On Beyond Zebra: The relation of linguistic and visual information*

RAY JACKENDOFF
Brandeis University

And I said "You can stop, if you want, with the Z
Because most people stop with the Z
But not me!

In the places I go there are things that I see
That I never could spell if I stopped with the Z.
I'm telling you this 'cause you're one of my friends:
My alphabet starts where your alphabet ends!

Dr. Seuss, On Beyond Zebra

Abstract

This paper addresses the problem of how the forms of information derived by the visual system can be translated into forms useable by the language capacity, so that it is possible to talk about what one sees. The hypothesis explored here is that there is a translation between the 3D model of Marr's (1982) visual theory and the semantic/conceptual structure of Jackendoff's (1983) theory of natural language semantics. It is shown that there are substantial points of correspondence through which the encoding of physical objects and their locations and motions can be coordinated between these two levels of representation, and that both of these levels are deeply involved in visual as well as linguistic understanding.

*This paper is excerpted from a work in progress entitled Consciousness and the Computational Mind. Much of it was drafted while the author was a Fellow at the Center for Advanced Study in the Behavioral Sciences in 1983–84, a year for which I will be forever grateful. I was encouraged in this work by discussions with Leonard Talmay, Lynn Cooper, Steven Pinker, Lucia Vaina, David Olson, Carol Krumhansl, James Cutting, Donald Hoffman, and Roger Shepard, among others. This research was supported in part by NSF Grants IST-8120403 and IST-8420073 to Brandeis University and by NSF Grant BNS-7622943 to the Center for Advanced Study in the Behavioral Sciences. Reprint requests should be sent to Ray Jackendoff, Program in Linguistics and Cognitive Science, Brandeis University, Waltham, MA 02254, U.S.A.

1. Introduction

One of the fundamental problems for a theory of natural language was stated most succinctly by John Macnamara in the title of his 1978 (unpublished) paper: "How Can We Talk About What We See?"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. The present paper develops such a point of connection, and shows that it not only helps answer Macnamara's question, but also solves certain outstanding problems in both theories.

I am interested in particular in psychological answers to Macnamara's question. So the following sort of solution, caricatured on Tarskian or Davidsonian semantics, will not do: Certain sentences are true of the world; speakers grasp the truth of these sentences by using their eyes, and thereby can warrantably assert them. In this sort of theory, the major problem is taken to be the explication of truth, a speaker-independent relation between language and the world. However, the psychological issue of what it means to grasp the relation is largely unaddressed. Barwise and Perry (1983) purport to address it by adverting to a primitive version of Gibsonian psychology of perception. But given that the possibilities look tremendously dim for constructing phonological and syntactic theories along Gibsonian lines, there seems little prospect of such an answer fitting into a unified theory of the language capacity.

Entering the realm of information-processing theories of mind, it is not enough to say, for example, 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 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. So the essential questions are: (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?

Let me be more specific about what I mean by a "form of information." One might think of each form as a discrete "language of the mind" (one of which is the "language of thought" in Fodor's, 1975, sense). More specifically, each form of information is a structured system of distinctions or oppositions, built up out of a finite set of primitives and principles of combination. More standard terms with essentially the same sense are level of representation as used in linguistics, Kosslyn's (1980) format, and Marr's (1982) stage of representation. The language faculty, for instance involves three distinct forms of information—phonological, syntactic, and semantic/conceptual. The phonological primitives include the phonological distinctive features, the notions of segment, syllable, and word, the elements of stress and timing systems, and
the elements of intonational patterns. The principles of combination include the combination of a set of phonological distinctive features into a segment (or speech sound), the concatenation of segments into syllables and words, and the principles for associating stress, timing, and intonation with strings of segments.

Similarly, the syntactic primitives include the syntactic categories (Noun, Verb, etc.), the phrasal categories (S, NP, etc.), and the elements of inflectional systems such as case, gender, number, tense and aspect. The principles of combination include the domination of one node by another, the linear ordering of nodes, and the elements of syntactic transformations (or alternative devices for capturing long-distance dependencies). We will discuss semantic/conceptual primitives and principles of combination in Section 3.

The overall logical organization of information used by the language faculty is shown in (1).

(1) *Organization of language*

```plaintext
auditory information

phonological formation rules

phonological structures

syntactic formation rules

syntactic structures

conceptual formation rules

semantic/conceptual structures

motor information
```

The primitives and principles of combination for each form of information are collectively labeled as “formation rules” for that level. In addition, though, there must be a set of principles for translating from acoustic information to meaning in order to accomplish language understanding, and similarly from meaning to vocal tract instructions in order to accomplish speech. These principles of translation are encoded as sets of “correspondence rules” that link adjacent levels; they are notated in (1) as arrows and double arrows between structures.

I should stress that (1) is the *logical organization* of the information, not a model of language processing. It is consistent with many different theories of the time-course of processing, as long as they respect the formal autonomy of the three kinds of linguistic information. For instance, it is consistent with a parallel processing system in which translation from acoustic form to phonological structure is taking place for one part of a sentence, while an
earlier part of the sentence is already undergoing translation from syntactic to semantic form. (Jackendoff (forthcoming, chapter 6) goes over various possibilities in detail, demonstrating the distinction between the theory of the logical organization of structure and the theory of the time-course of processing. The point is not new, but it is still widely misunderstood.)

The need to explain the function of the visual system in terms of the forms of information it derives is most explicitly articulated in the work of Marr (1982). (Two good introductions are Marr and Nishihara, 1978, and Pinker, 1984.) Marr calls such a description a computational theory, a theoretical prerequisite to the descriptions of processing (algorithmic theory) and neurological implementation. According to Marr’s theory, information transduced by the retinas must be translated successively into three distinct forms before object recognition can be achieved: the primal sketch, which encodes the local organization of boundary elements in the visual field; the $2\frac{1}{2}$D sketch, which encodes the geometry of the surfaces visible to the viewer; and the 3D model, which encodes the spatial configuration of viewed objects in viewer-independent terms.

The overall logical organization of Marr’s theory of visual information is given in (2).

\begin{itemize}
  \item \textbf{Organization of vision (Marr)}
\end{itemize}

\begin{center}
\begin{tikzpicture}
  \node[draw] (retinal) at (0,0) {retinal array};
  \node[draw] (primal) at (3,0) {primal sketch};
  \node[draw] (2_2D) at (6,0) {$2\frac{1}{2}$D sketch};
  \node[draw] (3D) at (9,0) {3D model};

  \node[draw] (formation) at (3,-2) {formation rules};
  \node[draw] (formation) at (6,-2) {formation rules};
  \node[draw] (formation) at (9,-2) {formation rules};

  \draw[->] (retinal) -- (primal);
  \draw[->] (primal) -- (2_2D);
  \draw[->] (2_2D) -- (3D);

  \draw[->] (1.5,0) -- (1.5,-2);
  \draw[->] (4.5,0) -- (4.5,-2);
  \draw[->] (7.5,0) -- (7.5,-2);
\end{tikzpicture}
\end{center}

Again, the formation rules characterize the formal elements out of which levels are built (for instance, “blobs,” “bars,” “terminations” in the primal sketch, “regions” and “contours” in the $2\frac{1}{2}$D sketch); the correspondence rules characterize the principles by which information of one form is translated into the next.

The problem of how we talk about what we see can be understood more clearly in terms of diagrams (1) and (2). What is necessary for this task to be possible at all is a set of correspondence rules linking forms of information in the two faculties. Through these correspondence rules, visual information can be translated into a form suitable to be couched as a linguistic expression. The hypothesis to be pursued here is that the desired correspondence rules are to be stated between semantic/conceptual structure and the 3D model,
The relation of linguistic and visual information

i.e. between linguistic meaning and "visual meaning"—the mind's reconstruction of "what is out there in the world."

From the point of view of the semantics of natural language, this possibility is tremendously exciting. For if a connection can be forged between semantic/conceptual structure and the information derived by the visual system, semantics will not have to back up into ever murkier levels of "interpretation" to explain our ability to talk about the visually perceived world. Rather, there will be a complete succession of levels of representation from the retina all the way to the vocal tract—a totally computational theory of how we talk about what we see. Information can pass up through the forms encoded by the visual system, be translated from 3D model organization into semantic/conceptual organization and proceed down the linguistic forms into speech.

The next two sections sketch some properties of the 3D model and semantic/conceptual structure; we then turn to the relation between them.

2. Properties of the 3D model

Marr explicitly designs the 3D model to be appropriate for encoding an object's shape in long-term memory, so that it may be recognized on subsequent occasions. He develops representations only for single objects, not for the complete configuration in the visual field; we will propose some extensions in Section 6.

There are three salient characteristics of the 3D model. First, it is volumetric: objects are represented as occupying volume in space, by contrast with the surface representations of the lower levels. Second, it is object-centered: it makes explicit the shape and size constancy of objects, regardless of the viewer's position, by contrast with the viewer-centered representations of the lower levels. Third, it is hierarchical: it represents the 3-dimensional structure of objects not just in terms of holistic shape (i.e., it is not a "statue in the head"), but rather in terms of a hierarchical decomposition of the object into parts and parts of parts.

For example, consider diagram (3), from Marr and Nishihara (1978), which suggests the organization of the 3D structure for a human figure. At the coarsest layer of description, the figure is represented simply by a cylinder, itself defined by a vertical axis down the middle and a cross-section. At the next layer of description, the cylinder is elaborated into a torso, a head, and four limbs, each of which is a cylinder defined by its own axis. The points of attachment of the head and limbs to the torso, and the angles that they form, are specified in terms of the coordinate system defined by the axis of the torso. In turn, each of the parts is subject to elaboration at finer levels
(3) A sample 3D model structure

of detail. For instance, the arm cylinder can be elaborated as two connected cylinders corresponding to the upper and lower arm; the lower arm can be elaborated into lower arm and hand; the hand can be elaborated into palm and fingers, the fingers into joints. Thus the 3D model is a sort of tree structure.

In each case of elaboration, the configuration of parts is specified in terms of the coordinate system defined by the axis of the main part in the configuration. Thus the positions of the fingers, for instance, are specified most directly in terms of their configurations within the hand. Their position with respect to the body as a whole is specified only indirectly, through the hand’s position on the arm and the arm’s on the torso.

This, then, is what Marr means by saying the description is object-centered: through recursive elaboration, the positions of the parts of an object are specified ultimately with respect to the main axes of the object as a whole, and can therefore be specified without mention of the viewer’s position. Of course, in integrating the object as a whole into the scene, its position with respect to the viewer must be specified. But this can be determined entirely by specifying the position of the object’s main axis, and the positions of the parts will fall into place automatically.¹

¹There seems to be no bar to locating the parts directly with respect to the viewer, but in the optimally hierarchical description, the one that permits identification of the object as a whole, this will be redundant. However, such description in the 3D model may well be used on occasion, either for particular tasks or for dealing with unfamiliar objects. For example, this appears to be a plausible account of Jolicoeur and Kosslyn’s (1983) experiments, in which subjects seemed sometimes to use viewer-centered descriptions for identification of rather bizarre objects. Jolicoeur and Kosslyn imply that subjects are storing a 2½D sketch. Closer attention to the formal details of Marr’s theory of the 2½D sketch would show that this is unlikely: recall that there is no notion of object in the 2½D sketch, through which identification could take place.
Though the formal characteristics of the 3D model representation are at present far from well understood, the general outlines are fairly clear. The primitive units of the representation are the coordinate axes and the means for generating simple shapes (such as the cylinders of (3)) around them. The principles of combination provide the means to combine simple shapes into more complex shapes. From (3) we can see that the principles of combination essentially must provide a way to elaborate a description of an object from a coarse description to the next higher degree of specificity, for example from the single cylinder to the torso plus head and limbs. Further layers of description will be possible simply by applying the principles of combination recursively.

Note that the principles are not of the form “such-and-such a part is connected to such-and-such a part.” Rather, the dispositions of the parts are specified in relation to the coordinate axes of the next coarser layer of description: “(the axis of) such-and-such a part at layer of description $L_i$ is disposed in such-and-such a configuration with respect to the axes of such-and-such an element of layer $L_{i-1}$.” Principles of this form are a great deal like the phrase structure rules of syntax, in which, for example, a relative clause is understood as an expansion not of the head noun, but of the next larger unit, the noun phrase.

It is moreover possible to discern a notion of “head” not unlike that in syntax. Just as one talks of the head noun—the principal element of a noun phrase, the one that determines the noun phrase’s categorical status—one can in many cases speak of the head constituent of a 3D model elaboration: that subsidiary element in an elaboration whose axes and volumetric form are identical to (or most similar to) the coarser-layer element being elaborated. In the human figure, for instance, the torso is the head of the initial elaboration; the palm is the head of the hand. (In other cases, though, such as the elaboration of the arm into upper and lower arm, the presence of a head constituent is less obvious. Perhaps there is an analogy here to coordination in language.)

Now consider the task of identifying an object that can change its shape, for example a human or animal. The long-term memory 3D model of such an object, to which its presented appearance can be compared, will not represent just a single pose of this object. Rather, the parameters specifying angle of attachment of the various parts in a decomposition like (3) will give a range of possible values, corresponding to the possible configurations of the joints. On the other hand, certain aspects will be kept rigid, for example the point of attachment. (If somehow the shoulder joint could slide up and down the torso, the way the carriage slides across a typewriter, that too could be specified by a variable.) Thus the object-centered decomposition of objects...
makes possible a description of an object’s possible variations in shape.

Similarly, the 3D model permits an intuitively attractive account of the representation of the action of moving figures, as Marr and Vaina (1982) point out. For instance, in describing the action of walking, the 3D model can specify the angular changes in each of the joints, and relate these changes in a temporal framework. Moreover, a rather natural segmentation of actions arises at points in time when joints start or stop changing in angle, or when angular change reverses direction. For instance, the swing of a leg in walking is bounded at its beginning and end by points of stasis in the change of angle at the hip joint. (This corresponds nicely to observations of Cutting, 1981 on the segmentation of events.) Notice that such specifications can be made only in an object-centered framework, since it is the angular change measured “objectively” at the hip that is significant, not the angular change in the visual field. Thus not only for object identification but for the description of action, an object-centered description along the lines of Marr’s seems a necessity.

A further move of abstraction permits the 3D level to encode variations in form among individuals of a common type. For instance, consider how a long-term representation might encode what horses look like. By allowing a range of permissible sizes, both overall and constituent-by-constituent, and by specifying ranges of proportions of size among parts, one can arrive at a geometrically parametrized encoding that can be matched with a class of presented objects more variable than a specific individual.

Notice how this conception differs from the widespread hypothesis that one stores a mental image of a stereotypical instance of a category. As often conceived, such an image is of a particular individual—a comparatively rigid template against which putative instances are matched by some unspecified computational process of “judging similarity”. The 3D model, by comparison, is decompositional and hierarchical, and thus capable of greater generality than a rigid template, in that the elaboration of the object into parts can specify proportions and tolerances rather than a single fixed value for the size, shape, and disposition of the parts. (The phenomenon of an experienced image of a prototypical instance probably arises in part through fixing typical or focal values for the variable parameters.) Moreover, the 3D model representation is not an arbitrary stipulated representation, but one motivated and developed on the grounds of its general adequacy for visual form recognition tasks.

The mention of stereotypes brings us into territory familiar from discussions of linguistic meaning. To put the issue into context, it is necessary first to make some remarks on the form of semantic/conceptual structure.
3. Outline of Conceptual Semantics

*Semantics and Cognition* (Jackendoff, 1983, henceforth *S&C*) develops a theory of the semantics of natural language called Conceptual Semantics, which answers to a number of demands not normally considered central by theories within philosophy, psychology, and AI. In particular, it is important that meanings be mapped perspicuously into the syntactic form in which they are expressed (the Grammatical Constraint) and that they be suitable for psychological purposes besides the interpretation of language (the Cognitive Constraint).

Among the leading points of this theory are the following:

1. **Meanings are mentally encoded.** This is the foundational assumption of Conceptual Semantics, necessary to embed the theory within a psychological framework. It is this assumption that makes the notion of a speaker grasping a meaning contentful. For example, it decisively rules out standard model-theoretic approaches, in which meanings are taken to be sets of individuals in possible worlds, extrinsic to the language user rather than encoded in the user’s mind.

2. **Meanings are decompositional.** That is, meanings have internal structure built up from a finite, innate stock of primitives and principles of combination. The argument for this point (*S&C*, chapters 5 and 7) is essentially built on the creativity of conceptual form, and in many respects it parallels Chomsky’s (1957) arguments concerning the creativity of syntax. This argument rules out, among other things, semantic network theories, in which meanings are unstructured nodes in a finite network of associations or inferences, and the monad-meaning postulate theory of Fodor, Garrett, Walker, and Parkes (1980).

3. **Meanings do not, however, decompose into necessary and sufficient conditions.** The theory acknowledges the numerous arguments that concepts have fuzzy borderlines and family resemblance properties. *S&C*, chapter 8, develops and defends in detail appropriate formal devices that incorporate such phenomena integrally into the possible decompositions of word meanings.

4. **There is no formal distinction of level between semantics and pragmatics.** It is shown that once one has the primitives and principles of combination appropriate for nonlinguistic tasks such as object categorization, the same machinery will account for many “purely semantic” tasks such as linguistic inference. It is concluded that there is no reason to posit a separate “semantic” component that deals exclusively with inference (*S&C*, chapter 6).
The formation rules for semantic/conceptual structure include, among other things, a vocabulary of primitive conceptual categories or "semantic parts of speech." Among these categories are such entities as Thing (or Object), Event, State, Action, Place, Path, Property, and Amount (S&C, chapter 3). Here are a number of the formation rules for expanding such basic categories into more complex expressions:

(4) (a) \( \text{PLACE} \rightarrow [\text{Place PLACE-FUNCTION (THING)}] \)

(b) \( \text{PATH} \rightarrow \begin{cases} \text{TO} \\ \text{FROM} \\ \text{TOWARD} \\ \text{AWAY-FROM} \\ \text{VIA} \end{cases} \) \( \{ \{\text{THING} \} \} \)

(c) \( \text{EVENT} \rightarrow \{ [\text{Event GO (THING, PATH)}] \\ [\text{Event STAY (THING, PLACE)}] \} \)

(d) \( \text{STATE} \rightarrow \{ [\text{State BE (THING, PLACE)}] \\ [\text{State ORIENT (THING, PLACE)}] \} \)

Let me briefly explain these rules. (4a) says that a conceptual constituent of the basic category Place can be expanded into a Place-function plus an argument of the function which is of the category Thing. The argument serves as a spatial reference point, in terms of which the Place-function defines a region. For example, in the expression *under the table, the table* designates a reference object, and *under* expresses a Place-function which maps the table into the region beneath it. (4b) similarly expands a Path, or trajectory, into one of five functions that map a reference Thing or Place into a related trajectory. An example of a Path with a reference Thing is *to the house*; a Path with reference Place is *from under the table*, where the trajectory begins at the Place "under the table."

(4c) says that a constituent of the category Event can be expanded into either of the two Event-functions GO or STAY, each of which takes two arguments. The arguments of GO, which denotes motion along a path, are the Thing in motion and the Path it traverses. This structure is seen most transparently in a sentence like *Bill went to New York*. The arguments of STAY, which denotes stasis over a period of time, are the Thing standing still and its Location, as seen in *Bill stayed in the kitchen*, for instance. (4d) gives two expansions of State; the first is used for specifying the location of objects (*the dog is in the park*), and the second for specifying the orientation of objects (*the sign points toward New York*).

The relation of syntactic and conceptual constituent structure can be illus-
The relation of linguistic and visual information

Treated for the simplest sort of case with a sentence like (5a), whose conceptual structure appears in (5b).

(5)  
(a) Syntactic structure  
\[ s \ [NP \ \text{John}] \ [VP \ \text{ran} \ [PP \ \text{into} \ \[NP \ \text{the room}]]] \]

(b) Conceptual structure  
\[ \text{Event} \ \text{GO} \ ([\text{Thing} \ \text{JOHN}] ; \ [\text{Path} \ \text{TO} \ ([\text{Place} \ \text{IN} \ ([\text{Thing} \ \text{ROOM}] ) ]])] \]

The sentence corresponds to the entire Event in conceptual structure. The verb corresponds to the Event-function GO, that is, this is a sentence expressing motion. The subject corresponds to the first argument of GO, and the PP corresponds to the second argument. This second argument is composite: the Path function TO takes a Place as its argument, and the Place in turn decomposes into the Place-function IN and a Thing expressed by the object of the preposition.

In order to see how (5b) is put together from its parts, it is necessary to look at the lexical entries for the two words in (5a) that have argument structure. In the entries in (6), the first line is the phonological structure; the second is the syntactic category; the third is the syntactic contextual feature (strict subcategorization); the fourth is the conceptual structure. Thus each lexical entry can be thought of as a small-scale correspondence rule for the three levels of linguistic structure.

(6)  
(a)  
\[ \text{Prep} \ mL.\] \[ -- \ [NP] \]
\[ -- \ [\text{Path} \ \text{TO} \ (\text{Place} \ \text{IN} \ ([\text{Thing} \ i])])] \]

(b)  
\[ \text{Verb} \ mL.\] \[ -- \ [PP] \]
\[ -- \ [\text{Event} \ \text{GO} \ ([\text{Thing} \ i] , \ [\text{Path} \ j])] \]

*Into* requires an NP object, which is coindexed with the argument position, the Thing, in conceptual structure. *Run* is slightly more complicated. Semantically, it requires two arguments, the Thing in motion and the Path that specifies the trajectory of motion. The first is indexed *i*, which we will take by convention to indicate subject position. The second argument is filled in with the reading of the PP following the verb. If no PP is syntactically present, the Path is simply unspecified: *John ran* means in part “John traversed some (unspecified) trajectory.” In other words, the well-formedness conditions on conceptual structure require this argument even if it is not expressed.

A similar conceptual structure can be expressed in different syntactic form, for example by a sentence like (7).
Here *enter* is an optionally transitive verb, with a lexical entry like (8).

(8) \[
\begin{array}{l}
\text{enter} \\
\text{Verb} \\
= \{ \text{Event GO ([Thing iI], [Path TO ([Place IN ([Thing j])])])} \}
\end{array}
\]

This verb incorporates into its meaning the Path- and Place-functions expressed separately by the preposition *into* in (5a). Notice that the intransitive version, *John entered*, means not just “John traversed some path,” but “John went into something.” That is, the sense of *into* appears even when the second argument is unspecified.

These examples show the elementary properties of the mapping between syntactic and conceptual structure. Words in general need not correspond to complete conceptual constituents—they correspond to constituents with open argument places. The argument places, however, may be embedded two or more functions down into the conceptual constituent, as seen in (6a) and (8); the only stipulation is that they themselves be full constituents. (This is codified as the “Lexical Variable Principle” in S&C, section 9.5.)

One last aspect of the formalism of Conceptual Semantics needs to be mentioned. The *type-token distinction* is formalized as a simple binary feature TYPE vs. TOKEN, applied to concepts that are otherwise formally similar. In this respect the theory differs from, e.g., predicate calculus, in which tokens are encoded as constants and types as predicates—formal expressions of entirely different type. Within the Conceptual Semantics formalism, three basic predication relations emerge as fundamentally similar:

(9) (a) Token-identity (e.g. *Clark Kent is Superman*)

\[
\text{IS-TOKEN-IDENTICAL-TO} \left( \begin{array}{c} \text{TOKEN}_X \\ \text{TOKEN}_Y \end{array} \right)
\]

(b) Categorization (e.g. *Rover is a dog*)

\[
\text{IS-AN-INSTANCE-OF} \left( \begin{array}{c} \text{TOKEN}_X \\ \text{TYPE}_Y \end{array} \right)
\]

(c) Category inclusion (e.g. *A dog is an animal*)

\[
\text{IS-INCLUDED-IN} \left( \begin{array}{c} \text{TYPE}_X \\ \text{TYPE}_Y \end{array} \right)
\]

The similarity in form among these conceptual structures parallels their similarity in linguistic expression, namely *NP is NP*. Moreover, it is shown that the three functions, IS-TOKEN-IDENTICAL-TO, IS-AN-INSTANCE-OF,
and IS-INCLUDED-IN, are at least closely related if not virtually identical—as revealed in the use of the same verb be to express all of them. (This treatment of predication is developed in detail in S&C, chapters 5 and 6.)

4. Preliminary points of correspondence

We are now in a position to begin to work out some principles of correspondence between the 3D model and conceptual structure. It is important to bear in mind that the former was motivated by the demands of the visual system, and the latter by the demands of language. One might not anticipate that any relationship between the two need exist. Any relationship that is found, therefore, is an unexpected windfall, and a vindication of the claims of both representations to psychological reality.

The first point of contact, of course, is the notion of physical object as a significant unit of both levels, though encoded in quite different fashions, to be sure. Thus a basic principle of correspondence maps constituents of the 3D level that encode objects into constituents of the conceptual level that encode objects; there is no other pair of levels between which this fundamental correspondence could be stated.

Another important correspondence that can be established is in the encoding of the relation of part-whole or inalienable possession. The 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 in structural terms. Basically, for physical objects X and Y, X IS A PART OF Y obtains in conceptual structure just in case the 3D object representation corresponding to X is an elaboration within the 3D object representation corresponding to Y.

Next, 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’re 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 somewhat 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; while one may on the other hand be very knowledgeable about the appearance of a very precisely 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 a little 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 an individual being perceived and a 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 "Höfding 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 identification is encoded conceptually by the relation IS-TOKEN-IDENTICAL-TO, an algebraic relation between two TOKEN concepts. Object categorization, however, is encoded by the relation IS-AN-INSTANCE-OF, a relation between a TOKEN and a TYPE. The overall forms of the two relations are given in (10); the vertical lines indicate associations or linkages between representations at the two levels.

(10) (a) Object identification

| conceptual level: IS-TOKEN-IDENTICAL-TO ([TOKEN]i, [TOKEN]j) |
| level: |
| 3D level: |
| visually derived |
| 3D model from memory |
The relation of linguistic and visual information

(b) Object categorization

conceptual IS-AN-INSTANCE-OF level: ([TOKEN]_i, [TYPE]_k)

3D level: visually 3D model derived from 3D model memory

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.

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 had cropped up occasionally in the literature. For instance, Marr and Vaina (1982) discuss how a few “basic action types” such as throwing, saluting, walking, etc. 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.

How are action categories to be encoded? Here is the problem, which in its essence goes back at least to the British empiricists. A 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 not available in the geometric representation—but it is available in conceptual structure, where we have the algebraically structured features that distinguish primitive conceptual categories. By linking the 3D figure in motion to an ACTION TYPE concept rather than to a THING 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 again a linkage of 3D and conceptual structures provides the right range of distinctions.

5. 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 to help encode 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.

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 a perceived 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 \([\pm \text{LONG NECK}]\) 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 distinctions in 3D model format, where they are not ad hoc at all, but precisely what this level of representation is designed to express.

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. 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 2, the 3D model provides a much more coherent account of what lies behind such an intuition that 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, verbs such as \(\text{walk, run, jog, lope, and sprint}\) differ from each other in much the same way as \(\text{duck, goose, and swan}\). 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 is this to 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.

While 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. Some of the essential algebraic features of conceptual structure were shown in the previous section 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.

6. 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 which play a de-
monstrable 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 sentences concerning spatial location and motion.

It has often been noted (Clark and Chase, 1972; Gruber, 1965; Jackendoff, 1976; Langacker, 1983; Miller and Johnson-Laird 1976; Talmy, 1983) 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 (11a), the table is the reference object and the book is figural. And it is intuitively clear that there is a distinct difference between (11a) and (11b), where the roles of reference object and figure have been exchanged.

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

The usual examples illustrating sentences of spatial relation, like (11), 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 (12).

(12) The book is
\[
\begin{array}{c}
\text{standing} \\
\text{lying} \\
\text{leaning} \\
\text{resting}
\end{array}
\] 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, while 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 (Herskovits, 1985; Jackendoff, 1978, 1983, chapter 9; Talmy 1983) is to make use of the formalism of Section 3, in particular the conceptual category Place. As seen there, location sentences like (11) and (12) 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 (13), paralleling (4d).

(13) \[\text{StateBE ([ThingBOOK], [Place ON ([ThingTABLE]])])}\]

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, \textit{in} expresses a function that (for a first approximation) maps an object into the region consisting of its interior. \textit{On} maps an object into the region consisting of its surface (in many cases, its upper surface). \textit{Near} maps an object into a region exterior to but contiguous with the object. \textit{Beside} is like \textit{near}, but restricts the region to roughly horizontal contiguity (a cloud above a mountaintop may be \textit{near} it; it cannot be \textit{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 \textit{beside} or \textit{along} a road, but one can only be \textit{beside}, not \textit{along}, a tree (unless the tree has been felled). Evidently the domain of the function expressed by \textit{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 \textit{in front of}, \textit{behind}, \textit{on top of}, and \textit{to the right of}. These get 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 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 \textit{in front of the house}, \textit{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 \textit{behind the sphere} picks out a region contiguous to the sphere and on the side of it not facing us.

As has been frequently noted (Clark, 1973; Olson and Bialystok, 1983; Talmy, 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 (14), 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).

\[ \text{front door} \]

(14) speaker → ○ house y x

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).

One of the points of 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.

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 phrase arguments. Consider the possibilities in (15), for instance.

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

As in the location sentences (12), the differences among the verbs reflect the internal dispositions and motions of the parts of John’s body, that is, they express the object-centered description of John himself. As observed in Section 5, 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 (15) 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) travelling along some externally specified trajectory. Thus the total description can be seen as hierarchical, the outer layer being the external trajectory of the object, the inner layer being an object-centered elaboration of the object-internal motion. Talmy (1980) points out that in some languages, for instance 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).

The Marr theory as it stands does not include any notion of “trajectory traversed by an object.” The object-external part of the motion appears in conceptual structure as an Event constituent of the sort in (5b):

\[
\text{(16) \quad [Event GO ([Thing \ JOHN], [Path \ TO ([Place \ IN ([Thing \ ROOM]))]))]}
\]

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 of 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.

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.

---

2On the other hand, there are verbs in English that express object-external motion, as in (i):

(i) John circled the tree.

The fly spiraled down to the table.

Thus the division is not grammatically rigid; nonetheless, it manifests itself in fairly clear semantic intuitions that in turn are often reflected grammatically.
for this case. (There are exceptions to these conditions, but I think the basic principles stand. This description may be thought of as “stereotypical across.”)

It is not difficult to imagine extending Marr’s notion of the 3D model to include such information as Paths. Just as with Places, in the cases where Path-functions make use of the geometry of the reference object, this seems 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.

7. Summary and methodological remarks

From this very preliminary investigation of the correspondence between conceptual structure and 3D model structure, then, a number of points emerge. First is that the two structures can indeed be brought into contact, and that 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, and action.

Second is that the theory of Conceptual Semantics developed in S&C and summarized in Section 3 contains many of the right sorts of elements to interface with spatial understanding. This is evidence that it is on the right track towards a properly mentalistic theory of the semantics of natural language, and in particular towards an answer to the question of how we can talk about what we see.

Third is that 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 which are already made available by the Marr theory, and which turn out on reflection to be
supported by the organism’s performance at nonlinguistic tasks. These enrichments moreover 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, while 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 information that one cannot 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 which one cannot hear and that one does not speak, but which 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.

I want to conclude by replying to two methodological objections raised by Zenon Pylyshyn in response to the present proposal. His first point is that it may be too premature in the development of linguistic and visual theory to responsibly attempt to connect them. While I agree with Pylyshyn that the studies of conceptual structure and the 3D model are both still in their infancy, I do not see this as reason to abstain from approaching their relationship. The necessity to translate back and forth between the two forms of structure is an important boundary condition on the adequacy of both theories, and thus each can provide a source of stimulation for the other—as we have already seen in the present study. It seems to me, in fact, that the

---

3These points were raised at the Conference on Mental Representation in Vancouver in February 1986, sponsored by Simon Fraser University and the University of British Columbia.
sooner this boundary condition can be brought to bear, the better, for it can help prevent false moves at fundamental choice points in the development of both semantic and visual theories.

Pylyshyn's second point stems from his longstanding (1973, 1980) opposition to theories that require image-like representations. (It should be noted that some of his standard objections do not pertain to the 3D model. For instance, the 3D model is highly structured, and well-defined computations can be carried out over it, so it is not the “holistic” sort of representation to which he rightly objects.)

The basic issue is one of parsimony: if all concepts can be expressed in algebraic (or propositional) form, why posit an additional geometrical level of representation? The use of 3D model structures for encoding long-term memory would seem to be redundant at best.

There are various levels of reply. Most superficially, it has not to my knowledge been demonstrated that all visual concepts can indeed be encoded propositionally. Moreover, by now it should be unquestionable that information cannot be mapped directly from retinal configuration to full object constancy, expressed propositionally. Rather, it seems inevitable that visual perception employs some intermediate stage(s) of encoding with geometric properties. Thus it would appear difficult, if not impossible, to dispense with geometric levels of representation altogether. In particular, Pylyshyn offers not a hint of how to replace the geometric representations of Marr's theory with something he finds more palatable.

This line of reply to Pylyshyn is rather brute force, in that it simply denies that propositional representations alone are adequate for accounting for how we talk about what we see. But the issue of parsimony itself deserves a closer look as well. Interestingly, the argument closely parallels the debate over Generative Semantics in the early 1970s. Postal (1972), like Pylyshyn, argued that the best theory is the one with the fewest levels in it. In particular, a theory of language that coalesced syntactic and semantic information into a single level (Generative Semantics) was to be preferred to a theory that had separate syntactic and semantic levels (Interpretive Semantics).

Chomsky's (1972) reply to Postal was that one does not measure the value of theories in terms of raw numbers of components, but in terms of their explanatory power. For instance, suppose that in order to describe some phenomenon in a one-component theory one must introduce ad hoc devices that perforce miss generalizations, whereas a two-component theory expresses these generalizations naturally—and suppose this situation arises time after time. In such a case, the two-component theory is clearly preferable. Chomsky cites numerous linguistic phenomena that cannot be described in their full generality unless one adopts a theory that separates syntactic from
semantic structure. It was arguments of this sort that led to the demise of Generative Semantics.

A reply much like Chomsky's can be raised in response to Pylyshyn. We have found here a natural division of labor between conceptual structure and the 3D model. There are algebraic characteristics of concepts that are totally inexpressible in the 3D model, for example the type-token distinction and the distinction between objects and actions. On the other hand, there are geometric characteristics of concepts that are at best blatantly ad hoc in conceptual structure, for example the distinction between the necks of ducks and geese and the distinction between jogging and loping. By sharing the encoding of spatial concepts between the two levels of representation, each level only has to do what it is good at; neither need be cluttered with ad hoc descriptive devices. Moreover, we have found substantial points of correspondence through which the two levels can be coordinated. This seems to me a promising beginning for an explanatory theory of spatial thought and its linguistic expression. The message of the present paper, then, is that in pursuit of such a theory it is necessary to find a synthesis that transcends the boundaries of pure linguistic theory and pure visual theory—in short, to go On Beyond Zebra.

References


Macnamara, J. (1978). How can we talk about what we see? unpublished mimeo, Department of Psychology, McGill University, Montreal.


**Acknowledgment**

Figure (3) is reprinted, by permission, from D. Marr and H.K. Nishihara, Representation and recognition of the spatial organization of three-dimensional shapes, *Proceedings of the Royal Society of London, B200* (1978), 269–294.

**Résumé**

Cet article étudie le problème de la traduction de l’information visuelle en de l’information utilisable par la faculté de langage, c’est-à-dire la question de savoir comment il est possible de parler de ce que l’on voit. Notre hypothèse est qu’il existe une traduction entre le modèle 3D de la théorie de la vision de Marr (1982) et la structure sémantique/conceptuelle de la théorie sémantique de Jackendoff (1983). Nous montrons qu’il existe de nombreuses correspondances par lesquelles le codage des objets physiques et de leur emplacement et déplacement peut être coordonné entre ces deux niveaux de représentation, et que ces deux niveaux jouent un rôle fondamental dans la compréhension aussi bien visuelle que linguistique.