between the 3D model and conceptual structure, let us establish the overall role of the 3D model in the computational mind by briefly noting some other uses to which it might be put.

First, one can identify the shapes of objects not only by looking at them but also by handling them—so-called *haptic* perception. This suggests that the 3D model notions of physical object, shape, and contour are shared between the visual and haptic modalities. However, the haptic route to the 3D model is informationally distinct from the visual route. It certainly does not make use of anything like the primal sketch, which encodes local

boundary segments in the retinal field. And although haptic perception is likely to involve some integrated representation of "touchable" surfaces, such a representation must be adapted to the shape and mobility of the hand rather than to the eye-centered coordinate system of the 2½D sketch. Thus, the primal and 2½D sketches, concerned with the recovery of shape information from a retinal image, are proprietary to the visual system and are not part of the haptic system. The parallel haptic levels will have to deal with recovering shape information from touch and pressure sensors in the skin. Through faculty-specific correspondence rules, both systems converge on a common representation of shape, the 3D model; this is what enables one to identify the same object by sight or by touch.

Next consider what is probably the most important use of information derived from the visual system: to help us find our way around in the world. From an evolutionary point of view this use is prior to the linguistic connection, since it is shared by nonlinguistic organisms. The eventual scope of the 3D model envisioned by Marr includes not just individual objects but the full spatial layout of the perceived world; we will be pursuing suitable extensions in section 10.5. Given such extensions, it is easy to imagine using the 3D model as an *input level for the capacity for physical action*: this level encodes what there is in the environment for the organism to approach or avoid, as the case may be.

Among the objects in the environment, of course, is one's own body. In order to compute what to do with one's body in order to act in the environment, one must have information about the spatial layout and motion of one's limbs and a sense of one's position and motion as a whole with respect to other objects in the environment. The 3D decomposition in figure 9.6 of the human body provides a natural way of encoding this information.

However, in the case of this particular object one has a privileged set of sensory cues that exist for no other object in the environment. Touch and pressure cues from the skin provide evidence about the direction and magnitude of forces (including gravity and acceleration) on the body; the semicircular canals in the ears provide evidence about rotation of the head in various directions; sensors in the muscles and joints provide information about the position of the limbs. In addition, one has visual cues from observation of the visible parts of the body.

Despite this plethora of sensory cues there is in awareness an essentially unified sense of one's spatial position and motion—one cannot consciously dissociate the sources of one's position sense. A series of experiments by James Lackner and associates (Lackner and Levine 1978; Lackner and Graybiel 1983, 1984; Lackner and DiZio 1984; Lackner and Taublieb 1984; Lackner 1981, 1984, 1985) has shown that all these sources play a role in the position sense and that they interact strongly with each other. For

instance, one can elicit in a subject an illusory sensation of arm movement by mechanically vibrating the biceps muscle while the arm is restrained. A subject watching his arm under such conditions visually experiences it as being in motion. However, watching the arm turns out to reduce the magnitude of illusory experienced motion, indicating that the visual and proprioceptive cues partially cancel each other out when there is a conflict. Lackner et al. find similar interactions between all combinations of cues for position sense.

In addition, a part of position sense that we more or less take for granted is the sense of the *size* and *shape* of our bodies. Lackner et al. have created illusions of change of size and shape as well as position. For instance, if a finger is placed on the nose and then the arm is subjected to illusory extension as above, one senses one's nose growing longer! Under various other conditions the legs seem to grow longer, the finger seems to be detached from the hand, or one seems to have multiple arms. Evidently the cues from position and motion must be integrated into a unified and consistent representation of the body. Under illusory conditions there is give and take among the various sources of information, resulting in a more or less self-consistent but anomalous perception of the body, with no conscious sensation regarding which sensory modality is providing which cues.

The usual story applies here. In order for information from a variety of sources to interact, there must be a common format into which all the sources of information are translated. The appropriate integrative level for the position sense would seem to be the 3D model, whose primitives of size, shape, and relative position of parts are well suited for the purpose. Thus, it is possible to envision a whole series of mappings from the different sensory sources to the 3D model of one's own body—a part of this level of representation that is specialized in terms of sources of information but that differs not at all in its formal structure as such.

More speculatively, notice that organisms such as bats and dolphins that use echolocation (sonar) to find their way around may have still another connection to the 3D structure, through that modality. Again, the lower levels of information structure that lead up to the 3D model in echolocation will be distinct from those for vision and touch. (Alternatively, this capacity might merge with vision at the 2½D sketch, since it does provide a viewer-centered image of sorts. I don't know at this point how one could tell the difference, but in principle it should be an empirically decidable question.)

The overall hypothesis that emerges from these considerations is that the 3D model, like conceptual structure, is one of the central interface languages of the computational mind. It is a general-purpose representation for all tasks involving spatial cognition, and language, vision, touch, action, and the body senses can all influence it and make use of it.

It is interesting to consider some consequences of such a hypothesis. For one thing, it says that the spatial information encoded by the blind, at its most central level of representation, is of the same nature—uses the same primitives and principles of combination—as spatial information available to the sighted. Furthermore, the way this information is put to use in finding one's way about in the world is the same. This is not to say that the same wealth of spatial information is available at any moment or even in toto to the blind, because of the inherent limitations of the haptic channel of input. But all such notions as object, distance, angle, trajectory, and so forth that play a role in spatial perception will be useable and derivable—not via the route of primal and 2½D sketches but via as yet unknown lower levels of representations specialized to the haptic capacity. (See Landau, Spelke, and Gleitman 1984 for confirming evidence.)

Of course, not all aspects of the 3D model need be equally available to all the capacities with which it exchanges information. For instance, color information is brought into the 3D model via the visual system, and it is available to conceptual structure. However, it obviously plays no role in the haptic or action systems; their correspondence rules are entirely silent on the translation of color information. (If we could somehow discern colors by touch, the story would be different.)

This raises a converse question: Is there spatial information supplied by touch that is unavailable to vision? Two candidates might be temperature and weight. How such properties might be formally integrated into a spatial representation remains open, though they need not present any more inherent difficulties than color.

It has sometimes been suggested to me that instead of Marr's 3D model serving general spatial cognition, one might reserve the 3D model specifically for visual cognition and introduce a separate amodal representation for space, fed by the 3D model and all the other sensory faculties. Note, however, that this would not change in any way the overall hypothesis of a level of spatial representation as a central level of cognition; it would only add an extra way station in the visual system.

In fact, at the moment I see no justification for the move. The essential primitives of the 3D model involve shape, contour, and motion; the essential principles of combination involve building up objects from parts. These are likewise essential for the haptic sense, for the body position sense, and, for that matter, for echolocation. What different primitives might be justifiably introduced for the supposed new level is altogether obscure. Even if each of these faculties contributes a subset of proprietary vocabulary in the 3D model (for example, color or weight), that does not negate their fundamental unity of formal structure. So for now there seems no point in an extra level.

10.3 Visual Identification and Categorization

We now come back to the central topic of the chapter, the interplay between the 3D model and conceptual structure.

Recall that Marr designed the 3D model level to encode long-term memory information suitable for either object *identification* or object *categorization*. But now let us ask, How is the long-term memory for a known individual distinguished from that for a known category? So far the two are not distinguished in formal structure—they are both 3D models—so what makes the difference?

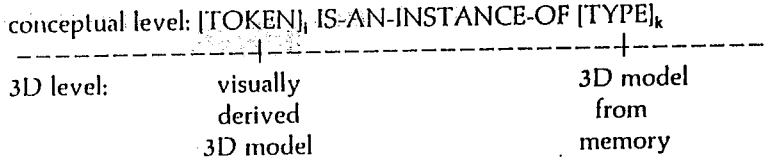
One's first impulse is to claim that memories of individuals and memories of categories differ in vagueness or generality, individuals being much more specific. But this will not do. One may be vague about the appearance of a slightly known individual—say, the car that hit mine and sped off into the night—and therefore encode it rather imprecisely in memory; on the other hand, one may be very knowledgeable about the appearance of an extremely delimited category—say, IBM PC keyboards—and therefore encode it in great detail and specificity.

Further reflection suggests that in fact there are *no* features of the 3D model, which is purely geometric in conception, that can distinguish representations of individuals from representations of categories. For example, the 3D models for the individual "my dog Rover" and for the category "dogs that look just like Rover" are necessarily identical, because of the way the category is defined. What is needed to distinguish the two kinds of representations is in fact the binary [TYPE/TOKEN] feature of *conceptual* structure, an *algebraic* form of representation. Only an algebraic structure can provide the proper sort of distinct two-way opposition.

Let me be more precise. The claim is that visual memory contains not just 3D representations but matched pairs of representations: a 3D model for how the thing looks and a conceptual structure that "annotates" the visual representation, specifying at least whether this is taken as a representation of a token or a type. The visual forms given by *perception* are automatically linked to the [TOKEN] feature in conceptual structure: that is, what one directly sees consists of particular individuals. On the other hand, what one *learns* and stores in memory can be linked either with [TOKEN] (if one is remembering an individual) or with [TYPE] (if one has learned a category).

Now consider the relation between the individual being perceived and the remembered individual or category. The two 3D model representations must be juxtaposed and compared, and the outcome of the matching process must be recorded. (This is the "Höffding step" of classical perception theory; see Neisser 1967.) But the outcome of a match cannot be represented visually: it is basically of the form "successful match" or "unsuccessful match." It can, however, be encoded in conceptual structure. A successful match in object categorization is encoded conceptually by our old friend IS-AN-INSTANCE-OF (section 8.1), an algebraic relation between a [TOKEN] and a [TYPE]. Object identification is encoded by a different relation, which may be called IS-TOKEN-IDENTICAL-TO, a relation between two [TOKEN] concepts. The overall form of the two relations is given in figure 10.1; the vertical lines indicate associations or linkages between representations at the two levels. Note that the 3D model part of the judgment is exactly the same in both cases: the comparison of a structure derived from vision with one from memory. The only difference between identification and categorization, then, lies in the conceptual level.

a. Object categorization



b. Object identification

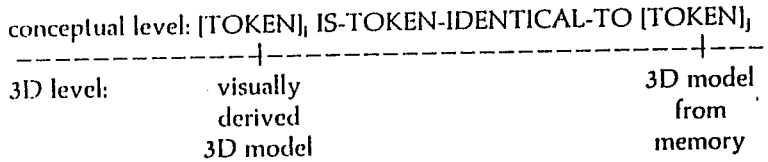


Figure 10.1
The roles of conceptual structure and the 3D model in (a) object categorization and (b) object identification

The notion of paired 3D and conceptual structures helps solve another, more often recognized problem concerning the visual encoding of categories of *actions*. The need for such categories has cropped up occasionally in the literature. For instance, Marr and Vaina (1982) discuss how a few "basic action types" such as throwing, saluting, and walking can be defined in terms of sequences of motions of body parts in the 3D model. Peterson (1985) suggests that there is a class of "natural actions" described by verbs like *throw* and *push*, analogous to "natural kinds" like *dog* and *banana*. Like natural kinds, natural actions are learned by ostension ("This is what it looks like") more than by definition. Moreover, section 8.4 observed that we can point to actions ("Can you do *that*?"), concluding that action information must be provided by the visual system.

How are action categories to be encoded? The problem goes back in its essence at least to the British empiricists. The visual representation of the action of walking, for example, requires by its very nature a walking *figure*—say, a generalized human. But then, what is to make it clear that this is a representation of *walking* rather than of *human*? The requisite distinction is simply not available in the geometric representation. However, it *is* available in conceptual structure, where we have the algebraically structured features that distinguish major ontological categories (section 8.4). By linking the 3D figure in motion to an [ACTION TYPE] concept rather than an [OBJECT TYPE] concept, we can encode the fact that the motion of the figure rather than its shape is taken as the significant information in the 3D model. Thus, a linkage of 3D and conceptual structures again provides the right range of distinctions.

10.4 The Use of 3D Models in Word Meanings

We have just seen that the visual system must make use of the level of conceptual structure in encoding long-term memories of individuals and categories. In this section we will see that language likely makes use of the 3D model in encoding distinctions among word meanings. This will reinforce the view of conceptual structure and the 3D model as partners in central representation.

First, there are distinctions of meaning among words that appear to be spelled out far more naturally in terms of spatial structure than in terms of conceptual structure. A good example (brought to my attention by Thomas Kuhn) is distinguishing among ducks, geese, and swans. In conceptual structure it is quite natural to make a taxonomy of these types, such that they are distinct from one another and together form a larger type "waterfowl," itself a subtype of birds. But how are the differences among these types to be expressed? Clearly, one of the most salient differences, and the one by which an individual is normally classified into one or the other of these categories, is how ducks, geese, and swans *look*—their relative sizes and the proportions and shapes of their respective parts.

Now the idea that these differences are represented in conceptual structure by features like [±LONG NECK] is implausible, because the features seem so ad hoc. Yet these have been the sorts of features to which descriptive semanticists have had to resort, for lack of anything better. (One suspects, in fact, that the evident need for such bizarre features is one of the major factors contributing to the suspicion with which lexical semantics has often been regarded.) However, notice that descriptions of size, proportion, and shape of parts, being purely geometric notions, are quite naturally expressed in the 3D model, which must include them in any event in order to accomplish object recognition. This suggests that conceptual structure may be divested of a large family of ad hoc descriptive features by encoding such information in 3D model format, where it is not ad hoc at all but precisely what this level of representation is designed to do.

An immediate implication is that *the representation of a word in long-term memory need not be just a triple of partial phonological, syntactic, and conceptual structures but may contain a partial 3D model structure as well.* From a different angle, a word for a spatial concept may be thought of as appending syntactic and phonological structures to a language-independent spatial concept, itself built from a linkup of conceptual and 3D structures.

This conclusion reflects the intuition that knowing the meaning of a word that denotes a physical object involves in part knowing what such an object looks like. It is the present theory's counterpart of the view that one's lexical entry may contain an image of a stereotypical instance. However, as observed in section 9.4, the 3D model provides a much more coherent account of what lies behind such an intuition than does a rigid "picture-in-the-head" notion of stereotype, allowing it to play a more interesting role in a formal lexical semantics.

Not only nouns benefit from 3D model representations. For instance, a group of verbs such as *walk, run, jog, lope,* and *sprint* differ from each other in much the same way as *duck, goose,* and *swan.* It is embarrassing even to consider a set of binary algebraic features that will distinguish them. However, since the 3D model level can encode actions, it can naturally provide the relevant distinctions in gait and speed as a part of the verbs' lexical entries.

Going further afield, consider *functional* definitions, which pick out objects that one can use in a certain way. For instance, an important component of the concept of *chair* is that it is something to sit in. How can this be encoded? Sitting is a "natural action," specifiable by an [ACTION TYPE] linked to a 3D model of what sitting looks like. The chair, in turn, can be specified as an auxiliary character in the action: it is the surface upon which the acting figure comes to rest. In the 3D model its appearance can be specified very coarsely, giving only its approximate size and the crucial horizontal surface that the figure makes contact with. Thus, as Vaina (1983) points out, a functional definition can be encoded by linking a particular object in a 3D action description with an [OBJECT TYPE] in conceptual structure—the 3D model encodes what one does with the object, plus only enough of the object's appearance to show how one does it.

Although the formal niceties of such word meanings are yet to be worked out, I think it is possible to see the germ of an important descriptive advance here. By using linkages of 3D models with conceptual structure, one can begin to circumvent the limitations of the purely algebraic

systems to which semantics has been largely confined and at the same time begin to see how language can make contact with the world as perceived.

This is not to say that *all* elements of linguistic meaning are conveyed by 3D models. Far from it. The essential algebraic features described in chapter 8 have been shown to be necessary even for *visual* memory, much less language. Moreover, such aspects of meaning as negation and quantification are fundamentally conceptual and cannot be translated into a 3D model. And of course there are plenty of words that express auditory, social, and theoretical concepts, for which no 3D counterpart should be expected. The point is only that when language has an opportunity to exploit the expressive power of the 3D model, it does so, and hence that one should expect words for spatial concepts to exhibit characteristically geometric distinctions as well as algebraic.

10.5 Enriching the Conceptual-3D Connection

If the 3D model is to be the component of the visual system most directly responsible for our ability to talk about what we see, it must be rich enough in expressive power to provide *all the visual distinctions that we can express in language*. (It may of course be even richer—there may be further 3D model distinctions that are not expressible in language but that play a demonstrable role in spatial cognition or the capacity for action.)

This section will use evidence from language to suggest some natural enrichments of Marr's theory. These will help extend the 3D model beyond the description of individual objects and actions to a description of the full spatial array. The evidence comes from the sorts of sentences concerning spatial location and motion discussed in section 8.5.

It has often been noted (Gruber 1965; Clark and Chase 1972; Miller and Johnson-Laird 1976; Jackendoff 1976; Talmy 1983; Langacker 1986) that spatial relationships between two objects are pretty much never expressed symmetrically in language. Rather, the language usually distinguishes two roles: a landmark or reference object, which often appears as the object of a preposition, and a figural object or theme, which often appears as grammatical subject. For instance, in (10.1a) *the table* is the reference object and *the book* is figural. And it is intuitively clear that there is a distinct difference between (10.1a) and (10.1b), where the roles of reference object and figure have been exchanged.

(10.1) a. The book is on the table.
b. The table is under the book.

The usual examples illustrating sentences of spatial relation, like (10.1), use only the neutral verb *be*. However, the asymmetry between the figural and reference objects is clearer in sentences of location that use more specific verbs, as in (10.2).

(10.2) The book is { standing / lying / leaning / resting } on the table.

Here the lexical distinctions among the verbs encode object-internal information about the figural object, the book—in particular, the spatial disposition of its major coordinate axis. In other words, whereas the neutral verb of location *be* gives only the very coarsest description of the subject, more specific verbs of motion and location elaborate some internal details of its 3D model. In turn, this supports the asymmetry of the relation between the figural object and the reference object.

The proper treatment of this asymmetry in conceptual structure (Jackendoff 1978; 1983, chapter 9; Talmy 1983; Herskovits 1985) is to make use of the formalism of section 8.5, in particular the conceptual category Place. As seen there, locational sentences like (10.1) and (10.2) assert, not a spatial relation between two objects, but the place at which the figural object is located. In turn, the place is specified as a function of the reference object, each choice of preposition specifying a different function and hence determining a different place. The conceptual structure of such sentences is therefore organized as in (10.3), paralleling (8.14).

(10.3) $[_{State}$ BE $([_{Object}$ BOOK$]$, $[_{Place}$ ON $([_{Object}$ TABLE$])])]$

If indeed conceptual structure is set in correspondence with a 3D model structure, the notion of Place ought to play a role in the latter level of representation. And in fact it seems quite natural to encode in the 3D model the notion of regions of space related to an object, determined in terms of the object's 3D representation. For example, the preposition *in* expresses a function that (for a first approximation) maps an object into the region consisting of its interior. *On* maps an object into the region consisting of its surface (in many cases, its upper surface). *Near* maps an object into a region exterior to but contiguous with the object. *Beside* is like *near* but restricts the region to roughly horizontal contiguity (a cloud above a mountaintop may be *near* it; it cannot be *beside* it).

More interesting are the prepositions that make use of the reference object's own intrinsic coordinate axes. For instance, one can be either *beside* or *along* a road, but one can only be *beside*, not *along*, a field or a tree (unless the tree has been felled). Evidently the domain of the function expressed by *along* is (roughly) objects of significant extension in one horizontal dimension only; this function maps such an object into an exterior contiguous region.

Another well-known class of examples of this sort consists of prepositions such as *in front of, behind, on top of,* and *to the right of.* These are used in two ways. Suppose the reference object is a person or a house. Then it has its own intrinsic axes, used in the 3D model to determine the internal layout of its parts. For instance, the head of a person and the roof of a house go on top, as specified by the directed up-down axis; the face and the front door go on the front, as specified by the directed front-back axis. These axes, independently necessary to establish the form of the object, may simply be extended beyond the surface of the object to determine regions that can be referred to by prepositional phrases such as *in front of the house, behind the man,* and so on.

On the other hand, some objects such as featureless spheres have no intrinsic axes. In such cases the position of the speaker (or hearer) extrinsically imposes a set of coordinate axes on the reference object: the front is the side facing me (or you), so that *behind the sphere* picks out a region contiguous to the sphere and on the side of it not facing us.

As has frequently been noted (Talmy 1983; Clark 1973; Olson and Bialystok 1983), ambiguities arise in the use of such prepositions in cases where two possible sets of coordinate axes are available and in conflict. For instance, in the situation depicted in figure 10.2, *The ball is behind the house* may be read as describing the ball at either position X (house's intrinsic axes) or position Y (speaker-imposed axes). Similarly, if Fred is lying down, *Fred's hat is on top of his head* can describe a configuration where the hat is in its normal position relative to his head (*on top of* is in terms of Fred's intrinsic axes) or one where it is covering his face (*on top of* is in terms of gravitational axes).

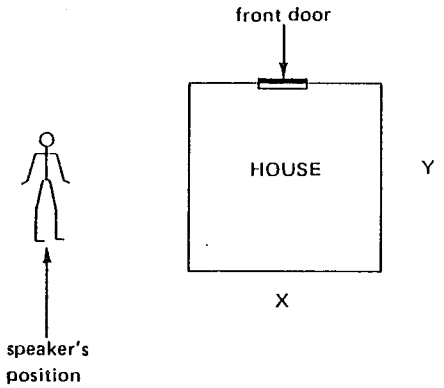


Figure 10.2
The sentence *The ball is behind the house* is ambiguous for a speaker at the position shown.

One of the points made by Olson and Bialystok (1983) is that such conflict between coordinate systems may occur even in purely spatial tasks, where language is not involved. This suggests, at the very least, that extending object-internal coordinate axes to the space exterior to an object plays a role in spatial understanding and that linguistic expressions of location are simply encoding information that is present for independent purposes. (See Shepard and Hurwitz 1984 for development of this point with respect to the preposition *up* and for further discussion of spatial axis systems.)

Let us turn next to the linguistic description of the motion of objects. This often divides rather nicely into two parts, which may be called *object-internal* and *object-external* aspects. (Lasher (1981) makes a similar distinction between "contour motion" and "locomotion.") Object-internal motion is in many cases expressed by lexical distinctions among verbs of motion; object-external motion, by the structure of accompanying prepositional phrases. Consider the possibilities in (10.4), for instance.

(10.4) John { walked / ran / squirmed / crawled / soared } { under the bridge. / into the room. / through the tunnel. / over the rug. / along the road. }

As in the location sentences (10.2), the differences among the verbs reflect the internal dispositions and motions of the parts of John's body; they express the object-centered description of John himself. As observed in the previous section, these differences are not easily characterized in terms of conceptual features. They are, however, rather naturally differentiated in terms of 3D action descriptions.

On the other hand, the differences among the prepositional phrases in (10.4) reflect the motion of John as a whole. For this part of the description John can be regarded as a very coarsely described object (a point or an undifferentiated lump) traveling along some externally specified trajectory. Thus, the total description can be seen as hierarchical: the outer layer, expressed by the prepositional phrase, is the external trajectory of the object; the inner layer, expressed by the verb, is an object-centered elaboration of the object-internal motion. (There are verbs in English that themselves express object-external motion, as in *John circled the tree* or *The fly spiraled down to the table,* so the conceptual division is not always reflected grammatically. The semantic intuitions are clear, however, and it is these that we are concerned with.)

Talmy (1980) points out that in some languages (such as French and Spanish) this bifurcation of motion description is even clearer, in that one cannot say "John squirmed through the tunnel" but must say, literally,

"John went through the tunnel squirming." Here the main clause expresses the object-external motion, while the dependent clause *squirming* expresses the object-internal motion. A similar constraint exists in Japanese (Mitsuaki Yoneyama, personal communication).

Object-external motion is encoded in conceptual structure as an Event constituent of the sort in (8.12).

(10.5) $[_{\text{Event}}$ GO $([_{\text{Object}}$ JOHN$], [_{\text{Path}}$ TO $([_{\text{Place}}$ IN $([_{\text{Object}}$ ROOM$])])])]$

In particular, the Path constituent corresponds to the trajectory of motion, and the preposition expresses the function that maps the reference object (here, the room) into the Path (here, a trajectory that begins outside the room and ends inside it).

Some of the prepositions that express such functions treat the reference object at the coarsest layer of description. *To*, for instance, treats its object essentially as a point (which may be elaborated, of course, by the object's own description). Many other prepositions of path, however, exploit the reference object's geometry to some degree or another. For instance, *into* and *through* describe paths traversing the interior of the reference object, and *onto* encodes a path that terminates on the reference object's surface.

293

One of the more complex cases is *across*. As pointed out by Talmy (1983), the kinds of objects one can go across usually have one horizontal axis longer than the other and edges roughly parallel to the long axis. Going across such an object involves traversing a path from one of these parallel edges to the other. For instance, one can go across a road or a river, and across the short dimension of a swimming pool, but not across the long dimension of a swimming pool—in fact, there happens to be no appropriate English preposition for this case. (There are exceptions to these conditions, due to the presence of a preference rule system as described in section 8.3, but I think the basic principles stand. This description may be thought of as "stereotypical *across*.")

Marr's theory as it stands does not include any notion of "trajectory traversed by an object". However, it is not difficult to imagine extending the 3D model to include such information as Paths. Just as with Places, the geometrical properties of the reference object that are referred to by Path-functions seem most often to be information that is independently necessary for a Marr-type description of the object itself—for instance, its major coordinate axes. Hence, the extension of the 3D model to include Paths seems quite natural, and yet another correspondence can be established between constituents of conceptual structure and 3D model structure.

There is nonlinguistic motivation as well for a notion of Path in spatial understanding. If an organism is to use something like a 3D model derived from vision to direct its action—to find its way around in the world—it will necessarily have to compute trajectories, both those of moving objects, to see where they will end up, and those that it plans to traverse in its own future action. Thus, the enrichment proposed in order to adequately account for motion expressions in language in fact serves an independent purpose. The language is just capitalizing on what is already present in spatial understanding.

10.6 Summary

From this very preliminary investigation of the correspondence between conceptual structure and 3D model structure, a number of points emerge. First, the two structures can indeed be brought into contact, and much of spatial thinking depends on aspects of both. Thus, these two levels of representation constitute a central core accessed by a number of different peripheral faculties, including visual perception, language, haptic perception, body perception and action.

Second, the theory of Conceptual Semantics developed in *S&C* and summarized in chapter 8 contains many of the right sorts of elements to interface with spatial understanding. This is evidence that it is on the right track toward a properly mentalistic theory of the semantics of natural language, and in particular toward an answer to the question of how we talk about what we see.

Third, language can provide evidence for the constitution of the 3D model level, in that it motivates elements such as Places and Paths that are not immediately apparent in intuitions about the theory of vision per se. Such enrichment is expressed naturally in terms of elements that are already made available by Marr's theory and that turn out on reflection to be supported by the organism's performance at nonlinguistic tasks. Moreover, these enrichments lead the theory toward an adequate description of the full spatial field, in that they encode relations among objects.

Fourth, the 3D model level, thus enriched, does not conform to one's intuitive stereotype of what information the visual system delivers. The "visual world" is not simply a collection of holistic objects—"statues in the head." Rather, the 3D representation is teeming with elements that one does not "see," such as the hierarchical part-whole structure and the coordinate axis systems proposed by Marr, and now regions determined by the axes of objects and trajectories being traversed by objects in motion. In other words, the information necessary to encode spatial understanding includes a great deal that, although still purely geometric rather than conceptual, is not part of visual appearance as such.

Some (though I hope not all) readers may question this idea: How can a theory of perception countenance the presence of abstract visual informa-

tion that one does not see? From the vantage point of linguistic theory, though, such a situation seems entirely as it should be. A major thrust of generative linguistic theory is that there are elements of hierarchical and abstract structure that one cannot hear and that one does not speak but that must play a role in explicating one's understanding of language. Moreover, there is nothing inherently suspect in investigating such entities: one can ask meaningful empirical questions about them and choose intelligently between alternative hypotheses. This is exactly the situation we have arrived at here in visual theory, paralleling language. If it calls for new methodology and new forms of argumentation, so be it; the flowering of linguistic theory in the last quarter century has been a direct result of giving up the expectation of overly concrete solutions.

10.7 Special-Purpose Capacities That Draw on Vision

To give a broader idea of the role of the 3D model in cognition, I conclude this chapter with extremely brief discussions of three special-purpose capacities that draw on it.

10.7.1 Face Recognition

Some research (Carey 1978; Carey and Diamond 1980, and references cited there) suggests that there is a specialized human capacity for face recognition and memory for faces. Like language, this capacity seems to have a characteristic developmental course (major growth between 6 and 10 years of age) and a brain localization (right posterior), damage to which produces deficits in face recognition but not in vision in general. Memories for faces are remakably differentiated and long-lasting: people remember many thousands of faces and often can recognize casual acquaintances they have not seen for many years—even despite changes due to aging. And this capacity is informationally not very well linked to language: try describing your mother's face so someone else could pick her out in a crowd, and compare that to how easy it is pick her out in a crowd yourself.

Where might this capacity fit into a computational theory? It would be silly to suppose that it is entirely separate from the visual system. Rather, it would seem most plausible to assume that it is a specialization within the visual system: most of the work of face recognition is done by the standard devices of vision, but as some point an additional component takes over the information and performs face-specific analysis on it. That is, we would ideally like face recognition to be "parasitic" on independently necessary characteristics of ordinary object recognition.

So the standard question arises: What is the form in which faces are encoded by this special-purpose capacity, and to which of the already known levels is this most closely related? Since face recognition is a specialized form of object recognition, the appropriate jumping-off point appears to be the 3D model representation. Like ordinary objects, faces are recognized, not at a particular single orientation or distance, but from a variety of orientations (though not upside down) and at any distance that permits adequate resolution of facial features. Thus, memory for faces should be encoded in a normalized form and in some sort of object-centered coordinate frame, altogether parallel to the ordinary 3D model.

This suggests that facial representation involves a set of primitives and principles of combination of the general form of those for the 3D model level, but very specific and refined for the task of differentiating human faces. Since faces are, after all, spatially integrated with bodies and with the rest of the world, these representations must be related to the more general 3D model structure via correspondence rules. In addition, there might well be some specialized correspondence rules that are designed to pick facial features more effectively out of the 2½D sketch. But since research on this capacity has primarily been directed at showing that it exists, and not at the structure of the information it processes, it is hard to go beyond speculation at this point.

It is worth mentioning, though, that this capacity is not specific to humans. Along with voice recognition, emotion recognition (from visual and/or auditory cues), and possibly other specialized capacities, it subserves an elaborate system of *social cognition* in many mammals and especially primates. Though it has long been recognized that animals have an elaborate social existence, little attention has been paid to the substantial computational capacities that social cognition presupposes. However, Cheney and Seyfarth (1985) observe that vervet monkeys deal routinely with complex problems in the social domain—say, three-term transitivity problems over the social dominance hierarchy—even though they are quite inept at problems of comparable logical structure in the nonsocial domain, the sort presented by the usual laboratory tests of animal intelligence. Cheney and Seyfarth go so far as to speculate that human intelligence might have arisen as a generalization of social cognition.

In any event, one of the essential foundations of social cognition is the ability to sharply distinguish individuals of one's species from one another. Thus, face recognition is not just a pleasant quirk in the system but an essential part of the biological capacity to play a role in society.

10.7.2 Reading

What forms of representation are involved in reading? Obviously some levels of the visual system must encode information from the page, and obviously the information provided by the visual system must eventually end up at some level in the language system. So the question is, Which levels?

Within the visual system it is clear that letters can be recognized and read regardless of their apparent size and, to a certain extent, regardless of orientation with respect to the reader. There are also tolerances with respect to letter shape, as seen by the effortless reading of new type fonts (within reason—German script, for instance, is relatively opaque to someone accustomed only to ordinary Roman letters). Above all, reading involves a recognition process—whether one is recognizing individual letters or larger units. These considerations suggest that the visual part of reading is a function of the 3D model level, where size and shape invariances can be specified. From the point of view of vision, then, letter and word recognition is a learned object discrimination task, not unlike, say, "automobile recognition": telling a B from an R is qualitatively similar to telling a Ford from a Buick.

The difference, of course, is that the objects discriminated in reading are translated into linguistic information. Again: At what level? This depends on the orthography. If it is an alphabetic script, like the Roman and Cyrillic alphabets, symbols of the orthography correspond more or less to phonological segments (there are plenty of exceptions, of course: *x* is phonetically *ks*, *th* is a single sound, often *e* stands for no sound at all; but segment-by-segment correspondence is the norm). In scripts like Hebrew and Arabic all consonants are represented in the orthography but vowels are omitted. In the Japanese *kana* script each symbol stands for a syllable. Whatever the differences among these scripts, they share the property of characterizing the *phonological* information of the language—how words are pronounced. By contrast, in Chinese orthography (much of which has been borrowed into Japanese as *kanji* script) and in ancient Egyptian hieroglyphics for the most part each symbol corresponds to a morpheme, regardless of its pronunciation. Here, then, the orthography is an indication of the *syntactic* or *semantic* level of representation.

This means that in learning to read, one must be establishing two things: a set of visual concepts that permit one to recognize elements of the orthography, and a set of correspondence rules between the orthography and the level of representation that the orthography symbolizes. In the case of morphemic scripts the correspondence is between the written symbol and the appropriate lexical item (perhaps at all of its levels), since one must learn such an orthography essentially one word at a time. In the case of a phonological script one may associate spellings with phonological representations in the lexicon (especially for idiosyncratic orthographies like that of English), but in addition there are general default principles for associating orthographic symbols with elements of the sound system, so that one can read ("sound out") words one does not know.

Thus, all levels of the visual system are implicated in the general in-

information is translated into the appropriate level of the linguistic system, whence it proceeds upward to the conceptual level. (I am not inclined to think there are any *new* levels of representation specialized for reading: how could one develop them?)

In processing, of course, information need not pass unidirectionally along this tortured route through the levels of representation. There can be the usual amount of top-down influence in perception—where this time "top-down" means down the informational path from conceptual structure through the linguistic system, thence into the visual system going down to at least the 2½D sketch. Think, for instance, of trying to read a messy handwriting, where one's identification of the letters is so strongly aided, even consciously, by one's knowledge of the language and one's guesses about what the writer must mean. And having read the scrawl, its segmentation into letters is visually obvious—the equivalent in reading of the phoneme restoration effect.

Interestingly, there is some neurological evidence for the distinction between phonological and morphological scripts. Tzeng and Wang (1983) report a number of studies in which reading and writing of phonological scripts is impaired by temporal lobe lesions, whereas reading and writing of morphological scripts is impaired by posterior, occipital-parietal lesions. Especially interesting is the case of Japanese, which intermixes the two kinds of script freely. According to Sasanuma (1974), deficits in reading and writing the two kinds of script can be independent of each other, each being associated with a different sort of aphasia. Here, then, is neurological evidence that the translation of visual symbols into linguistic forms varies with the kind of orthography, circumstantially supporting my claim that the two types invoke different types of correspondence rules.

10.7.3 Sign Language

American Sign Language (ASL) is the primary language of the deaf community in North America. It is a language independent of English ("signed English" imposes English word order and grammatical conventions on ASL vocabulary and is quite distinct), fully as expressive as any spoken language.

The acquisition of ASL by children is governed by principles similar to those for the acquisition of spoken languages. In particular, it is typically not taught to deaf children by (hearing) parents or teachers. Rather, until recently its use was actively discouraged by educators; the language is most often "picked up" from peers in the dormitories of schools for the deaf (Klima and Bellugi 1979, chapter 3). Moreover, sign language aphasias appear quite comparable to those in spoken language (Bellugi, Poizner, and Klima 1983).

Although it seems once to have been fashionable to claim that ASL has

no grammatical structure, recent investigation (for example, Klima and Bellugi 1979; Padden 1983; Supalla 1982; Gee and Kegl 1982; Newport 1982) has revealed a rich syntactic and morphological structure altogether comparable to that of natural languages; Elissa Newport (personal communication) finds its grammar not unlike that of Navajo. The difference, of course, is that instead of having a phonological structure that leads to the auditory and vocal modalities, ASL connects to the visual and gestural systems.

Again we can ask what levels are involved in this mixture of modalities. The evidence at the moment points to ASL certainly having a level of syntactic structure. Since we are regarding morphology as the word-interior aspect of syntax, the existence of ASL morphology fits in well here. On the other hand, there is no evidence for much beyond a rudimentary phonological level: there are words, and there are aspects that correspond to the suprasegmental information of stress and rhythm, but there is certainly no syllabic and segmental organization. Rather, at this point the information slips over into the visual-gestural modality, in which the usual criteria (object-centered descriptions, categorial recognition) implicate the 3D model representation. In ASL perception the 3D model will be derived via the lower visual levels; in production the 3D model will serve as input to the production of gesture, via whatever levels of representation are appropriate for that. One might hope, in fact, that the rich and yet contained system of action made use of by sign language could provide interesting evidence toward a theory of motor representation and of temporal segmentation in both vision and action.

The point of bringing up these specialized capacities, even if much too briefly and speculatively, is to suggest how accounts of them are to mesh with the primary theoretical construct of the present theory, the notion of levels of representation. To the extent that their information demands can be framed in terms of independently justified levels, this confirms the overall form of the theory. To the extent that such capacities can provide evidence for refinement of the theories of various levels, or suggest new sorts of connections among levels, or even suggest new levels, this too is useful. The overall goal, of course, is to keep the number of independent forms of representation small, not to have to invoke brand new levels for each task, and yet to recognize distinctions among levels when necessary.