The Structure Grounding Problem

Michael Gasser
Departments of Computer Science and Linguistics
Indiana University
Bloomington, IN 47405
gasser@cs.indiana.edu

Abstract

Work on grounding has made a start towards an understanding of where simple perceptual categories come from. But human concepts are made up of more than the simple categories of these models; concepts have internal structure. Within the visual/spatial domain, it is necessary to go beyond an account of how "square" and "above" are grounded to an account of how "here is a square above a circle which is to the left of a triangle" is grounded. Conceptual/linguistic structure is not just arbitrary patterning which falls out once the object and relation categories have been identified. Rather, it reflects fundamental aspects of the perception of objects and relations. Thus there is a need to ground the structure as well as the categories which make up concepts.

Grounding and Structure

Since Harnad's important paper on the need to ground symbols in perception or action (Harnad, 1990), there have been various attempts to train systems to categorize visual inputs in terms of noun or noun-like categories or simple one- or two-place relations (Dorffner, 1990; Nenov, 1991; Regier, 1992). While the progress made by these researchers should not be minimized, grounded categories by themselves are only the beginning. One of the hallmarks of human concepts is their **structure**. On the traditional view, symbols combine to form an effectively unlimited set of possible symbol structures, which are interpretable by virtue of the context-freeness of the symbols and the small number of rules of composition.

There are at least two ways in which conceptual structure may relate to the symbol grounding problem.

1. As suggested by Harnad (1990), once symbols are grounded, it is the job of the symbol system to see that they are arranged in ways that make the results meaningful. Symbol structure is a purely symbolic matter.
2. The structure of concepts, like the symbols that are combined in the structures, gets its significance from the learned or wired-in connections between perceptual phenomena and structural distinctions. Structure is grounded.

My purpose here is to argue for the second alternative.

The Structure of a Scene Description

I will be considering only the limited case of what is involved in generating a description of a simple scene, such as that shown in Figure 1.

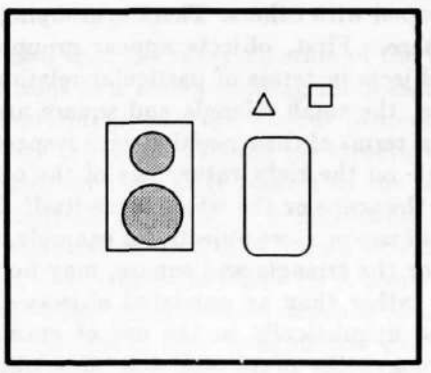

Figure 1: A Scene to Describe

One way in which a speaker might choose to describe this scene is as follows:

(1) In the middle of the scene are two rectangles. (2) Inside the one on the left are two circles. (3) They are more or less the same color. (4) The rectangle on the right has two small figures above it. (5) The one on the left is a triangle, (6) and the one on the right is a

square. (7) The rectangle on the right also has rounded corners.

At one level, this description can be viewed as structured in terms of what each sentence is "about", or what is relatively given, and what it introduces that is new. I will refer to the more given element somewhat sloppily as the "topic", though givenness and topicality are probably not the same thing. In the first sentence, the scene, taken to be given, is used as a setting to introduce the rectangles. The rectangles, in the form of one or the other of the members of the pair, become the topic of sentences (2), (4), and (7). Sentences (2) and (4) also introduce new objects which the succeeding sentences take as their topics. At the level of the entire description, there is a hierarchical structure, defined in terms of the topics which are associated with sets of sentences.[1] Note how the use of *also* in the last sentence reflects this structure; this word says, in effect: "now I'm saying something more about the rectangle on the right." The distinction between given and new also plays a role in the structure of individual clauses; the topic appears early in the clause, the new information later on. One possible way to represent the structure of the description is through a directed acyclic graph like that shown in Figure 2. The sequential nature of the description is indicated by the thick arrows, and the height of the nodes representing the assertions in the description reflects the hierarchical structure.

A further aspect of the structure of the description concerns the way in which particular objects are grouped with others. There is grouping of two kinds here. First, objects appear grouped with other objects in terms of particular relations. For example, the small triangle and square are introduced in terms of their position with respect to the rectangle on the right rather one of the other figures in the scene or the whole scene itself. Second, groups of two or more objects, for example, the two circles or the triangle and square, may be treated as sets rather than as unrelated objects. This is reflected linguistically in the use of grammatical morphology: the plural and the noun phrase *the one*.

[1]There is nothing new about these claims; the hierarchical nature of discourse is familiar in discourse functional linguistics and in the natural language processing literature concerned with discourse comprehension and production.

Scene Description and Perception

How does the description relate to the scene described? It seems clear how each of the lexical items in the description (e.g, *rectangle*, *middle*, *above*, *corner*) corresponds to some aspect of the scene. But the structure of the description, in particular, the ordering of the constituents and the sentences and the way in which the topic shifts, seems to bear no direct relationship to the scene itself. These features of the discourse derive instead from the manner in which the scene was processed by the system.

The selection of a new object to introduce, and possibly to continue discussing in further clauses, depends in particular on its relative salience. But it may also depend on a top-down strategy for scanning a scene in a consistent direction. Such as a strategy tends to be adopted in descriptions of rooms, for example. The sequential arrangement of the arguments of a clause, and in the case of asymmetric relations such as *above* the categorization of the relation itself, depends on which of the objects in question is treated as more given than the other.

The grouping of objects in terms of particular lexical relations or as sets with the same category label is based on judgements of the relative similarity and proximity of the objects and possibly also on similarity judgements of a higher order. For example, if there were two small figures above the triangle on the left as well as that on the right, this would increase the likelihood of these particular groupings because of the analogy that could be made. It is also possible to group relations, though this is not reflected in the above description. For example, one could refer to an instance of jumping in terms of crouching, leg straightening, becoming airborne, and landing, or simply in terms of jumping. Here it is not only the similarity of the actions involved that is relevant; it is their close temporal proximity and the fact that one flows smoothly into the next. In this case, the availability of a lexical category (*jump*) for the sequence of actions probably also plays a role in the grouping, at least in the context of language production.

Finally, there is another aspect of the scene description which though not, strictly speaking, structural is a reflection of a basic perceptual processing mode. Recognizing that the two circles in the scene are the same color is not a matter of categorizing some aspects of the scene as *same* and *color*, whatever that would mean. Rather it involves selective attention, the ability to attend to only one of a set of perceptual dimensions that characterize the cir-

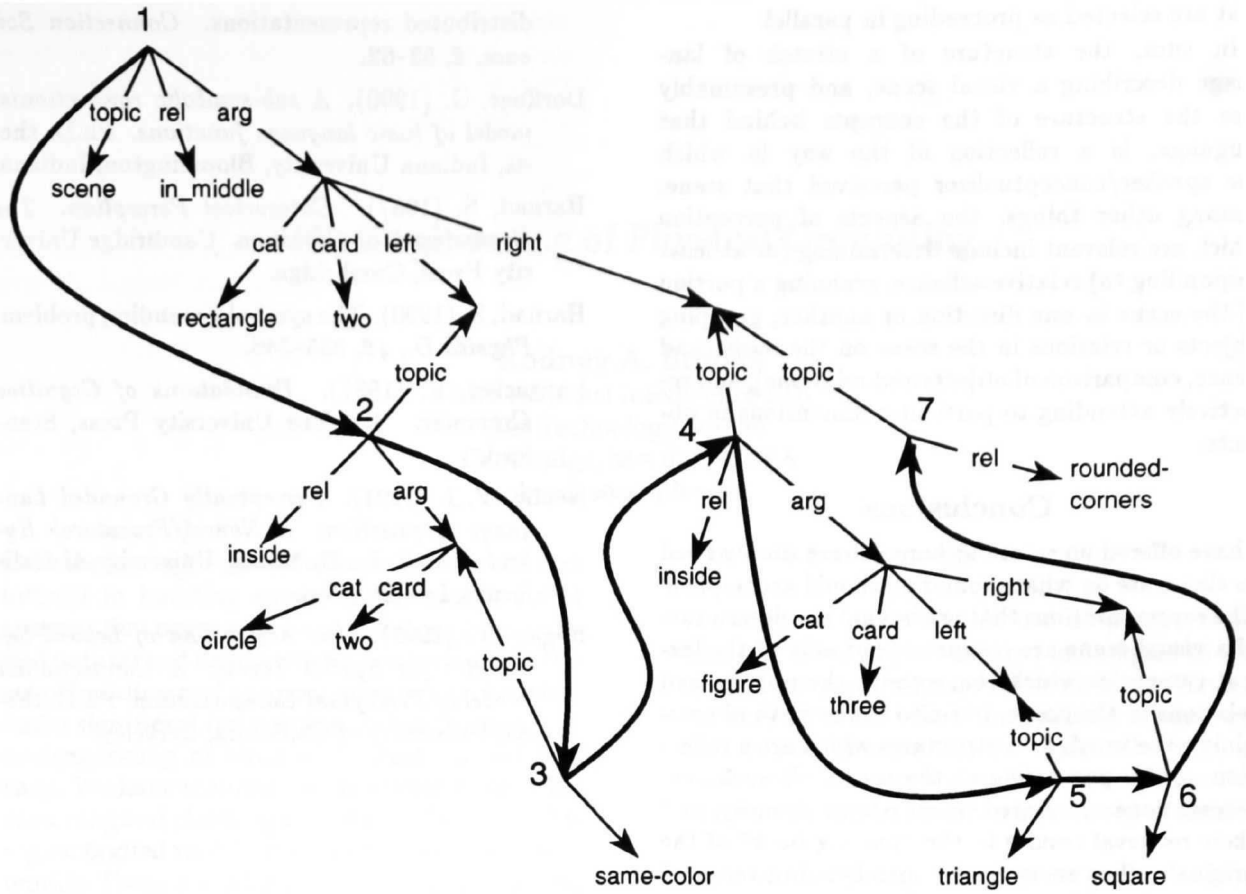

Figure 2: Structure of a Scene Description

cles (a relatively difficult task, incidentally, and one that children do not learn early on).

None of these ideas is new. The notion that linguistic form, and the conceptual structure that is behind it, is a reflection of the way in which states and events are perceived or otherwise experienced is one of the basic tenets of cognitive linguistics. See Langacker (1987) for an extensive discussion, including many other kinds of examples.

Note that the kinds of distinctions that appear in the structure of discourse, like those that relate to categories, are all-or-none, but that they map onto differences in the image, or an "iconic representation" of it, which are continuous. And it appears likely that there would be category effects for the structural distinctions, just as there are for the categories themselves (Harnad, 1987). That is, given two objects which might or might not be treated as a group, we might expect people to judge pairs which are on either side of the boundary between these two types of treatments to be more different than pairs which are on the same size of the boundary, all else being equal.

Note also that the categorization of the objects and relations in a scene, yielding the lexical items in the description, and the various processing actions which are taken, yielding the structure of the description, are intimately related to one another. For asymmetric relations, it is impossible to categorize the relation until it is known which of the objects is to be treated as topic. (Is X *above* Y or Y *below* X?) Thus categorization depends on the processing correlates of topic assignment. Conversely, a decision concerning whether to treat two objects as a set may depend on whether there is an available lexical category which refers to both, that is, a noun that could be used in the plural to refer to the set. Thus it seems best to imagine the processing that is reflected in the structure of discourse and the categorization that is reflected in the lexical items

that are selected as proceeding in parallel.

In sum, the structure of a stretch of language describing a visual scene, and presumably also the structure of the concepts behind that language, is a reflection of the way in which the speaker/conceptualizer perceived that scene. Among other things, the aspects of perception which are relevant include determining (or at least responding to) relative salience, scanning a portion of the scene in one direction or another, grouping objects or relations in the scene on the basis (and hence, comparison of objects and relations), and selectively attending to particular dimensions in objects.

Conclusions

I have offered no solutions here. I have simply tried to elaborate on what grounding should accomplish. The representations that are behind the description of a visual scene are comprised not only of the lexical categories which characterize the objects and relations in the scene, but also of the ways of combining the words into structures which are a reflection of the processing of the scene. If such representations are stored in long-term memory, and their retrieval results in the "playing back" of the originals, then scenes are remembered in terms of the way in which they were originally processed. And the processing actions discussed here are significant for other reasons. Grouping of objects and relations, in particular, is at the heart of what goes on in visual analogy, for example.

What does all of this have to say about the nature of conceptual structure? While concepts could still look like symbol structures, this is somewhat harder to imagine than it would be in a system where only the categories are grounded. With conceptual structure (at least for visual/spatial concepts) now grounded in perception, the motivation for a symbolic component is weakened considerably. It seems more reasonable to imagine a system which is basically connectionist throughout, but with relatively localized representations at the level of lexical categories, where the system stands to gain from the efficiency of a finite, and more or less fixed, set of labels. These categories would be combined into distributed representations of structure, with their capacity to generalize; to model human short-term memory limitations; and, incidentally, to behave systematically too (Chalmers, 1990).

References

Chalmers, D. (1990). Syntactic transformations on distributed representations. *Connection Science, 2*, 53–62.

Dorffner, G. (1990). *A sub-symbolic connectionist model of basic language functions.* Ph.D. thesis, Indiana University, Bloomington, Indiana.

Harnad, S. (1987). *Categorical Perception: The Groundwork of Cognition.* Cambridge University Press, Cambridge.

Harnad, S. (1990). The symbol grounding problem. *Physica D., 42*, 335–346.

Langacker, R. (1987). *Foundations of Cognitive Grammar.* Stanford University Press, Stanford.

Nenov, V. I. (1991). *Perceptually Grounded Language Acquisition: A Neural/Procedural Hybrid Model.* Ph.D. thesis, University of California, Los Angeles.

Regier, T. (1992). *The Acquisition of Lexical Semantics for Spatial Terms: A Connectionist Model of Perceptual Categorization.* Ph.D. thesis, University of California, Berkeley.