Neural Network Models and Cognitive Neuropsychology

by MARTHA J. FARAH, PhD, and JAMES L. McCLELLAND, PhD

Patterns of dissociation among abilities after brain damage are prime sources of evidence that psychologists use to infer the organization of perceptual and cognitive processes. In using dissociations for this purpose, it is virtually always assumed that, after damage to one part of the system, the remaining parts continue to function as they did before damage. This assumption, which has not been tested directly, sometimes leads to counterintuitive results. One such example is presented here: A double dissociation between knowledge of living and nonliving things, which seems to imply that the organization of knowledge in the brain is categorical (eg, living versus nonliving) rather than modality-specific (eg, visual, motoric, etc) as is generally believed.

Dr. Farah is associate professor of psychology, and Dr. McClelland is professor of psychology and computer science, Carnegie Mellon University, Pittsburgh, Pennsylvania.

Address reprint requests to Martha J. Farah, PhD, Carnegie Mellon University, Department of Psychology, Pittsburgh, PA 15213.

Neural network models provide an alternative way to think about the effects of local brain damage on mental functioning.

Neural network models provide an alternative way to think about the effects of local brain damage on mental functioning, and lead to different conclusions about the structure of the mind. We present a model of semantic memory that shows how a fundamentally modality-specific organization for knowledge could, after damage, produce category-specific impairments.

INTRODUCTION

Cognitive Neuropsychology: Inferring Mental Structure from Neuropsychological Dissociations

Brain damage often has rather selective effects on cognitive functioning, impairing some abilities while sparing others. Psychologists interested in describing the "functional architecture" of the mind (ie, the relatively independent information-processing subsystems that underlie human intelligence) have recognized that patterns of cognitive deficit and sparing after brain damage are a potentially useful source of constraints on the functional architecture.

In particular, neuropsychological dissociations provide a direct way of delineating and identifying components of the functional architecture. For example, if it is possible to impair ability A without affecting ability B, for reasons other than A being harder than B, this suggests that A relies upon at least some

distinct components of the functional architecture not relied upon by B. If A can be selectively impaired relative to all other abilities, and reasons such as differential difficulty can again be ruled out, then a natural difference is that there is an "A component" of the functional architecture that is dedicated to some aspect of cognitive processing uniquely associated with ability A.

The Transparency Assumption

Of course, even these apparently simple and straightforward inferences are based on many background assumptions. One such assumption is what Caramazza[1,2] called the "transparency assumption." The transparency assumption is that the cognitive system of a brain-damaged patient is essentially the same as that of a normal subject, except for a "local" modification of the system. In other words, the remaining components will not work differently after damage.

The transparency assumption simplifies the process of inferring the functional architecture of the mind from neuropsychological dissociations. Indeed, the only justification offered for this assumption is that without it the problem of inferring the normal architecture from the behavior of the damaged system would be intractable. Nevertheless, the assumption may be wrong and lead to wrong conclusions. Neurologists have long noted the highly interactive nature of brain organization and the consequent tendency for local damage to unleash new emergent organizations or modes of functioning in the remaining system.[3,4]

With the advent of neural network models, psychologists have a new conceptual tool with which to reason about the effects of damage in highly interactive systems, and need not rely on the transparency assumption to make their inferences.

Principles of Nontransparency

Neural network models can be used as a source of principled constraints on the ways in which the remaining parts of the system behave after local damage.

Brain damage often has rather selective effects on cognitive functioning.

For present purposes, the relevant principles of operations in neural network systems are:

- *Distributed representation of knowledge.* In network systems, representations consist of a pattern of activation distributed over a population of units. Long-term memory knowledge is encoded in the pattern of connection strengths distributed among a population of units.
- *Graded nature of information processing.* In network systems, processing is not all or none; representations can be partially active, for example, by the partial activation of some of those units that would normally be active. Partial knowledge can be embodied in connection strengths, either before learning has been completed or after partial damage.
- *Interactivity.* The units in network models are highly interconnected; thus, mutual influence among different parts of the system is the rule rather than the exception. These influences can be excitatory, as when one part of a distributed representation activates the remaining parts (pattern completion), or they can be inhibitory, as when different representations compete with one another to become active or maintain their activation. Note that interactivity is the aspect of the neural network framework that is most directly incompatible with the transparency assumption. If the normal operation of a given part of the system depends on the influence of some other part, it may not operate normally after that other part has been damaged.

FUNCTIONAL ARCHITECTURE OF SEMANTIC MEMORY: CATEGORY-SPECIFIC OR MODALITY-SPECIFIC?

One incentive for exploring alternatives to the transparency assumption is that this assumption sometimes leads to bizarre or counterintuitive conclusions. One such case concerns a double dissociation between two types of impairment in semantic memory knowledge. Beginning in the 1980s, Warrington and her colleagues began to report the existence of patients with selective impairments in knowledge of either living or nonliving things.[5-7] In subsequent years, many other researchers corroborated these surprising observations in patients recovering from a variety of brain lesions, including postviral encephalitis, brain trauma and cerebrovascular accidents.[8-11]

Although these patients are not entirely normal in their knowledge of the relatively spared category, they are markedly worse at recognizing, defining, or answering questions about items from the impaired category. The existence of a double dissociation makes it unlikely that a sheer difference in difficulty underlies the apparent selectivity of the deficits. Some of the studies cited above tested several alternative explanations of the impairments in terms of factors other than semantic category (such as name frequency or familiarity) and failed to support them.

Interpretation of "Living" and "Nonliving Things" Deficits Relative to the Functional Architecture of Semantic Memory

Using the transparency assumption, the most straightforward interpretation of the double dissociation between knowledge of living and nonliving things is that these two bodies of knowledge are represented by two separate category-specific components of the functional architecture of semantic memory. Figure 1 represents a

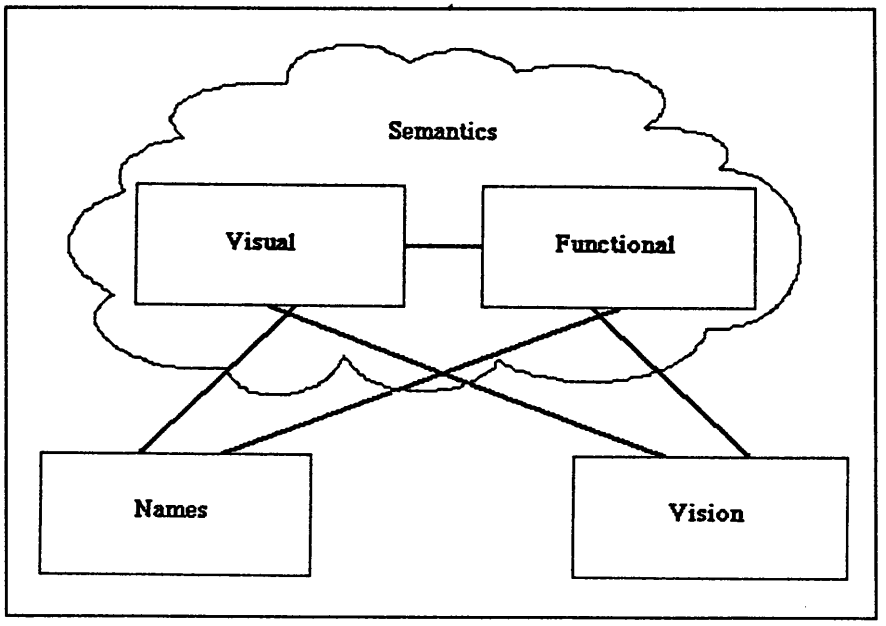

Figure 1. Outline of a category-specific model of semantic memory.

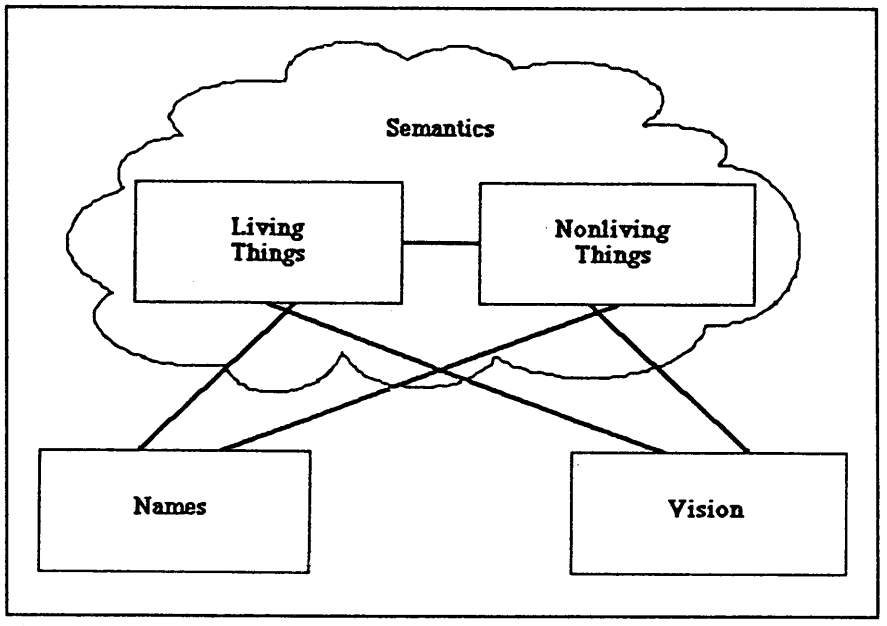

Figure 2. Outline of a model of modality-specific semantic memory.

model of semantic memory and its relation to visual perception and language.

Warrington and colleagues have suggested an alternative interpretation, however, according to which semantic memory is fundamentally modality-specific. They argue that selective deficits in knowledge of living and nonliving things may reflect the differential weighting of information from different sensorimotor channels in representing knowledge about these two categories. They have pointed out that living things are distinguished primarily by their sensory attributes, whereas nonliving things are distinguished primarily by their functional attributes.

For example, our knowledge of an animal such as a leopard, by which we distinguish it from other similar creatures, is predominantly visual. In contrast, our knowledge of a desk, by which we distinguish it from other furniture, is predominantly functional (ie, what it is used for). Thus, the distinctions between impaired and preserved knowledge in the cases reviewed earlier may not be "living/nonliving" distinctions per se, but "sensory/functional" distinctions, as illustrated in Figure 2.

The sensory/functional hypothesis seems preferable to a strict living/nonliving hypothesis for two reasons. First, it is more consistent with what is already known about brain organization. It is well known that different brain areas are dedicated to representing information from specific sensory and motor channels. Functional knowledge could conceivably be tied to the motor system.

A second reason for preferring the sensory/functional hypothesis to the living/nonliving hypothesis is that exceptions to the living/nonliving distinction have been observed in certain cases. For example, Warrington and Shallice[5] report that their patients, who were deficient in their knowledge of living things, also had impaired knowledge of gem stones and fabrics. Warrington and McCarthy's patient, whose knowledge of most nonliving things was impaired, seemed to have retained good knowledge of very large outdoor objects such as bridges or windmills.[6] It is at least possible that our knowledge of these aberrant categories of nonliving things is primarily visual.

Unfortunately, there appears to be a problem with the hypothesis that "living things impairments" are just impairments in sensory knowledge, and "nonliving things impairments" are just impairments in functional knowledge. This hypothesis seems to predict that cases of "living things impairment" should show good knowledge of

the functional attributes of living things, and cases of "nonliving things impairment" should show good knowledge of the visual attributes of nonliving things.

The evidence available in cases of "nonliving things impairment" is limited to performance in matching-to-sample tasks, which does not allow us to distinguish knowledge of visual or sensory attributes from knowledge of functional attributes. However, there does appear to be adequate evidence available in cases of "living things impairment," and in several cases it disconfirms these predictions.[12] For example, in the study by Newcombe et al,[9] a subject was impaired relative to normal subjects in his ability to sort living things according to such nonsensory attributes as whether or not they were generally found in the United Kingdom, in contrast to his normal performance when the task involved nonliving things.

In sum, the sensory/functional hypothesis seems preferable to the living/nonliving hypothesis because it is more in keeping with what we already know about brain organization. However, it does not seem able to account for the impaired ability of these patients to retrieve nonvisual information about living things.

Accounting for Category-Specific Impairments with a Modality-Specific Architecture

We have modeled the double dissociation between knowledge of living and nonliving things using a simple autoassociative memory architecture with modality-specific components.[12] We found that lesions in a semantic memory system consisting of visual and functional components produced selective impairments in knowledge of living things and nonliving things. More importantly, we found that such a model could account for the impairment of both visual and functional knowledge of living things.

The basic architecture of the model is shown in Figure 2. There

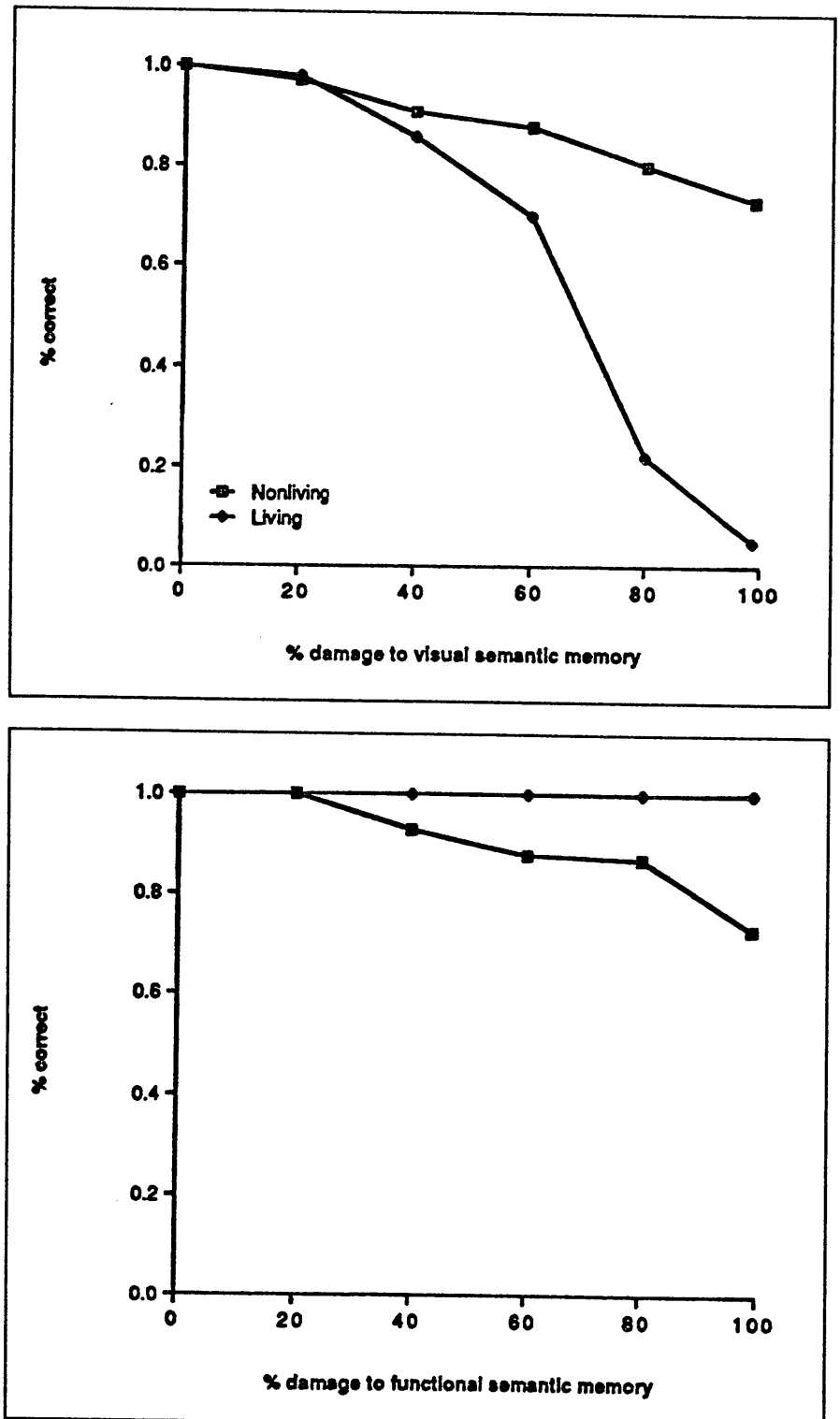

Figure 3. Top graph shows the averaged picture-to-name and name-to-picture matching performance of the model for living and nonliving items under varying degrees of damage to visual semantics. Bottom graph shows that the opposite dissociation is obtained when functional semantics is damaged.

are three pools of units, representing the names of items, the perceived appearances of items, and the semantic memory representations of items. The semantic memory pool is subdivided into visual

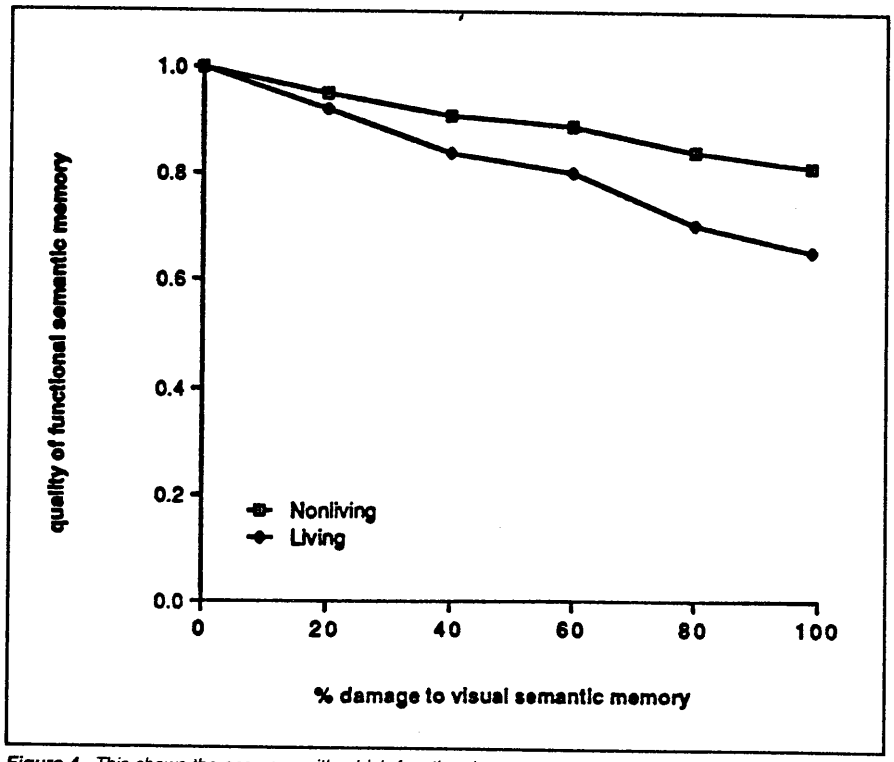

Figure 4. This shows the accuracy with which functional semantic memory information could be activated for living and nonliving things after different degrees of damage to visual semantics.

semantic memory and functional semantic memory. An item, living or nonliving, is represented by a pattern of +1 and −1 activations over the name and visual units, and a pattern of +1 and −1 activations over a subset of the semantics units. The relative proportion of visual and functional information comprising the semantic memory representation of living and nonliving things was derived empirically.

Normal subjects identified terms in dictionary definitions of the living and nonliving items used by Warrington and Shallice[5] as referring to visual properties or functional properties. This experiment confirmed that visual and functional information was differentially weighted in the definitions of living and nonliving things, and the results were used to determine the average proportions of visual and functional units in the semantic memory representations of the living and nonliving items. For the living items, about seven times as many visual semantic units as functional semantic units participated in the semantic memory pattern; for nonliving items, the proportions were closer to equal. Units of semantic memory not involved in a particular item's representation took the activation value of zero.

The model was trained using the delta rule[13] to associate the correct semantic and name portions of its pattern when presented with the visual portion of its pattern as input, and the correct semantic and visual portions of its pattern when presented with the name portion as input. It was then damaged by eliminating different proportions of functional or visual semantic units, and its performance was assessed in a simulated picture-name matching task. In this task, each item's visual input representation is presented to the network. The pattern that is activated in the name units is assessed, or each pattern's name is presented and the resultant visual pattern is assessed. The resultant pattern is scored as correct if it is more similar to the correct pattern than to any of the other 19 patterns.

Figure 3 (top) shows the averaged picture-to-name and name-to-picture matching performance of the model for living and nonliving items under varying degrees of damage to visual semantics. With increased damage, the model's performance drops. Performance drops more precipitously for living things, in effect showing an impairment for living things of comparable selectivity to the patients in the literature. Figure 3 (bottom) shows that the opposite dissociation is obtained when damage to functional semantics occurs.

The critical challenge for a modality-specific model of semantic memory is to explain how damage could create an impairment in knowledge of living things, which includes functional knowledge of living things. To evaluate the model's ability to access functional semantic knowledge, we presented either name or visual input patterns as before, but instead of assessing the match between the resulting output pattern and the correct output pattern, we assessed the match between the resulting pattern in functional semantics and the correct pattern in functional semantics. The normalized dot product of these two patterns, which provides a measure between zero (completely dissimilar) and one (identical), served as the dependent measure.

Figure 4 shows the accuracy with which functional semantic memory information could be activated for living and nonliving things after different degrees of damage to visual semantics. At all levels of damage, the ability to retrieve functional semantic knowledge of living things is disproportionately impaired.

These dissociations can be understood as follows. In the case of picture-name matching, the ability of a given output unit (eg, a name unit, in the case of picture-to-name matching) to attain its correct activation value depends on the input it receives from the units to which it

is connected. These consist of other name units (collateral connections) and both visual and functional semantics units. Therefore, the more semantics units that have been eliminated, the more the output units are deprived of the incoming activation they need to attain their correct activation values. Because most of the semantic input to the name units of living things is from visual semantics, whereas the same is not true for nonliving things, damage to visual semantics will eliminate a greater portion of the activation needed to activate the name patterns for living things than for nonliving things, and will therefore have a more severe impact on performance.

The same principle applies to the task of activating functional semantics, although in this case the units are being deprived of collateral activation. Thus, when visual semantic units are destroyed, one of the sources of input to the functional semantics units is eliminated. For living things, visual semantics comprises a proportionately larger source of input to functional semantics units than for nonliving things, hence the larger effect for these items.

Relation to the Transparency Assumption

Contrary to the transparency assumption, when the visual semantics component is damaged, the remaining parts of the system do not continue to function as before. In particular, functional semantics, which is part of the nondamaged residual system, becomes impaired in its ability to achieve the correct patterns of activation when given input from vision or language. This is caused by the loss of collateral support from visual semantics. The ability of this model to account for the impairment in accessing functional knowledge of living things depends critically upon this nontransparent aspect of its response to damage.

CONCLUSIONS

The question of whether the functional architecture of the mind is "transparent" after brain damage, that is, whether the remaining, nondamaged components continue to function normally, is a fundamental one, and cannot be answered decisively on the basis of one study. Nevertheless, we can draw at least two conclusions from the study presented here. First, that in the realm of semantic memory impairments, one arrives at more sensible inferences about the functional architecture of the mind by abandoning the transparency assumption and instead conceiving of the brain as a highly interactive system. Second, that even a high degree of interactivity does not make the problem of inferring functional architecture from the behavior of the damaged system an intractable one. Neural network models give us principled ways of reasoning about the response of highly interactive systems to damage.

REFERENCES

1. Caramazza A. The logic of neuropsychological research and the problem of patient classification in aphasia. *Brain Lang.* 1984; 21:9-20.
2. Caramazza A. On drawing inferences about the structure of normal cognitive systems from the analysis of patterns of impaired performance: the case for single-patient studies. *Brain Cogn.* 1986; 5:41-66.
3. Ferrier D. *The Functions of the Brain.* London, England: Smith, Elder & Co; 1896.
4. Jackson JH. On the anatomical and physiological localization of movements in the brain. *Lancet.* 1873; 1:84-85, 162-164, 232-234.
5. Warrington EK, Shallice T. Category-specific semantic impairment. *Brain.* 1984; 107:829-854.
6. Warrington EK, McCarthy R. Category-specific access dysphasia. *Brain.* 1983; 106:859-878.
7. Warrington EK, McCarthy R. Categories of knowledge: further fractionation and an attempted integration. *Brain.* 1987; 110:1273-1296.
8. Farah MJ, McMullen PA, Meyer MM. Can recognition of living things be selectively impaired? *Neuropsychologia.* 1991; 29:185-193.
9. Newcombe F, Mehta Z, de Haan EF. Category specificity in visual recognition. In: Farah MJ, Ratcliff G, eds. *The Neuropsychology of High-Level Vision: Collected Tutorial Essays.* Hillsdale, NJ: Erlbaum. In press.
10. Sartori P, Job R. The oyster with four legs: A neuropsychological study on the interaction of visual and semantic information. *Cognitive Neuropsychology.* 1988; 5:105-132.
11. Silveri C, Gainotti G. Interaction between vision and language in category-specific semantic impairment. In: Coltheart M, Sartori G, Job R, eds. *The Cognitive Neuropsychology of Language.* London, England: Erlbaum; 1988.
12. Farah MJ, McClelland JL. A computational model of semantic memory impairment: modality-specificity and emergent category-specificity. *J Exp Psychol (Gen).* 1991; 122:339-357.
13. Rumelhart DE, McClelland JL. *Parallel Distributed Processing, Vol 1: Foundations.* Cambridge, Mass: MIT Press; 1986.