Psychological Review
2004, Vol. 111, No. 1, 205–235

Copyright 2004 by the American Psychological Association, Inc.
0033-295X/04/$12.00 DOI: 10.1037/0033-295X.111.1.205

Structure and Deterioration of Semantic Memory: A Neuropsychological and Computational Investigation

Timothy T. Rogers
Medical Research Council Cognition and Brain Sciences Unit,
Carnegie Mellon University, and Center
for the Neural Basis of Cognition

Matthew A. Lambon Ralph
University of Manchester

Peter Garrard
National Hospital for Neurology and Neurosurgery

Sasha Bozeat
Medical Research Council Cognition and Brain Sciences Unit

James L. McClelland
Carnegie Mellon University and Center for the
Neural Basis of Cognition

John R. Hodges and Karalyn Patterson
Medical Research Council Cognition and Brain Sciences Unit

Wernicke (1900, as cited in G. H. Eggert, 1977) suggested that semantic knowledge arises from the interaction of perceptual representations of objects and words. The authors present a parallel distributed processing implementation of this theory, in which semantic representations emerge from mechanisms that acquire the mappings between visual representations of objects and their verbal descriptions. To test the theory, they trained the model to associate names, verbal descriptions, and visual representations of objects. When its inputs and outputs are constructed to capture aspects of structure apparent in attribute-norming experiments, the model provides an intuitive account of semantic task performance. The authors then used the model to understand the structure of impaired performance in patients with selective and progressive impairments of conceptual knowledge. Data from 4 well-known semantic tasks revealed consistent patterns that find a ready explanation in the model. The relationship between the model and related theories of semantic representation is discussed.

Human beings live in a world infused with meaning. This capacity is so fundamental that it seems to reside near the core of what we intend when we speak of human cognition. Small wonder, then, that scientists and philosophers have concerned themselves with questions about the nature of semantic knowledge for centuries. How is meaning stored, represented, and retrieved in the mind and brain? How is it acquired in development, and how might it be disturbed by pathology?

Neuropsychology has long been used as one tool for answering these kinds of questions. More than 100 years ago, the German neurologist Carl Wernicke sketched out a theory of semantic memory based on his study of neuroanatomy and disorders of language. Wernicke (1900, as cited in G. H. Eggert, 1977) proposed that semantic knowledge arises from the interactions among modality-specific perceptual representations of objects and of the words we use to describe these objects. He suggested that these interactions are mediated by transcortical or association areas of cortex, which have anatomical links to perceptual and language areas.

> The concept of a rose is composed of a “tactile memory image”—“an image of touch”—in the central projection field of the somesthetic cortex. It is also composed of a visual memory image located in the visual projection field of the cortex. The continuous repetition of similar sensory impressions results in such a firm association between those different memory images that the mere stimulation of one sensory avenue by means of the object is adequate to call up the concept of the object. In some cases, many memory images of different sensory areas and in others only a few correspond to a single concept. However, by the very nature of the object, a firmly associated constellation of such memory images which form the anatomic substrate of each concept is established. This sum total of closely associated memory images must “be aroused into consciousness” for

Timothy T. Rogers, Medical Research Council Cognition and Brain Sciences Unit, Cambridge, United Kingdom; Department of Psychology, Carnegie Mellon University; and Center for the Neural Basis of Cognition, Pittsburgh, PA. Matthew A. Lambon Ralph, Department of Psychology, University of Manchester, Manchester, United Kingdom. Peter Garrard, National Hospital for Neurology and Neurosurgery, London. Sasha Bozeat, John R. Hodges, and Karalyn Patterson, Medical Research Council Cognition and Brain Sciences Unit. James L. McClelland, Department of Psychology, Carnegie Mellon University, and Center for the Neural Basis of Cognition.

The research was supported by National Institute of Mental Health Grant MH47566 to James L. McClelland. Thanks are due to Jim Tanaka for allowing us to investigate his feature norms in an earlier version of this work.

Correspondence concerning this article should be addressed to Timothy T. Rogers, Medical Research Council Cognition and Brain Sciences Unit, 15 Chaucer Road, Cambridge CB2 2EF, United Kingdom. E-mail: tim.rogers@mrc-cbu.cam.ac.uk

> perception not merely of sounds of the corresponding words but also for comprehension of their meaning. Following our anatomic mode of interpretation, we also postulate for this process the existence of anatomic tracts, fibers, connections, or association tracts between the sensory speech center of word-sound-comprehension and those projection fields which participate in the formation of the concept. (Wernicke, 1900, as cited in Eggert, 1977, p. 237)

This framework allowed Wernicke (1900, as cited in Eggert, 1977) to account for the range of language disorders observed in his clinical experience and to predict the occurrence of generalized semantic disorders arising from damage to the hypothesized transcortical areas, which were thought to mediate interactions between peripheral sensory zones. Such generalized impairments have now been documented, as the result of a variety of etiologies including dementia of the Alzheimer's type (DAT), herpes simplex virus encephalitis (HSVE), cerebral vascular accident (CVA), and head injury. In this article, we focus on a neurodegenerative disorder that, in many respects, provides the clearest patient model of semantic disruption—semantic dementia (Hodges, Patterson, Oxbury, & Funnell, 1992; Snowden, Goulding, & Neary, 1989). Patients with semantic dementia exhibit a profound and progressive impairment of semantic knowledge in a variety of tasks, including picture naming, word-to-picture matching, sorting, drawing and copying, and category matching (Hodges, Graham, & Patterson, 1995; Schwartz, Marin, & Saffran, 1979; Warrington, 1975). Despite these difficulties, their remaining cognitive faculties seem remarkably spared. They show little difficulty on tests of spatial memory such as the Rey Complex Figure, are well oriented in space and time, and have good recognition memory, normal visual perception, and unimpaired digit spans (Patterson & Hodges, 2000; Snowden, Neary, & Mann, 1996). Their speech, though marked with severe word-finding difficulties, is otherwise grammatical and fluent.

The cognitive impairments witnessed in semantic dementia arise from progressive focal atrophy of the anterior and inferolateral aspects of the temporal cortex bilaterally (Lambon Ralph, McClelland, Patterson, Galton, & Hodges, 2001; Mummery et al., 2000). The affected region is a plausible anatomical locus for Wernicke's (1900, as cited in Eggert, 1977) proposed "transcortical association area," in that these areas are known to receive convergent input from and send output to all sensory and motor systems (Gainotti, Silveri, Daniele, & Giustolisi, 1995; Gloor, 1997; Grey & Bannister, 1995). For example, the temporal pole, the region almost invariably affected in the earliest stages of semantic dementia, has extensive connections with all three temporal gyri, which in turn receive projections from earlier sensory processing centers. Specifically, the anterior part of the inferior temporal gyrus is thought to be the terminus of the ventral visual processing stream; the middle temporal gyrus is generally thought to integrate input from somatosensory, visual, and auditory processing streams; and the superior temporal gyrus as well as the superior temporal sulcus play important roles in auditory and speech perception. The cortex of the temporal pole and the anterior portion of the inferior temporal gyrus send projections to orbitofrontal and prefrontal cortex as well (Grey & Bannister, 1995).

The syndrome of semantic dementia provides the clearest evidence of a relatively pure semantic impairment that affects all modalities of testing and all conceptual domains, which suggests that semantic memory may be largely subserved by a unitary and relatively homogeneous neural system in the anterior and lateral aspects of the temporal cortices bilaterally (Bozeat, Lambon Ralph, Patterson, Garrard, & Hodges, 2000; Lambon Ralph, Graham, Patterson, & Hodges, 1999). Although the observed deficits are typically neither category nor modality specific, the dissolution of semantic knowledge is nevertheless structured. For example, patients with semantic impairment consistently show more robust memory for the general properties of objects than for their more specific features (Done & Gale, 1997; Hodges et al., 1995; Warrington, 1975) and frequently overextend familiar or typical labels to semantically related objects in tests of confrontation naming (Hodges et al., 1995). It seems reasonable to suppose that such patterns reflect representational structure in the semantic system. But where does such structure come from?

In this article, in agreement with Wernicke (1900, as cited in Eggert, 1977) and many others (e.g., A. R. Damasio, 1989; Kellenbach, Brett, & Patterson, 2001; Martin & Chao, 2001; McClelland & Rogers, 2003; Warrington & Shallice, 1984), we suggest that the representations and processes underlying semantic memory are best understood within a theory in which semantic knowledge emerges from the interactive activation of modality-specific perceptual representations of objects and statements about objects. In contrast to some contemporary approaches, we argue that semantic representations do not need to extract, store, and retrieve attributes, facts, or propositions about objects to fulfill this role; they need only to allow such information to be produced as overt responses in particular task contexts.

We further argue that abstract semantic representations emerge as a product of statistical learning mechanisms in a region of cortex suited to performing cross-modal mappings by virtue of its many interconnections with different perceptual-motor areas. In this view, modality-specific perceptual representations provide the input to semantics, and modality-specific response systems permit the expression of semantic knowledge. The content of semantic memory is represented in the same regions of cortex that directly encode modality-specific regularities in the environment during perception and action. Domain-general learning mechanisms operate to allow the semantic system, when presented with information about an object in some perceptual modality, to make correct inferences about the object's unspecified attributes. As a consequence, the system acquires abstract representations whose similarity relations are not tied to any individual modality but capture the deep structure across modalities.

To support these arguments, we consider a simple computational implementation of the theory, in which visual representations of objects and perceptual representations of verbal statements about these objects interact with one another by means of an intermediating semantic system. In this model, mediating semantic representations are not prespecified but emerge as the network learns to map between verbal descriptions, names, and visual appearances of objects. These acquired representations do not code explicit semantic content, but they are structured in ways that facilitate the system's ability to generate appropriate responses when given perceptual inputs. When the inputs and outputs capture aspects of structure apparent from visual and verbal attribute norms, the model acquires internal representations whose similarity structure is not reflected in either modality independently.

To evaluate the model, we investigate its ability to perform analogs of semantic tasks as it is subjected to increasingly severe simulated lesions. The particular representations discovered by the model when trained with empirically motivated input patterns, along with the processing assumptions embodied in the model, provide an intuitive a posteriori account of many previously reported findings in the study of semantic dementia. The model also yields novel and counterintuitive predictions about patient performance that are supported by the results of studies reported here for the first time.

Parallel Distributed Processing (PDP) Implementation of the Theory

The model implementation of our theory is shown in Figure 1. It consists of sets of nonlinear processing units organized into groups and connected, as shown in the illustration. Associated with each unit is an activation state, which varies along a sigmoid function bounded at 0 and 1. The state of a given unit at any point in time is determined by the strength of its input. Each unit group (or layer) represents an anatomically distinct region of cortex, specialized to subserve a particular function by virtue of its connectivity. For example, the layer labeled *visual* is specialized to represent high-level visual information, as a result of receiving input from earlier visual processing streams. In the model, these perceptual signals are presented as external inputs to the units in the visual layer—that is, the states of the units in the visual layer can be set directly by visual stimuli in the environment. Similarly, the *verbal* layer represents areas of cortex that subserve linguistic processes; these unit states may be set directly by linguistic stimuli in the environment, such as an object's name or its verbal description. Because the visual and verbal units receive external inputs, these are also referred to as *visible units*.

All of the units in the visual and verbal layers are bidirectionally connected with the set of units in the layer labeled *semantic*. The semantic units do not receive direct, external inputs from the environment. Their states may be set only by the activity of the units to which they are connected, as weighted by the strength of the intervening connections. Consequently, these are referred to as *hidden units* (Rumelhart, McClelland, & PDP Research Group, 1986).

The units in the visible layers each represent a particular, explicit property in the corresponding modality. For instance, each unit in the verbal layer represents a verbal statement that describes an object, such as a name (*animal*, *bird*, *goose*), a visual property (*has eyes*, *has wheels*), a functional property (*can fly*, *can roll*), or an encyclopedic property (*lives in Africa*, *found in kitchen*). Thus, verbal descriptions of objects can be represented as patterns of activity across these units, and objects to which similar predicates apply receive similar representations in this layer. In the illustration, the verbal units are divided into four pools, corresponding to four different kinds of information that may be expressed verbally (i.e., perceptual, functional, encyclopedic, and name information). We have arranged the units this way as a reminder that verbal propositions can refer to any of several different kinds of information. However, in our model, all verbal statements are construed as first activating the same regions of cortex, regardless of the kind of information to which they refer; and the arrangement of verbal units into separate pools in Figure 1 has no functional consequence in the model.

Each unit in the visual layer represents a unique visual property, such as *has limbs* or *is round*. Visual representations of objects correspond to patterns of activity across this assembly of visual features, such that objects with similar visual appearances give rise to similar visual representations. Note that the units standing for visual properties are quite separate from the units that stand for verbal propositions about visual properties in the model.

A stimulus is presented to the network by directly setting (or clamping) the states of the visible units that correspond to the features apparent in the stimulus item. For example, to present a picture of a canary to the network, we turn on the units in the visual layer that represent the visual attributes of canaries and turn the remaining visual units off. While these units are clamped, their states are not affected by the activation of the other units in the model. To perform correctly, the network must then activate the

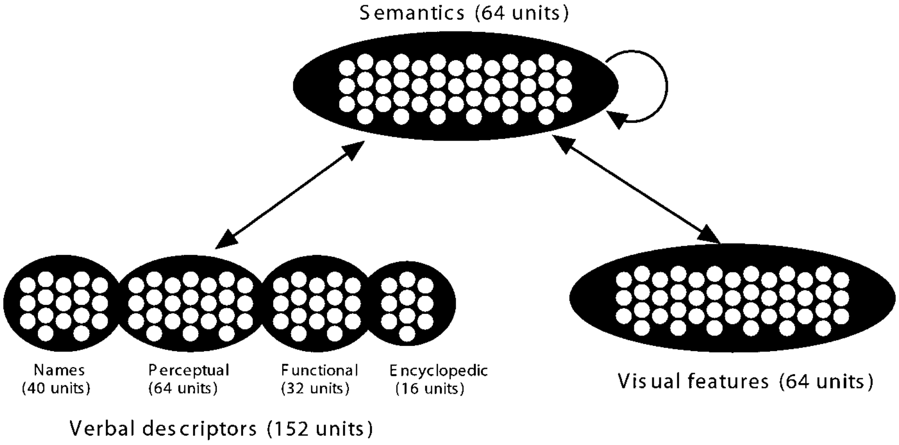

Figure 1. Architecture of the model. Verbal descriptor and visual feature units receive input directly from the environment.

verbal proposition units that describe the canary (including its name). Alternately, we might present the name *canary* to the network by turning on the corresponding unit in the verbal layer and turning off all of the other name units. In this case, the network must activate the other verbal propositions that describe the canary and the visual attributes that correspond to the canary's appearance. We might give the network a verbal description as input by clamping on the appropriate proposition units and requiring it to produce the corresponding name or visual pattern as output. Thus, in its trained state, the model is capable of carrying out analogs of semantic tasks such as picture naming, naming to definition, verbal description, drawing and copying, category and property verification, and so on.

Information is processed in the model through the successive updating of unit activation states over time. When a verbal or visual stimulus item is presented to the network, the states of the units throughout the network change gradually in response to this input. On a given time step, all the units calculate their new states once by summing the activation of the units from which they receive projections, weighted by the magnitude of the intervening connection, and passing the result through a sigmoidal squashing function. The unit activations are then updated simultaneously, and the algorithm begins again, with each unit calculating a new activation state in response to the changes from the last pass. This process is reiterated until the unit states stop changing, at which point the network is said to have settled into a steady state (Rumelhart et al., 1986).

The steady state (or attractor) into which the network settles when given a particular input depends on the values of the interconnecting weights. When these weights are configured to allow the network to perform correctly, the model can be said to "know" the domain; hence, the model's semantic knowledge is stored in its weights (McClelland, Rumelhart, & PDP Research Group, 1986). To find an appropriate set of weights, the model is exposed to visual and verbal patterns that are associated with the different objects in its virtual environment and is trained with a variant of the backpropagation learning algorithm suited to learning in a recurrent network (Rumelhart, Hinton, & Williams, 1986). In each training instance, an input is presented to the model for a fixed period of time, and the activity is allowed to spread through the network. The inputs are then removed, and the network is permitted to cycle for several more time steps. Finally, the actual states of the visible units (the visual and verbal units) are compared with their desired states, and all the weights throughout the network are adjusted by a small amount to reduce the discrepancy between the observed and target states.

The application of inputs and targets to the network can be considered analogous to natural semantic tasks imposed by the environment. For example, when a child learns to name an object, we might assume that the child is first directed to look at it and then is told its name. In the model, we simulate this process by clamping a visual input pattern that corresponds to the object's appearance, allowing the network to cycle, and then removing the inputs and applying the target pattern to the name units. Similarly, we might imagine a mother and child looking at a picture book. The mother asks, "Can you show me the piggie?" The child must activate some of his or her knowledge about the visual features of a pig and use this to direct the choice of animals in the book. Feedback from the mother allows the child to know whether the choice was correct. In the model, we simulate this kind of activity by clamping the name input unit, allowing the network to settle, and comparing the states of the visual units to the desired target states provided by the environment. Thus, on any given trial, any of the visible units (visual features, verbal propositions, or names) may serve either as input or output units. We assume the environment provides both the input and the target states (see Rogers & McClelland, in press, for further discussion).

As it learns the associations among names, appearances, and descriptions, the model assigns to each input a stable pattern of activity across its hidden units. These patterns are not directly constrained to represent the presence or absence of particular features in the environment. The model is free to use whatever representations emerge from the learning algorithm. However, the representations acquired by the model are influenced by the similarity structure of representations in visual and verbal modalities (Plaut, 2002; Rogers & McClelland, in press). In the next subsection, we consider the extent to which different objects tend to share the same verbal descriptors and visual attributes, according to recent attribute norming and drawing studies. We then derive a simple model training environment that captures the important aspects of structure apparent from these data, and we examine the representations that arise across hidden units in the model when trained with these patterns. The structure of these representations, together with the processing assumptions embodied in the model, form the basis of our account of data from semantic dementia in the Understanding the Structured Deterioration of Conceptual Knowledge in Semantic Dementia section, below.

Assessing Verbal and Visual Structure in the Environment

Many different efforts to assess the attribute structure of concepts have appeared in the literature (e.g., McRae & Cree, 2002; Devlin, Gonnerman, Andersen, & Seidenberg, 1998; Garrard, Lambon Ralph, Hodges, & Patterson, 2001; McRae, De Sa, & Seidenberg, 1997; Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976; Tyler, Moss, Durrant-Peatfield, & Levy, 2000), and they typically use verbal attribute-listing methods. Participants are given a series of object names, and for each object they are asked to list all the properties they can think of that are characteristic of the object. The results are often interpreted as providing a useful proxy measure of the attribute structure of semantic representations themselves. Some researchers, however, have rightly pointed out a number of difficulties encountered by this interpretation (e.g., Murphy & Medin, 1985; Sloman & Rips, 1998):

1. The number and type of attributes generated in the task can vary substantially as a function of the amount of time devoted to each exemplar and whether the listing task is largely instruction-free or is accompanied by specific prompts from the experimenter.

2. There are many degrees of freedom in the way that responses are coded and analyzed. For example, the same attribute may be described by different words (e.g., a horse may *whinny* or *neigh*); most researchers choose, but to varying degrees, to collapse such responses into a single feature.

3. Some features common to all or almost all members of a category are simply not the kinds of attributes that spring readily to mind in this task. For example, when asked to list the properties of a duck, participants may be unlikely to say that it "has DNA," "has blood," or "has eyes"; but these attributes are presumably an important part of the concept *duck* and link it strongly to the domain of animals.

4. The names and propositions that are used to describe objects may fail to capture aspects of structure in the environment that are more directly apparent through other modes of perception. For example, although people may seldom use the proposition *has eyes* when describing a duck, they may directly observe that a duck has eyes (and that in this respect it is similar to other animals) whenever they encounter one in the environment. Similarly, objects with similar visual appearances are often described with different verbal labels (e.g., *sharks* and *dolphins*), whereas objects or object parts described by the same verbal label may differ substantially in appearance (e.g., the *handle* on a hammer and the *handle* on a teacup).

We argue, however, that these difficulties are substantially less vexing if one conceives of the feature-listing task differently. Specifically, we construe the task as a verbal act that is driven by abstract semantic representations that do not themselves encode explicit content. Hence, the data from such tasks do not provide a window on underlying feature-based semantic representations; they simply indicate the words that people are likely to use when referring to objects in speech. In this view, there is no way to probe directly the compositional structure of conceptual knowledge, although the properties given in attribute-listing tasks do provide a useful measure of one source of such information about similarity available from the environment, namely language. Other sources of similarity information are available through other perceptual channels: in the sounds produced by objects, the actions they afford, the behaviors they exhibit, and in their visual appearances.

A comprehensive assessment of the claim that representational structure in semantics can be derived from the perceptual structure of the environment would require measurement of the perceptual similarities that exist among a wide range of objects for all relevant modalities. Here, we provide a more modest consideration of measures of two different sources of information about similarity: what people say about objects in verbal attribute-listing studies and what visual features of objects people depict when drawing them. These analyses allow us to discover whether the similarities captured in verbal descriptions and visual reproductions are congruent with one another and provide an empirical basis for generating patterns for use in the simulation work.

Verbal Attribute Structure

The verbal attribute-listing data that form the primary basis for our analysis were described in detail by Garrard et al. (2001). In this study, norms were collected for 62 object concepts drawn from six semantic categories. These items also form the basis for the battery of neuropsychological tasks developed by our group to assess semantic memory. Half of the items are living things (land animals, birds, and fruits) and half are nonliving (household objects, vehicles, and tools). Garrard et al. (2001) asked the participants to list as many properties as they could that were true of each item and provided them with prompting questions to help them think of attributes that might not otherwise spring to mind. Lists for all items were collected from 20 participants. The responses from each participant were concatenated into a single list. Properties that were listed by fewer than 2 participants for a given item were discarded; property names that referred to the same underlying semantic attribute (e.g., *is big* and *is large*) were collapsed into a single feature. This yielded a total of 618 different properties across all the 62 items. From these data, Garrard et al. (2001) derived many interesting observations, but we focus on three questions of special interest for our theory.

First, to what extent do the items in the battery seem similar to one another, considering only their propensity to share verbal descriptors? Under the PDP theory, the similarity structure of the hidden semantic representations depends on the similarities apparent in the input and output across different modalities (Plaut, 2002), including similarities apparent in spoken references to objects. It is important, therefore, to determine what similarities may be discerned among the verbal descriptions of objects captured in the norms.

To accomplish this, we performed a hierarchical cluster analysis of items from Garrard et al.'s (2001) data. The similarity between every pair of items in the set was calculated by taking the total number of attributes held in common by both items as a proportion of the total number of unique attributes listed across the pair of items, a measure known as *Jaccard's distance*. The similarity matrix was entered into a hierarchical clustering algorithm (Everitt, 1974) to yield the results shown in Figure 2A. In this plot, each node (indicated by a horizontal branch) joins two subordinate nodes, with the bottom-most nodes joining two individual items. The vertical height of each node indicates the mean similarity between the two joined subordinate nodes, with distal nodes joined near the top and proximal nodes joined near the bottom.

Three aspects of the data are of interest. First, a substantial degree of semantic structure was recovered: the clustering algorithm found three broad groups, corresponding well to the semantic categories animal, artifact, and fruits. Second, there are considerable differences in the degree of subordinate structure apparent within these broader groups. Among the animals, the subcategories of birds and land animals are well differentiated from one another, whereas subcategories in the domain of fruits and artifacts are less well differentiated. Third, the man-made objects are much less similar to one another as a group than are the various animals or the set of fruits, which suggests that, considering verbal descriptions alone, artifacts form a much less cohesive grouping.

Next, we inquired to what degree items within each of the three broad domains identified previously tended to share the same verbal descriptors. To accomplish this, we adopted a method described by Garrard et al. (2001) and McRae et al. (1997). For each attribute listed for a given item, we calculated the proportion of items in the same general category (animals, artifacts, or fruits) that also share the property. This measure produces a rating on a

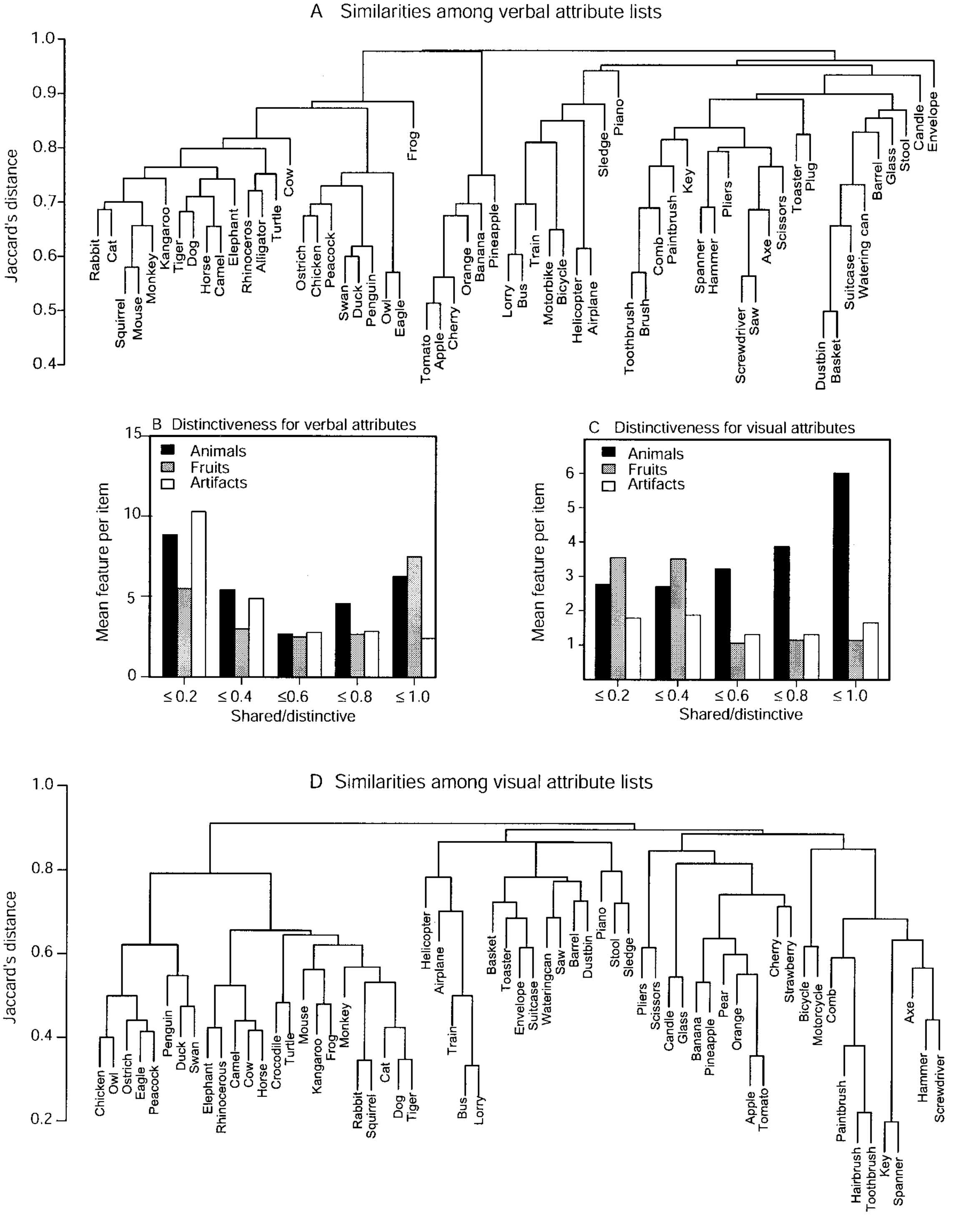

Figure 2. A: Hierarchical cluster analysis of the feature vectors for 62 items described by Garrard et al. (2001). B: Mean number of features per item across the shared–distinctive continuum from verbal attribute lists from Garrard et al. C: Distribution across the shared–distinctive continuum from the visual attributes that appear in drawings of animals, fruits, and artifacts. D: Hierarchical cluster analysis of the visual similarities among drawings of the same items.

distinctive-to-shared continuum. Low values indicate properties that apply only to a small number of category exemplars, and high values indicate properties that tend to be shared by many items in the same category. For the three broad categories, we examined the average distribution of properties by distinctiveness by dividing the distinctive–shared range (0–1) into five bins and assigning each of the properties of each item to the appropriate bin. Finally, we counted the number of properties in each bin for each individual item and averaged these figures across all items in each of the three broad categories.

Figure 2B shows the average number of features per bin for animals, fruits, and artifacts. It is clear that items in the three domains show considerably different propensities to share verbal descriptors. The artifacts show a strong positive skew, with most properties falling in the distinctive end of the distribution and with few shared properties. By contrast, the animal and fruit distributions are strongly bimodal, with almost as many features falling in the shared half of the range as in the distinctive half.

From these simple observations, the Garrard et al. (2001) data suggest that five important aspects of structure are apparent from verbal descriptions.

1. The general semantic categories, animals, fruits, and man-made objects, are easily discriminable from one another.

2. Fruits constitute a cluster that is distinct both from man-made objects and from animals. There is little similarity in verbal descriptions of fruits and those of animals, even though items in both categories are, in a sense, living things (or at least natural kinds).

3. Within the domain of animals, the subcategories birds and land animals are readily discriminable, whereas no very obvious subgroupings are apparent within either the artifacts or the fruits.

4. As a group, the artifacts are much less similar to one another than are the various animals or the fruits.

5. A comparatively high proportion of the verbal descriptors applied to a given animal are common to other animals, whereas few verbal descriptors are shared by the majority of artifacts.

Finally, if we are to base our model training environment on these observations, it is important to determine to what degree the data from Garrard et al. (2001) are robust to variations in the testing and scoring methods and to what extent the observations drawn from these data are representative of the object categories of interest. A full assessment of these issues would require analysis beyond the remit of this article, but evidence from the literature suggests that the observations from Garrard et al. (2001) are indeed reliable. All five points are consistent with other recent attribute-listing experiments, including a study of 93 object concepts described by Moss, Tyler, and Devlin (2002); an analysis of 60 concepts reported by Devlin et al. (1998); and a large norming study of 549 object concepts conducted by McRae and Cree (2002). All groups found that broad semantic domains are discriminable from verbal attribute lists, and all reported a higher degree of similarity among animals than among man-made objects, more distinct subclusters within the animal domain than within the artifact domain, and many more properties shared by animals than by artifacts. Neither Moss et al. (2002) nor Devlin et al. (1998) indicated whether fruits cluster with animals or artifacts, but McRae and Cree (2002) reported that these items (along with man-made foods, plants, and roots) tend to form a tight cluster that, as in Garrard et al. (2001), is quite distinct from animals and artifacts (although ultimately patterning with artifacts rather than animals). This consistency across studies suggests that the patterns observed in the aforementioned data will be apparent across a larger corpus of items and from different data collection methods. Hence, these are the aspects of structure we have captured in the verbal descriptions that appear in our simplified model-training environment.

Assessment of Visual Attribute Structure

To assess the degree of visual similarity that exists among various objects, we considered data that may be viewed as a visuospatial version of the attribute-listing task. Instead of verbally listing object properties, participants were asked to draw them. We then investigated the tendency for the 64 items from our semantic battery (all 62 items from Garrard et al., 2001, plus two additional vegetables) to share visual attributes with the drawings.

The method used to score the drawings was originally designed to assess the content of drawings produced by semantically impaired patients, and it is described in detail by Bozeat et al. (2003). Briefly, drawings of all 64 items were collected from 8 control participants. Two independent raters who were blind to the purpose of the study examined the drawings and decomposed them into lists of visual attributes. For example, in a drawing of a zebra, the raters might list the properties of body, neck, mane, head, ears, eyes, tail, legs, hooves, stripes. For each item, properties that were drawn by only a single control were dropped from the set, and the remaining properties were concatenated into a single list. Because we wanted to assess the visual similarities that exist among items in the battery independent of the words used to describe their parts, we examined the visual feature lists side-by-side with the original drawings. Visual features that had been labeled with different words but that were judged to be similar in appearance were classified as instances of the same feature, whereas those that had received the same label but were judged to be visually dissimilar were classified as distinct features.

From this large set of visual properties, we constructed a score sheet of the features appearing across all 64 items. Each individual drawing was then scored by ticking off the visual features that could be identified within it, yielding a vector of visual features for each drawing. The visual similarity between each pair of items was then calculated from the visual feature overlap, just as was done for the verbal attribute lists described previously.

Figure 2C shows the mean distribution of visual features across the shared–distinctive continuum for animal, artifact, and fruit categories. For animals, the majority of features were shared by category members, and there are relatively few distinctive properties on average. By comparison, artifacts and fruits are visually less complex, both having far fewer visual properties overall; and

Visual prototype patterns

```
     birds ++++++++ ++++++ 000000 000000000000 -- ---- ---- ----------------------
   mammals ++++++++ 000000 ++++++ 000000000000 -- ---- ---- ----------------------
  vehicles -------- ------ ------ ------------ ++ ++++ 0000 0000000000000000000000
hh objects -------- ------ ------ ------------ ++ 0000 0000 0000000000000000000000
     tools -------- ------ ------ ------------ ++ 0000 ++++ 0000000000000000000000
    fruits -------- ------ ------ ------------ ++ 0000 ++00 0000000000000000000000
```

Verbal prototype patterns

Perceptual descriptors

```
     birds +++++ +++++ 00000 0000000000000000 - ---- -------------------------
   mammals +++++ 00000 +++++ 0000000000000000 - ---- -------------------------
  vehicles ----- ----- ----- ---------------- + ++++ 0000000000000000000000000
hh objects ----- ----- ----- ---------------- + 0000 0000000000000000000000000
     tools ----- ----- ----- ---------------- + 0000 0000000000000000000000000
    fruits ----- +---- ----- 0000000000000000 + ---- 0000000000000000000000000
```

Functional descriptors / *Encyclopedic descriptors*

```
     birds +++ +++ 00 00000000 - - - -------------       +-+-----0000000000
   mammals +++ 000 ++ 00000000 - - - -------------       +--+----0000000000
  vehicles --- --- -- -------- + + + 0000000000000       -+--+---0000000000
hh objects --- --- -- -------- + 0 0 0000000000000       -+---+--0000000000
     tools --- --- -- -------- + 0 + 0000000000000       -+----+-0000000000
    fruits --- --- +- 00000000 - - + 0000000000000       +------+000000++++
```

Figure 3. Prototype feature vectors used to generate visual (top) and verbal (bottom) representation patterns for the model. Plus signs indicate units likely to be active for items in the category (turned on with $p = .8$), zeros indicate idiosyncratic units that are less likely to be active for items in the category ($p = .2$), and dashes indicate units that are never active for items in the category. hh = household.

both groups have far fewer properties shared by members of the same domain.

A hierarchical clustering analysis of these data is shown in Figure 2D. There are three points to note. First, the algorithm successfully discriminated animate from inanimate objects (i.e., artifacts and fruits), which suggests that these broad groupings are apparent from the visual appearance of the items alone. Second, in contrast to the verbal attribute-listing results, fruits were not well differentiated from the artifacts in this dataset. Third, animal subgroups are again well differentiated, whereas artifact subgroups are not: The individual birds are quite distinct from the individual land animals. In the domain of artifacts, the clustering algorithm identified several poorly differentiated clusters that do not correspond well to intuitive semantic categories.

The analysis suggests that, for the items we have examined, visual resemblances largely recapitulate information about similarity relations apparent from verbal attribute-listing studies: Man-made objects can be reliably discriminated from animals, more specific subcategories may be reliably discriminated within the domain of animals but not artifacts, more visual features are shared across animals than artifacts or fruits, and artifacts are less similar to one another as a group than are animals or fruits and vegetables. However, an important difference was found in the case of fruits and vegetables: Rather than comprising a distinct cluster well-separated from animals and artifacts, these items were not well differentiated from the man-made objects on the basis of visual features. The results thus suggest that, although visual appearances and verbal descriptions may provide useful sources of information about semantic similarity relations, the two channels can yield somewhat incongruent information for certain kinds of objects. In the simulation work, we see how this incongruity can lead the semantic system to acquire internal representations for such objects whose similarity relations differ from those expressed in either modality independently.

Constructing a Model Environment

On the basis of these analyses, we constructed a simple virtual environment to train the model, which captured the important aspects of structure apparent in visual appearances and verbal statements for the categories of animals, artifacts, and fruits. The environment consisted of 48 objects, half corresponding to living things and half to artifacts. These broad domains were further subdivided into the categories birds, mammals, fruits, vehicles, household objects, and tools. Associated with each item was a set of visual attributes, a name, and a verbal description, which were generated as follows.

Visual Representations

To create visual representations, we began with prototype patterns[1] for each of the six categories in the model's environment, shown in Figure 3. The prototypes were created to capture the finding from the norming data that some visual properties are likely to be shared by most items in a semantic domain, some are likely to be shared by items in the same subcategory within a domain, and others are likely to be idiosyncratic to individual items. The plus-marks in Figure 3 indicate visual properties that are likely to be observed in members of the category. Among the animal properties, some units are common to all animals, some are

[1] Our use of the word *prototype* is intended to refer only to the general pattern from which individual instance patterns were generated for the model and not to a supposed representational construct.

common only to the bird or the mammal categories, and others are idiosyncratic to individual items. Among the artifact properties, two are common to all artifacts, four are common to the vehicles, two are common to the tools, and the remaining properties are assigned as idiosyncratic. No property is held in common between animal and artifact items. Finally, the prototype for fruit items indicates that the fruits are likely to share five properties with one another, to share one visual property with artifacts generally, and to share one visual property with the tools.

To generate unique patterns to represent individual items in each category, we applied a mild distortion to the category prototype: Taking each feature in turn, we altered its state with likelihood 0.2. Thus, it was possible for a property typically shared by animals to be turned off for a particular animal (corresponding, for instance, to an animal with no legs, such as a snake) or for a property shared by birds (such as wings) to be turned on for a particular nonbird animal (e.g., a bat). This distortion was also applied to the idiosyncratic units: Each idiosyncratic unit appropriate to the domain was turned on with probability 0.2 to give each instance some unique identifying features. Eight exemplars were generated in this way for each of the six categories, with the added constraint that no two objects could have identical visual or verbal representations. For the category of fruits, the distortion of the prototype was applied solely across the artifact–fruit properties, to ensure that individual fruits were visually more similar to the artifacts than to the animals.

Verbal Representations

Verbal descriptions were created in the same manner as visual representations but with different prototype patterns shown in Figure 3. The prototypes are divided into different pools for visual, functional, and encyclopedic verbal descriptors. As noted previously, these distinctions have no functional consequence in the model. We arranged them this way to indicate that all verbal descriptors, regardless of the kind of information they refer to, are coded in the same manner. In contrast to the visual prototype, the verbal prototype for the fruits shares some properties with each of the other categories including the birds and the mammals. To generate individual fruit items, all properties in this prototype were distorted with likelihood 0.2, so that individual fruits could share idiosyncratic properties with both artifacts and animals. Both of these measures served to render fruits somewhat distinct from both animals and artifacts.

Finally, we ensured that there existed one verbal attribute that uniquely identified each of the categories (bird, mammal, vehicle, household object, tool) and domain (animal, artifact, fruit). These were intended as analogs of predicates that describe groupings of objects at different levels of specificity (e.g., *is living*, *is manmade*, *lives on land*, etc.), which would allow us to simulate sorting tasks as described in the next section.

This procedure provided us with a simple means of capturing the various aspects of similarity structure identified from the norming data reviewed in the previous section. Hierarchical cluster plots of the visual and verbal input representations are shown in the top and bottom parts of Figure 4, and the distribution of visual and verbal properties by distinctiveness is shown in the middle of this figure. For purposes of comparison to the norms discussed previously, note the following:

1. On the basis of property overlap, animals are distinct from artifacts in both input modalities.

2. Subcategories are distinct from one another within the domain of animals, but are less so within the domain of artifacts, in both modalities.

3. Animals are more similar to one another as a group than are artifacts.

4. Animals have a greater proportion of shared properties than do artifacts in both modalities.

5. The category of fruits forms a distinct cluster in the verbal-feature inputs, separate from both animals and artifacts, but is integrated with the artifact items in the visual-feature inputs.

Naming in the Model

In addition to visual attributes and verbal descriptors, each item was also given a name. In assigning names to objects in the model's environment, we considered two issues. First, although a given item may be named at any of several levels of specificity, participants in free-naming experiments are typically fairly consistent with respect to the level of specificity they choose for a particular item (e.g., they are likely to choose "dog" instead of "animal" or "collie"; see Brown, 1958). Second, the level of specificity at which participants prefer to name can vary across items. For example, robins, sparrows, and blue-jays are usually named as *birds* in confrontation-naming tasks with nonexperts (Rosch et al., 1976), whereas atypical but familiar category exemplars (like ducks, penguins, and ostriches) are usually named at this more specific level (Jolicoeur, Gluck, & Kosslyn, 1984) and completely unfamiliar objects may be named at a more superordinate level (e.g., wildebeests, ocelots, and marmosets may simply be named as *animals*).

To capture both of these aspects of naming behavior in the model, we assigned each item a single label, corresponding to the name it would be given in a free picture-naming task. However, across instances, the level at which an object was named could vary. For example, among the birds, three were given the same basic name (*bird*), and five were given a more specific name (*chicken*, *raven*, *swan*, *ostrich*, and *penguin*). Similarly among the mammals, five were assigned unique basic names (*cat*, *dog*, *mouse*, *goat*, and *pig*) and three were given the same superordinate label (*animal*); among the tools, five were given unique basic names (*hammer*, *screwdriver*, *wrench*, *saw*, *drill*), and three were given the superordinate name *tool*; and among the vehicles, five were given specific names (*car*, *lorry*, *boat*, *sledge*, *train*), and three were given the general name *vehicle*. Household objects and fruits were given a unique basic name and no more general name. When required to name an item from verbal description or visual input, the network was trained to activate a single name unit corresponding to the verbal label the item was assigned.

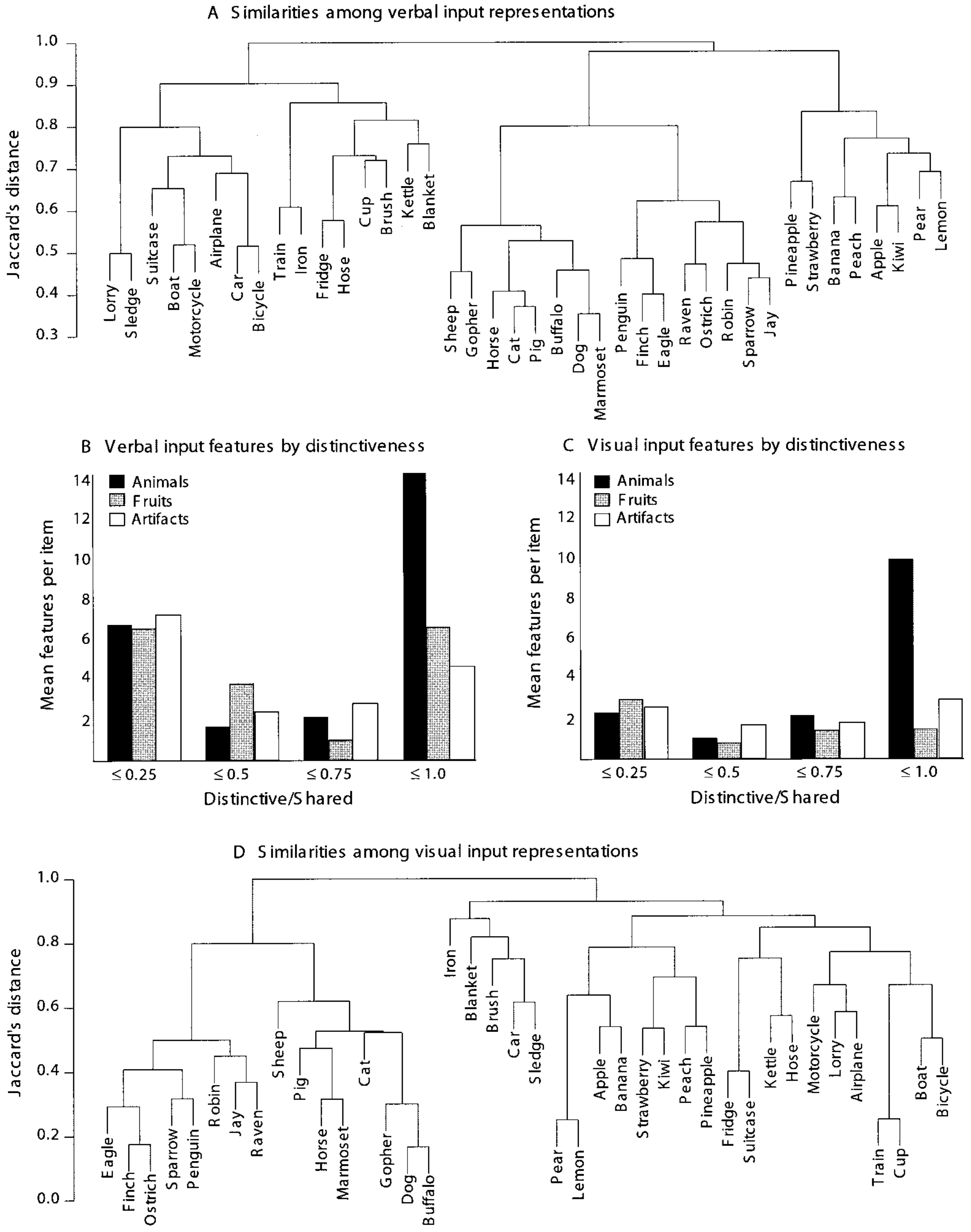

Figure 4. A: Similarities revealed by a hierarchical cluster analysis on the verbal input patterns constructed for the model. B: Mean distribution of features by distinctiveness for the verbal patterns. C: Mean distribution of features by distinctiveness for the visual patterns in the model. D: Hierarchical cluster analysis of the visual input patterns.

Second, the network was required to produce either a verbal description or a visual representation when given a name as input. Although most names uniquely identify a particular object in the network's environment, the four general labels (*bird*, *animal*, *vehicle*, *tool*) included in the training set do not. When one of these more general names was used as input, we trained the model by presenting it with a target pattern selected at random on each trial from among the set of items to which the name applied. For example, whenever the model was given the word *animal* as input, an individual animal was selected at random from among the 16 birds and mammals, and its target values were applied to visual and verbal units. As a consequence, the model learned to generate visual and verbal properties common to most animals when given the name *animal* as input. The names *bird*, *tool*, and *vehicle* were treated the same way.

Learning, Processing, and Representation in the Intact Model

Training Details

All units in the network were assigned a fixed, untrainable bias of −2, a parameter that has the effect of deducting 2 from each unit's net input. Thus, in the absence of input, each unit's activation settles to the low end of its activation range (approximately 0.19).

In each training trial, the model was presented with either a single name, a visual pattern, or a verbal pattern as input. Units in the corresponding input layer were hard-clamped to their input values, and the network was permitted to cycle for three time steps (in each time step, all units update their states four times). Inputs were then removed, and the model was permitted to cycle for two more time steps, at which time target values were applied across all visual and verbal units in the model, including the units acting as input during the trial. The model was permitted to cycle for two more time steps, recording the error on the relevant target units, at which point the error derivatives for all the weights in the network were calculated and the weights adjusted by a small amount to improve performance.

Every training pattern appeared once in each epoch, with the order randomized within epochs. The model was trained with a learning rate of 0.005, without momentum, and with a decay parameter set to 0.001 to prevent individual weights from growing disproportionately large. Training proceeded for 400 epochs, at which point the model had learned to generate a steady state for all inputs in which all verbal and visual units were within 0.05 of their target states.

Semantic Representation and Performance in the Intact Model

When the model has finished learning, it can take an input representation in any surface form (name, description, or image) and settle into a steady state in which all of the visible units are in the appropriate states. Associated with each input is a unique attractor state, corresponding to the appropriate pattern of activity across all visual and verbal units and some abstract pattern of activity across hidden units. Although the network is not trained to produce any particular pattern of activity across its hidden units, the representations it derives from the learning algorithm are structured in interesting and useful ways.

Figure 5 shows the results of a hierarchical clustering analysis performed on the network's internal representations for all of the name inputs. The four general names are printed in uppercase letters. There are three points of interest to note. First, aspects of similarity structure apparent in both verbal and visual modalities are recapitulated in the model's internal representations: Animal and artifact domains are well separated from one another, and subcategories within the domain of animals are also well differentiated, whereas subcategories of artifacts are less so. Second, the model has learned representations of general words that are similar to the individual item representations to which the general word applies. For example, its representation of *bird* is similar to its representations of *chicken*, *raven*, and so on. Third, the model has discovered representations for the fruit items that capture similarity relations not expressed in either the visual or verbal modality independently. Recall that, on the basis of verbal descriptions, fruits appear to constitute a separate cluster that is well separated both from artifacts and animals (although slightly more similar to animals; see Figure 4). On the basis of visual appearance, fruits are integrated with the cluster of artifacts. The model has learned representations of fruits informed by both of these structures. In Figure 5, the fruits form a cluster that is well separated from both artifact and animal items but that is considerably more similar to the artifacts than to the animals. Although not shown, visual inputs give rise to similarly structured internal representations.

Why are the model's learned internal representations structured in this way? The reason is that the information about similarity

Similarities among learned representations of names

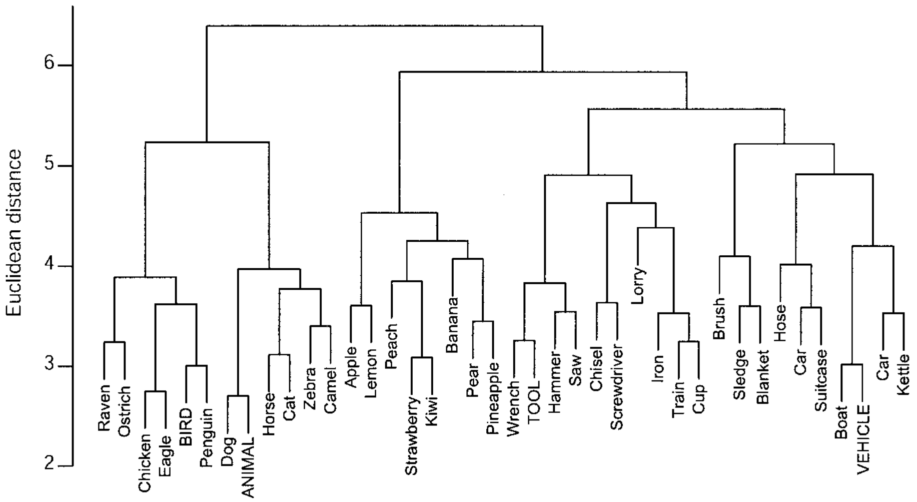

Figure 5. Hierarchical clustering analysis of the model's learned internal representations, with Euclidean distance as the measure of similarity. General names, which apply to more than one individual item in the training environment, are indicated with uppercase letters.

available to the semantic units, in its inputs and in the target signals that drive learning, is always filtered through the visual and verbal perceptual channels. Objects that have overlapping visual attributes, and that are described with overlapping sets of propositions, contribute similar inputs and targets to the semantic system throughout training. Such items produce similar weight changes throughout the network, with the consequence that they continue to generate similar internal states during training. Items that share few verbal descriptors and have different visual appearances contribute different inputs and targets to the system, and they generate differing internal representations (Rumelhart & Todd, 1993; Plaut, 2002; McClelland & Rogers, 2003; Rogers & McClelland, in press). When the model has learned, its representations recapitulate in abstract form the similarities among objects that are apparent across both domains. When different input modalities capture somewhat discrepant similarities among the same set of items (as is the case for the fruits), the structure apparent across the two modalities can differ from that expressed in either modality alone.

The ability of connectionist networks to extract and represent the similarity relations latent in their inputs and outputs has been treated extensively elsewhere (Hinton, 1986; Rumelhart & Todd, 1993; Rogers & McClelland, in press). The application of these ideas to the domain of semantics in the current simulation is of interest for the following reasons.

First, the simulation demonstrates that information about similarity that is apparent from verbal descriptions and drawings of objects is sufficient to support the acquisition of representations that capture intuitions about semantic similarity relations for animal and artifact categories. This is not particularly surprising, because the visual and verbal representations constructed for the model directly encoded these similarities. Indeed, several investigators have used attribute lists as the basis for constructing semantic representations in similar models (Cree, McRae, & McNorgan, 1999; Tyler et al., 2000). In our model, however, the semantic representations do not code these features explicitly. The activation of a given hidden unit does not directly correspond to an attribute such as *has eyes* or *can reproduce*, for instance. Instead, the hidden units may be understood as semantic units solely by virtue of the function they subserve: intermediating between visual representations and verbal statements and/or names. There is no explicit semantic store, except insofar as the configuration of weights permits the system to produce the appropriate response when probed with a particular input (see also Rogers & McClelland, in press). To determine what the model "knows," it is necessary to determine what visual or verbal responses it generates when given particular inputs; and we believe the same is true for the human semantic system. Thus, we propose that semantic representations are defined with reference to the function that they perform and not the content that they encode.

We believe that this approach represents an advance over feature-based models for two reasons. First, it circumvents certain difficulties of interpretation that are raised whenever semantic features are invoked. Which of an object's properties count as semantic and which are merely perceptual? Which are sufficiently useful or important to be included in a feature-based semantic representation? Should the dog's bark and the cat's meow count as different semantic features or as different instances of the same feature (e.g., makes a distinctive sound)? Such questions are critical to feature-based theories, but they are rendered moot when the representation units are not themselves construed as encoding interpretable information. Second, abstract semantic models are constrained to be somewhat more explicit about how particular semantic tasks are carried out. Because the activation of semantic units is not itself interpretable, the model must incorporate additional representations and processes designed to explain how semantic representations receive input and generate output. In principle, such constraints can allow the approach to address empirical phenomena from particular tasks in somewhat more specific detail, as we show in the next section.

The second point of interest raised by the model is the demonstration that the extraction of structure across modalities can lead to the discovery of new representational structure not reflected in either modality independently. The demonstration is important because it addresses the common-sense objection to association-based theories of semantic knowledge acquisition—semantic similarity relations are not always evident for all classes of objects from their visual appearances or from other perceptual information. Indeed, the utility of representational structure in semantics (under any theory) seems to lie precisely in the fact that semantically related items may be treated as similar, even when they have few directly perceptible properties in common. We believe that such useful similarity relations do consist in overlapping perceptual inputs, provided that these similarities are assessed across all modes of perceptual experience including language and across a broad range of episodes and events. Taking this perspective, the simulation illustrates how simple perceptual-learning mechanisms can give rise to representations whose structure differs from that apparent in any individual modality.

Finally, the particular representations discovered by the network, because they derive from training patterns modeled on the norming data described earlier, provide a basis for interpreting and predicting behavior in semantic tasks that use stimulus items from animal, artifact, and fruit categories. In the following section, we consider how the model can account for a range of phenomena in the study of disturbed semantic cognition by focusing on three aspects of the model's behavior that depend on the structure of the internal representations it acquires: (a) the high degree of similarity structure within the domain of animals relative to artifacts and the influence of such structure on the behavior of the model as the knowledge stored in its weights is degraded; (b) the mappings acquired by the model between its internal states and particular visual and verbal attributes, with an emphasis on understanding how the interactions between semantic and peripheral representations affect the model's behavior under damage; and (c) the implications of the counterintuitive suggestion that fruits are represented as similar to artifacts, despite sharing some important properties with animals.

Understanding the Structured Deterioration of Conceptual Knowledge in Semantic Dementia

In this section, we consider the behavior of patients with semantic dementia on each of four common tests of semantic memory: confrontation naming, word and picture sorting, word-to-picture matching, and drawing. A comparison of performance across the four tasks reveals remarkably consistent patterns in patient performance, which we take to reflect representational structure and processing mechanisms similar to those embodied in our model. To illustrate why, we also examine the model's performance on analogs of the same tasks, under simulated lesions of varying severity. In each case, we see that the model allows us to explain the patterns apparent in the patient data and that the explanation proffered by the model leads to new predictions about patient behavior.

Confrontation Naming

The most pervasive and self-evident impairment observed in semantic dementia is a marked anomia that grows increasingly severe as the disease progresses. Patients exhibit profound word-finding difficulties, and the patients' confrontation naming is dominated by two types of production errors. First, patients often produce a name that is correct but is more general than the label usually given by age-matched controls for the same object (e.g., *animal* instead of *dog*), which suggests that superordinate category names are more robust than are more specific labels (Hodges et al., 1995; Warrington, 1975). Second, highly familiar or typical names are often inappropriately extended to semantically related objects (e.g., *dog* for *pig*, *goat*, and *sheep*; Hodges et al., 1995).

These patterns have been well documented in longitudinal case studies and cross-sectional group studies (Lambon Ralph, Graham, Ellis, & Hodges, 1998; Lambon Ralph et al., 2001; Schwartz et al., 1979; Snowden et al., 1989; Warrington, 1975). In the first experiment of the current work, we analyzed the longitudinal naming performance of 15 patients with semantic dementia to determine how the distribution of observed naming errors varies as a function of the severity of semantic impairment.

Patient Method

Except where noted, all patients in this and subsequent experiments were identified at the Memory and Cognitive Disorders Clinic at Addenbrooke's Hospital, Cambridge, United Kingdom, and were diagnosed according to criteria described previously (Hodges et al., 1992, 1995). Specifically, all patients presented with anomia, impairment in single word comprehension, and impoverished semantic knowledge, with relative preservation of phonology, syntax, visuospatial abilities, and day-to-day memory. Structural brain imaging by magnetic resonance imaging showed focal atrophy that involved the polar and inferolateral regions of one or both of the temporal lobes in all cases.

Naming was assessed using the 48 line drawing from the original Hodges semantic battery (see Hodges, Salmon, & Butters, 1992), which depicts items drawn from three categories of animate objects (birds, water creatures, and land animals) and three categories of artifacts (household objects, vehicles, and musical instruments). Patients were shown each drawing in a random order and were asked to name it. Responses were classified as (a) correct when the patient gave the same response typically provided by controls, (b) a superordinate error when the patient gave a correct but more general response than that provided by controls, (c) a semantic error when the patient gave an incorrect response from the same semantic domain as the correct response, (d) a cross-domain error when the patient gave an incorrect response from the wrong semantic domain, and (e) an omission error for all remaining errors.

The vast majority of omission errors were cases in which the patient was unable to provide any name for the objects (although sometimes they would attempt to describe it). A very small number of visual errors were also grouped into this category.

Each patient was tested at least once, and in most cases patients were tested several times over a span of several years. On average, each patient participated in 3.8 testing sessions; the greatest number of testing sessions with a single patient was 10. The total number of sessions across patients was 57.

Model Method

To simulate the cortical atrophy underlying semantic dementia, we simply removed an increasing proportion of all the weights in the model, a choice motivated by the fact that all weights are either intrinsic to the semantic layer or project into or out of this layer. To simulate confrontation naming, we presented the model with the visual input pattern corresponding to one of the items in the model's environment, allowed the model to cycle for three time steps, then removed the inputs and let the model settle to a steady state. We chose as the model's response whichever name unit was most active above a threshold of 0.5. If no unit exceeded this threshold by the time the network settled, it was considered to have given no response. Under this procedure, the trained, undamaged network always yielded the correct response.

The model was tested only with those items in its environment that had been assigned a unique specific name, excluding the fruits (which were not included in the patient testing materials). The model's and the patients' responses were classified in the same way. The trained model was lesioned 100 times at each of five levels of increasing severity, and the data were averaged across the runs at each level to ensure that the results did not depend on a chance lesioning of particularly informative weights (see Plaut, 1995, for discussion).

Results: Naming Errors for All Items

To examine how response patterns varied with severity of impairment on average, we tabulated the patient data in the following way. The results from a given patient in a single testing session were treated as an independent observation, and all such observations were divided into quartiles on the basis of the patient's overall naming accuracy during that testing session. Fifty-seven total observations were collected from 15 individual patients; hence, each quartile contained 14 observations, except for the lowest, which contained 15. Within each quartile, the total number of responses of each type (correct, superordinate error, semantic error, cross-category error, or omission error) was calculated across all items and patients. We then converted these sums to proportions by dividing them by the total number of naming responses made by all patients within the quartile.

The left side of Figure 6 shows proportions of each of the four error types plotted against overall naming accuracy (by quartiles) for the patients. The right side shows the same data for the model, also plotted against accuracy at the points where the model's total proportion correct most closely matched that of the patients in each quartile (with 10%, 20%, 25%, and 35% of connections lesioned, respectively).

The patients and the model show a qualitatively similar pattern of behavior: As the degree of impairment increases, so do the observed proportions of omission errors and to a lesser degree superordinate errors. By contrast, semantic errors initially rise with severity but then decline, with patients in the fourth quartile making fewer semantic errors on average than patients in the third. Relatively few cross-domain errors are observed at all.

The increasing proportion of omissions observed in the model's naming behavior results from the mappings the model has learned

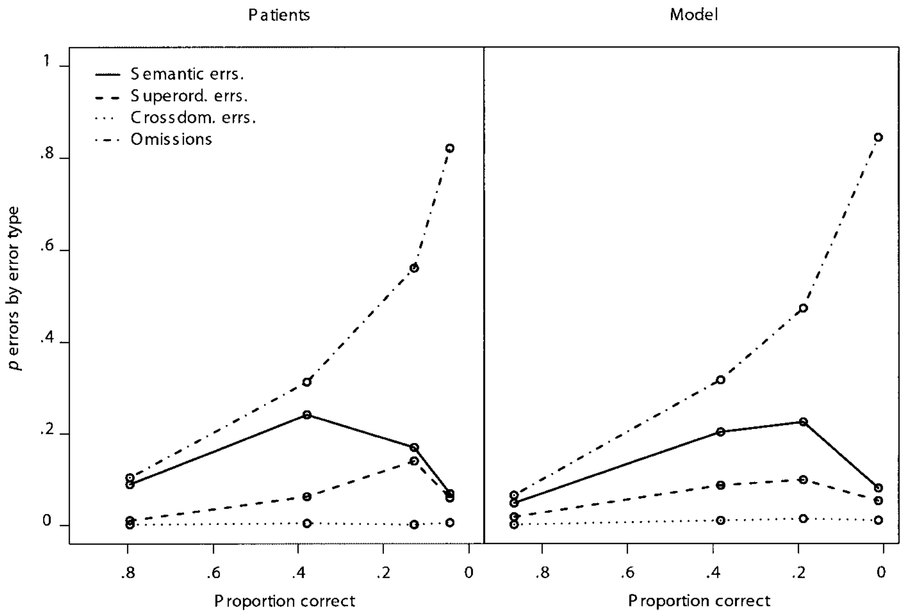

Figure 6. Picture-naming responses from averaged patient data and from the model, plotted against overall accuracy. Model responses were obtained with 10%, 20%, 25%, and 35% of connections lesioned. Semantic errs. = semantic errors; Superord. errs. = superordinate errors; Crossdom. errs. = cross-domain errors.

between its internal representations and the various name units. To name correctly, the model must activate a single unambiguous name unit when its internal state is proximal to a given item's representation, but it must refrain from activating this unit in response to other items in the immediate representational neighborhood. For example, the name *zebra* applies to the zebra, but not to other animals that receive similar representations in the model. Hence, the trained model activates this name only when its internal state is very similar to the learned representations of zebra. When the model's internal representations degrade as a consequence of lesioning some of its connections, the attractor into which it settles when given a visual input may drift out of the limited region of representation space from which it has learned to produce the correct naming response, and the model fails to activate the correct name unit or indeed any other.

Errors of commission result in the model as a consequence of an asymmetry in the vulnerability of idiosyncratic relative to general semantic information, coupled with the recurrent processing dynamics assumed by the theory. As the system's internal state drifts away from the circumscribed region to which an idiosyncratic property applies, the attractor corresponding to the item's representation may collapse, and the model can fall into a neighboring attractor. For example, to maintain stable internal states that distinguish zebra and horse representations, the semantic system must engage interactions with peripheral layers in which the visual features and verbal descriptors that differentiate zebras from horses are activated (e.g., *has stripes*). However, the features that are specific to the zebra, and not shared by the horse, are only active when the model's internal state is proximal to its learned representation of zebra. If, as a result of damage, the model moves away from this state when given the input for zebra, it generates patterns across visual and verbal layers that are indistinguishable from the patterns appropriate to the horse, and recurrent processing pushes the model's internal state toward its representation of horse instead of zebra. As a consequence, under moderate amounts of damage, the model occasionally produces an incorrect but semantically related name.

This account explains not only the increasing proportion of omissions and semantic errors with moderate semantic impairment but also the observed drop-off in semantic naming errors with severe impairment. With increasing damage, the model becomes unable to generate any information that individuates items from the same broad domain, and representations within a given domain collapse into a single general attractor from which the model produces only those properties common to the majority of items in the domain. From this degraded state, the model is unable to generate any individual name. However, the state is similar to that produced in the intact network by very general category names (e.g., *animal*, *bird*, *tool*, *vehicle*), and so even under severe amounts of damage the network is occasionally capable of producing these labels as output. The model never names an object with a completely unrelated label, because such names apply only to objects with very distal internal representations.

We should note that the naming performance of patients with semantic dementia is strongly influenced by concept familiarity and word frequency. We did not manipulate these parameters in the simulation, because we were primarily concerned with demonstrating effects of representational structure on naming in the model. Much past research, however, has shown that neural-network models that acquire internal representations are also strongly sensitive to frequency information, and in general we would expect such effects in the model to parallel those observed in patient data (see e.g., Rogers & McClelland, in press).

Prediction: Naming Errors Should Vary in Different Domains

The explanation of naming errors offered by the model relies on the similarity structure of the representations acquired by semantics and the ability of the recurrent processing dynamics to generate output responses from different regions of the semantic representation space. One prediction offered by this explanation is that error types should vary depending on the density of the semantic neighborhood. Specifically, domains with a high degree of similarity structure offer more opportunities for the semantic system to be "captured" by incorrect attractors and, hence, more opportunities to make errors of commission. Unstructured domains offer fewer such opportunities, and consequently we would expect to see a greater proportion of omission errors in such domains.

Results: Naming Errors by Domain

To test this prediction, we tabulated the proportion of errors of each type separately for animal and artifact items in the model and in the patient data. Both are shown in Figure 7. The model data confirm the intuitions articulated earlier: As damage increases, errors of omission are more likely to occur in the domain of artifacts at all levels of severity, whereas errors of commission

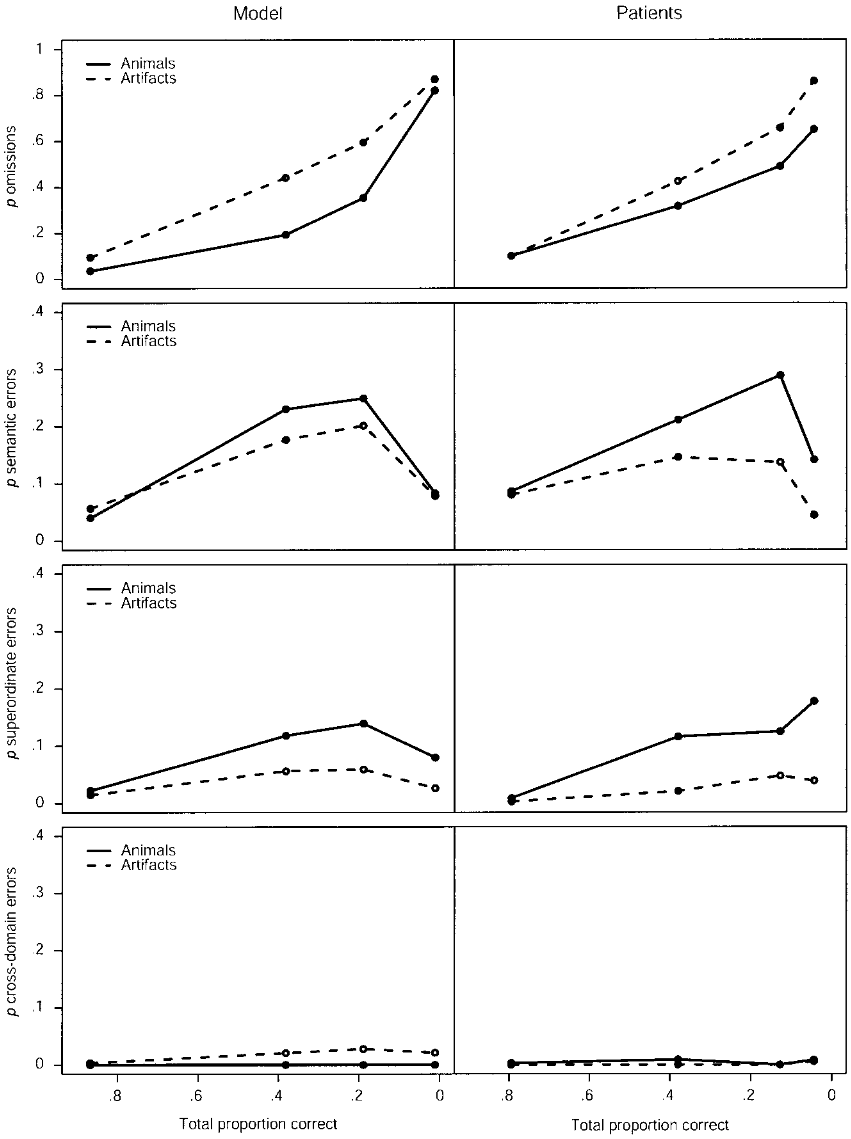

Figure 7. Picture-naming errors split by domain (animals and artifacts) for the model and for the patient data, which show qualitatively similar patterns of errors.

occur relatively more frequently for animal items. The results from the patient analysis closely match the predictions of the model.

Sorting Words and Pictures

According to the model, errors of commission in naming occur when the semantic system has difficulty maintaining distinct internal representations for semantically related items. This explanation is consistent with studies of sorting, which typically find that semantically impaired patients perform better when sorting items into more general relative to more specific semantic categories (Hodges et al., 1995; Warrington, 1975). This suggests that they have more robust access to information that distinguishes broad semantic domains than to information that individuates more specific groupings. As in naming, these results have been reported in longitudinal case studies and in cross-sectional group studies. We examine word and picture sorting data collected longitudinally from a group of patients with semantic dementia to determine how sorting performance varies with the magnitude of semantic impairment, the level of specificity of the sorting criteria, and the modality of testing (i.e., words or pictures).

Patient Method

Twelve patients with semantic dementia participated in the picture-sorting task, and 8 participated in the word-sorting task. As in the confrontation-naming task, multiple observations were collected from each patient, but these were treated as independent in the data analysis. Forty-one total observations were collected across patients in the picture-sorting task, with a maximum of 10 observations from a single patient. In the word-sorting study, 23 observations were collected, with a maximum of 7 from a single patient.

Stimuli for the sorting task consisted of the same 48 items used in confrontation naming, half animals and half man-made objects. Line drawings of the objects were used in picture sorting, and cards with the objects' names printed on them were used in word sorting. Both tasks used the same procedure, which included two testing conditions: a general sorting and a specific sorting condition. In the general sorting condition, patients were asked to sort all 48 items into the categories living thing and man-made object. In the specific condition, they first sorted the animals into the categories air creature, land animal, or water creature, and next sorted the artifacts into the categories vehicle, household object, and musical instrument. In each case, written category labels were placed in view of the participants and on every trial the experimenter verbally stated the category name and simultaneously pointed to the corresponding label on the table top. If the patient was unable or unwilling to make a response, the experimenter provided a prompt, such as "Do you think this is something that lives in the water, in the air, or on land?" We scored performance on each test by calculating the proportion of items that were placed in the correct category.

Model Method

When we constructed the verbal-description patterns for the model, we ensured that there was one verbal descriptor that uniquely identified each semantic domain (living or man-made) and category (bird, mammal, fruit, vehicle, household object, or tool). These units were intended to stand as proxies for the verbal category labels provided by the experimenter in the sorting task. For convenience, we refer to them as *domain* and *category units*, respectively. To simulate sorting, we presented the model with an input pattern (either a visual image or a name), allowed it to settle, and inspected the states of the domain and category units to determine to which group the network assigned the item. For example, to simulate general picture sorting, the network was presented with a visual input pattern, and the stimulus was categorized as living or man-made depending on which of the corresponding domain units was most active. The same procedure was used to simulate specific sorting, but activation was assessed across the more specific category units rather than the domain units. Because the patient experiment did not use fruit items, we excluded these from consideration in the first simulation; however, we consider how the network sorts fruits in the next simulation. Hence, we assessed sorting in the model for 16 animals and 24 artifacts at general and specific levels for both words and pictures under increasingly severe simulated lesions. Fifty damage trials were conducted, and the results we report are averaged across these runs.

Results: Sorting at Different Levels

Each patient-testing session was treated as an independent observation, which yielded a separate score for sorting at general and specific levels. Patients were ordered according to the degree of their semantic impairment at the time their sorting behavior was assessed, as measured by the patients' word-to-picture matching score during the same testing session. To examine the central tendencies of the data across the spectrum of severity, we then averaged together every four such observations at each level of specificity, as well as the corresponding word-to-picture matching score. Thus, each data point shown in the results corresponds to the average of four individual observations, each having comparably severe semantic impairment.

The results for both the patients and the model are shown in Figure 8. The right column shows smoothed patient data plotted against severity of impairment for sorting both words and pictures at the general and more specific level of granularity.

There are four aspects of the data to note. First, overall performance deteriorates with the severity of impairment. Second, in both the word- and picture-sorting tasks, patient performance is less impaired for general relative to specific sorting at all degrees of severity. Third, picture sorting is more robust than word sorting at all levels of severity, especially for the general level. Fourth, the degree of discrepancy between general and specific levels is greater in the picture-sorting task than the word-sorting task. These patterns also characterize the simulation data.

The patient data show that the robust preservation of more general knowledge about objects reported in past studies may be observed across the spectrum of disease severity in semantic dementia. The model links the phenomenon to the same factors that contribute to production errors in naming. The expression and comprehension of general facts about broad semantic categories (such as whether an object is living or man-made) is relatively robust, because the intact system has learned to generate such information from a broad range of contiguous internal states. Hence, the effect of damage must be quite severe before the system begins to generate incorrect verbal information about such properties. Properties that reliably discriminate narrower categories by definition span a more restricted range of the representation space and are more vulnerable to damage, as are specific names.

The model also explains why sorting performance is better for pictures than for words at all degrees of impairment. Relative to a picture, a single word provides less constraint on the representa-

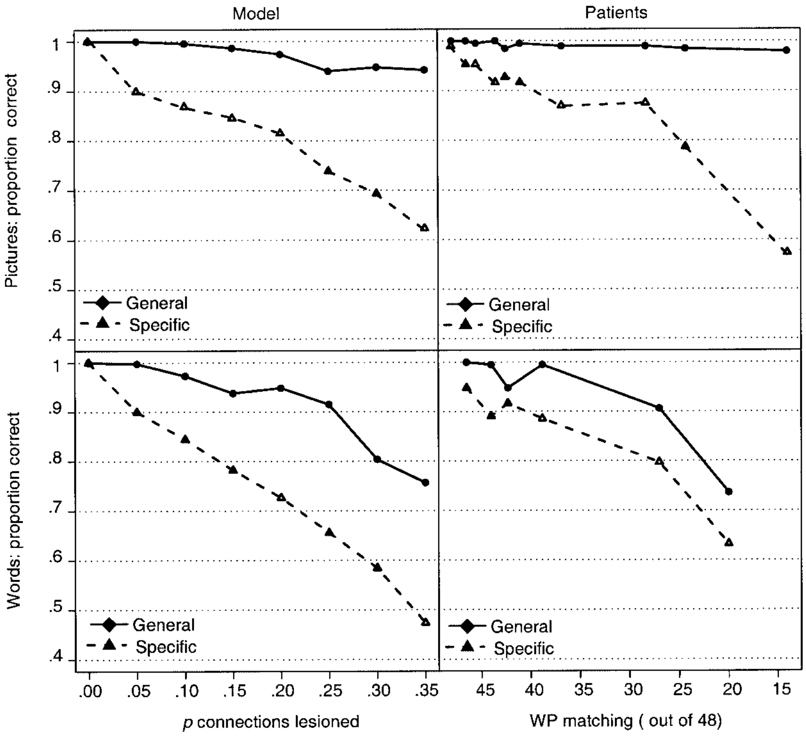

Figure 8. Average patient and model data for sorting pictures (top) or words (bottom) at two levels of granularity, general and specific. WP = word-to-picture.

tions that arise in semantics. Whereas individual visual features in an image can independently provide some information to the system about an object's identity, the subcomponents of an individual name (such as its constituent letters or phonemes) are not systematically related to the identity of the object the word denotes and thus do not provide such constraints. In the model, this difference is implemented by representing individual words with single units and representing images with distributed patterns of activity across visual feature units.[2] As a consequence, the relationship between visual and semantic representations is partially systematic, in that items with similar sets of visual attributes are also likely to have similar semantic representations. This systematicity arises in the model because its learned internal representations derive from the similarities expressed across visual and verbal inputs: For animals and artifacts, visual images and verbal descriptions capture the same similarity relations, which are simply recapitulated in the model's internal representations. The relationship between individual names and semantic representations, however, is arbitrary, because single names are represented with single units in the model. Hence any two individual names are represented with perfectly nonoverlapping inputs and capture no degree of similarity. It is this difference in the nature of the mapping between surface form and conceptual representations that underpins the difference in performance for word and picture sorting. Arbitrary mappings are more vulnerable to damage than are systematic mappings, as has been demonstrated in many other domains (Lambon Ralph & Howard, 2000; McGuire & Plaut, 1997; Plaut, McClelland, Seidenberg, & Patterson, 1996).

Prediction: Fruits Are Sorted With Artifacts

The explanation of sorting data proffered by the model suggests that knowledge about the more general semantic properties of objects may not be more robustly preserved for unusual groups of objects whose semantic neighbors do not share the property in question. The category of fruits constitutes one example of such a group. Although fruits share some general properties with animals (e.g., they are *living*, or at least *natural kinds*, and not *man-made*), our model suggests that, by virtue of having visual attributes in common with simple man-made objects, they may be represented as more similar to artifacts than to animals by the semantic system. This suggestion has counterintuitive implications for sorting of fruits at different levels of specificity. Under damage, representations of individual fruits may migrate toward artifact representations and away from representations of other living things such as animals. If this happens, the system may incorrectly generate verbal descriptors that are generally true of artifacts when given a fruit as input, but should rarely generate descriptors appropriate to animals. In other words, the usual finding that patients are better at

[2] In reality, we believe that the arbitrary mapping between individual words and semantic representations arises from the lack of systematicity between phonological (or orthographic) surface representations of words and semantics and not from the local representation of individual lexemes (e.g., Plaut et al., 1996). However, the use of local word representations in the model provides a useful proxy for capturing this arbitrary mapping.

sorting into general rather than specific categories should be reversed for the category of fruits.

We tested this prediction in the model and in the patients by replicating the previous sorting experiment with the 64-item semantic battery used by Garrard et al. (2001) that, in addition to categories of animals and artifacts, includes a category of fruit and vegetable items. As before, patients were first asked to sort all 64 items into living and man-made categories, and then were asked to sort items from each domain into more specific categories (land animals, air creatures, or fruits and vegetables or vehicles, household objects, and tools). We then tabulated accuracy separately for fruit items and for animal and artifact items.

Results: Sorting With Fruits Included

Data from the model and from the patients are shown in Figure 9. As predicted, the model shows a reversal in the tendency for general information to be more robust than specific information for the fruit items, both in word and picture sorting. At all levels of severity, the model performs better at specific relative to general sorting. For animal and artifact items, the reverse is true, that is, the usual advantage for the general level applies.

This pattern is also strongly evident in the patient picture-sorting data. Indeed, the predicted effect is much stronger in the patients than in the model, with the most severe participants scoring at or below chance when categorizing pictures of fruits as living or man-made but at ceiling when categorizing them as land animals, air creatures, or fruits and vegetables. The predicted pattern is less clearly apparent in the word-sorting data, although this may be due partly to the paucity of the data from severely impaired patients. With only two and three observations in the lowest two quartiles, there is little power to detect differences in accuracy between sorting levels. The difference between general and specific levels was in the predicted direction for the 2 patients in the third quartile but was not in the most severe quartile.

The magnitude of the effect in the picture-sorting data may reflect the influence of other perceptual factors that contribute to representational structure in semantics, which are not implemented in the model. For example, the actions afforded by fruits, and the contexts in which they are encountered, may contribute structure to acquired semantic representations that render them even more similar to small manipulable artifacts than is captured by our model (see McRae & Cree, 2002, for evidence supporting this idea). Nevertheless, the data clearly support the prediction that the typical pattern of spared sorting for more general semantic categories may be reversed when the sorting criterion does not map systematically onto the similarity structure of the domain (as assessed across multiple modalities).

Word-to-Picture Matching

A third common measure of semantic impairment is the word–picture matching task, in which patients are presented with a spoken word and an array of pictures and are asked to choose the picture that matches the word. Word–picture matching deteriorates with the disease progression in semantic dementia and correlates strongly with performance on many other tasks that tap semantic memory (Patterson & Hodges, 2000; Bozeat et al., 2000). In our work, and following many other researchers, we have frequently taken word–picture matching performance as a general measure of word comprehension under semantic impairment (e.g., Rogers, Lambon Ralph, Hodges, & Patterson, in press b).

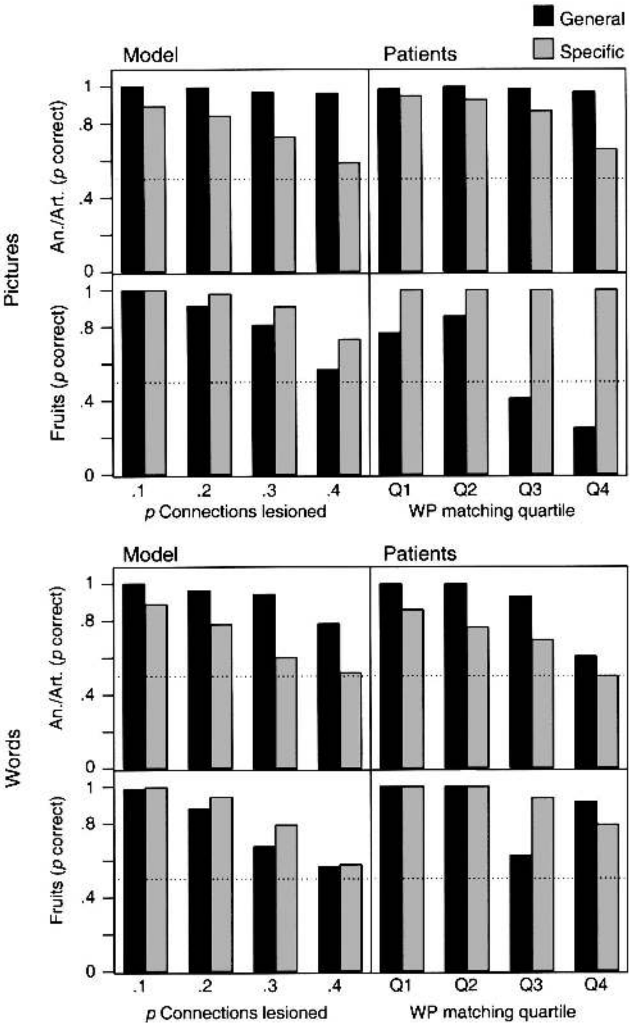

Figure 9. Model and patient data showing mean proportion correct across animals and artifacts (An./Art.) or across fruits, with sorting at general (living/man-made) and more specific (for living things: land animal, air creature, or fruit–vegetable; for artifacts: vehicle, household object, or tool) levels for pictures (top) and words (bottom). The dashed lines indicate chance performance for the general sorting condition. WP = word-to-picture; Q = quartile.

We have suggested that impairments in sorting and naming arise because patients have difficulty maintaining the distinctions between items represented as similar in the semantic system. It is easy to see that this explanation extends to word–picture matching as well. In this view, both words and pictures give rise to some internal state in semantics, even when semantic knowledge is severely degraded. Deficits in word–picture matching arise when the target word and some or all of the pictures in the array produce indistinguishable internal states. This explanation supports the prediction that the comprehension of a given word as assessed by word–picture matching may appear to be greatly compromised when items in the picture array are semantically related to the target, but it may seem relatively spared when distractor items are semantically distal to the target. Funnell (1996) has described results consistent with this prediction: Patient E.P. performed better at a word–picture matching test when distractors were unrelated to the target word than when they were drawn from the same category. In the current work, we tested the prediction using a two-alternative forced-choice word-to-picture matching paradigm in which we systematically varied the semantic distance between the target and distractor items.

Patient Method

Two patients with severe semantic impairments were selected for testing (M.S. and D.C., described previously by Graham, Lambon Ralph, & Hodges, 1997, and Lambon Ralph, Ellis, & Franklin, 1995, respectively). Eighty target stimuli (e.g., *duck*) were presented to each patient on four separate occasions, with a close (e.g., *penguin*), a dissimilar (e.g., *frog*), a distant (e.g., *goat*), or an unrelated (e.g., *trumpet*) foil. For comparison to the model data, we report performance in the close, distant, and unrelated conditions.

On each trial, patients were shown one written word and two line drawings (selected from a variety of corpora), including the target and one foil. The experimenter read the word aloud and asked the patient to decide which of the two pictures matched the word. We scored each trial as correct or incorrect, and we calculated the mean proportion correct in each distance condition (close, dissimilar, distant, or unrelated) for both patients. In a given session, patients saw each target item only once, and the foils in each semantic distance condition were counterbalanced across sessions. At least 1week elapsed between testing sessions.

Model Method

The task was simulated by presenting the model with a target name as input and then with a series of visual inputs corresponding to the target and distractor pictures in the task. In each case, we recorded the states of the semantic units after the model had settled. We chose as the network's response the visual input that generated an internal representation most similar to that produced by the name (using Euclidean distance as the measure of similarity).

The model was tested with visual distractors that varied in their degree of contrast with the target item. In the close condition, the target and distractor items were drawn from the same category (e.g., both birds). In the distant condition, the distractor was selected from an alternate category in the same domain. In the unrelated condition, the target and distractor were selected from different domains. We tested only those items that were given a unique name in the network's environment, and we also excluded fruit items as these did not appear in the patient-testing materials. In all conditions, we tested every possible pairing of target and distractor and calculated the proportion of trials on which the model performed correctly. Because there are a larger number of artifacts than animals that have a uniquely identifying name (18 instead of 10), we first calculated the mean proportion correct within these domains, and then averaged these proportions (to balance the influence of animal and artifact items). We again performed the simulation 50 times at each level of damage and report the model's average behavior across these trials.

Results

The left panel of Figure 10 shows the model's performance longitudinally, when semantic distractors are chosen from the close condition (as is typically the case in word–picture matching tests). Its performance deteriorates with increasingly severe lesions. The middle panel shows the model's performance in all three conditions when 20% of its connections have been lesioned, and the right panel shows the mean accuracy in each trial condition for M.S. and D.C. Both patients and the model performed the worst for trials that required them to differentiate objects at a specific grain, they were best when required to differentiate unrelated objects, and they were somewhere in between in the intermediate condition.

The patient data suggest that word comprehension under semantic impairment is not an all-or-none phenomenon, with particular words losing all meaning as the concepts to which they refer

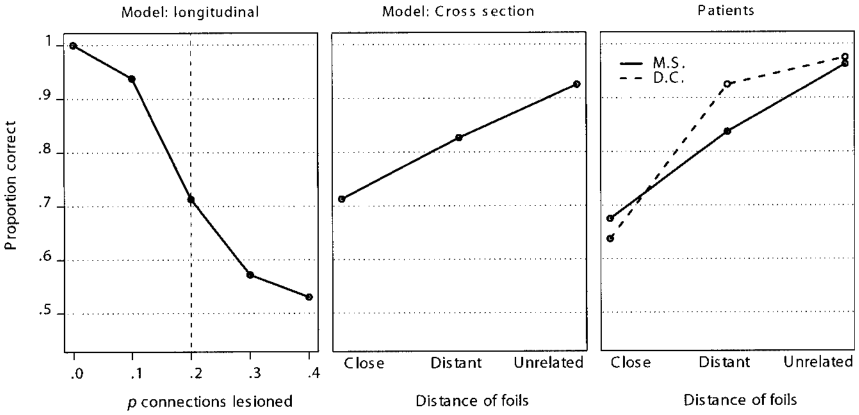

Figure 10. Word-to-picture matching data for model and patients in a two-alternative forced-choice paradigm in which the distractor varied in its semantic relatedness to the target. The left panel shows longitudinal performance of the model in the close condition; the middle panel shows performance of the model in all three conditions when 20% of its connections are lesioned; and the right panel shows cross-sectional data from patients M.S. and D.C. in all three distance conditions. The dashed vertical line in the left panel indicates the point at which the cross-sectional data in the middle panel were taken.

degrade, but is somewhat more graded in nature. The distributed representations that inhere in the model provide a natural way of thinking about this phenomenon. All verbal and visual inputs give rise to some internal state in semantics. When knowledge degrades, this state may deviate strongly from the correct representation, but it still carries with it information that may be used in comprehension. Although the recurrent dynamics familiar from previous simulations lead the degraded system to generate similar internal states for semantically related objects, items from very different semantic domains continue to produce discriminable representations in semantics until very late in the disease progression. In this sense, words and pictures continue to generate "meanings" even when knowledge is severely degraded, but the meanings derived for semantically related objects grow increasingly indistinguishable from one another.

Drawing and Delayed Copying

Thus far, we have considered tasks that tax verbal production and verbal and visual comprehension. In the final set of experiments, we consider a semantic task that requires the production of visual information, namely drawing and copying of line drawings of objects. Drawing is widely used as a clinical assessment tool for investigating such disorders as constructional apraxia or neglect, but it has been rarely used to test semantic memory. A few studies have revealed that patients with semantic dementia can have great difficulty producing drawings of meaningful objects when given their name or when required to copy them under conditions of delay (Lambon Ralph & Howard, 2000; Lambon Ralph, Howard, Nightingale, & Ellis, 1998). A full assessment of drawing was not the primary aim of these experiments, and the data they report were not analyzed quantitatively. In this section, we describe one attempt at such an assessment and consider whether the patterns of behavior we have seen in naming, sorting, and word–picture matching may also be found in drawing.

1. Are the rates of omission and commission errors different for items in different semantic domains, as has been shown in naming?

2. Is specific visual information more vulnerable than information about visual properties shared by items in the same domain, as found in naming, sorting, and word–picture matching?

3. Are intrusions in drawing more likely to occur for visual properties shared among an item's semantic neighbors, as suggested by our account of production errors in naming?

Patient Method

Data were collected in conjunction with another drawing study conducted simultaneously, which is described in detail in Bozeat et al. (2003). Drawings were solicited from 4 control participants and 3 semantic dementia patients under three task conditions: immediate copy, in which the participant was permitted to look at the stimulus while drawing it; delayed copy, in which the participant was asked to reproduce the drawing from memory after counting from 1 to 15; and drawing to name. Control participants were 4 volunteers recruited at the Medical Research Council, Cognition and Brain Sciences Unit in Cambridge, who were matched in age to the patients. The 3 patients, (D.S., D.C.,[3] and I.F.) were selected as representative of the early, middle, and late stages of the progressive disorder, respectively. Drawings were collected in several sessions over the course of a year. We assessed both the patients' and the control participants' ability to draw 56 items taken from two animal categories (16 land animals and 8 birds) and 3 artifact categories (16 household objects, 8 vehicles, and 8 tools). These were the same items used to test sorting, excluding the fruits (see Bozeat et al., 2003, for further detail).

Both patient and control data were scored using the same visual feature score sheets used to assess visual feature overlap in the *Assessing Verbal and Visual Structure in the Environment* section, above. Any visual feature on the score sheet that could be identified in a given drawing was checked off on the score sheet. Features that appeared in a drawing but that could not be identified were noted separately, but such features were rare both in the control and patient data and are not considered further. Thus, the visual features that appeared in each individual drawing were coded in a vector of length 279, with each element indicating whether a particular feature was evident in the drawing.

The patients' performance was assessed relative to the control data in the following way. For each item, we examined the four control drawings and discarded any features that were included in some control drawings and not others. The remaining features were designated as targets if all four controls included the feature in their drawings and as nontarget features if the feature was omitted by all four controls. For example, the feature *eye* was designated as a target feature for drawings of individual animals and as a nontarget feature for drawings of individual artifacts.

The features appearing in each patient drawing were then classified in the following way: (a) as correct if it was a target feature that the patient included (e.g., drawing wings on a swan), (b) as an omission if it was a target feature that the patient failed to include (e.g., drawing a swan without wings), and (c) as an intrusion if it was a nontarget feature that the patient included (e.g., drawing four legs on a swan).

Model Method

To simulate drawing from long-term memory in response to an object's name and delayed copying in the model, we gave the network an input (either a name or a visual pattern), allowed it to cycle for three time steps, and then removed the inputs and allowed the model to settle to a steady state. We then examined the pattern of activity across visual units and considered the network to have "drawn" those attributes whose activity exceeded a threshold of 0.5. The model's performance was scored in a manner analogous to the patients' performance. Target features were defined as those visual units activated by the intact network for a given item, and the remaining visual units were considered nontarget features. For each stimulus, features were classified as correct, an omission, or an intrusion, exactly as was done for the patient data. In the drawing task, the model was only assessed on those name inputs that uniquely identified a given item.

In the model, the presentation of a stimulus is simulated by hard-clamping the corresponding input units. The analog of the immediate copy task in the model, then, would be to look at the states of the visual units while they are hard-clamped in the damaged model and to compare them to the hard-clamped states of the same units in the intact model. However, because the model is always given the correct input, it would always be right in this case. For this reason, we report the model's performance on just the drawing and delayed copy tasks.

[3] The patient D.C. reported here is not the same patient D.C. reported in the word-to-picture matching study in the *Word-to-Picture Matching* section, above.

We tested the model's drawing and delayed copy performance 50 times at each of three levels of damage. The data reported here are averaged across these runs.

Results

Overall accuracy by task. To assess overall accuracy, we calculated the total number of errors (omissions and intrusions) for each picture, for both the model and the patients. The results are shown in Figure 11. The patients made comparatively few errors in the immediate copy task, which demonstrates that they have visuospatial and executive resources sufficient to the task. All patients made considerably more errors in the delayed copy and drawing-to-name tasks.

The data were assessed with a univariate analysis of variance (ANOVA) in which each picture was treated as a single data record, with the total number of errors per picture as the dependent variable and with task (immediate copy, delayed copy, or drawing) and patient (D.S., D.C., or I.F.) as fixed, independent between-case factors. There was a strong main effect of drawing condition, $F(2, 425) = 96.4$, $p < .001$, with all 3 patients making the fewest errors in the immediate copy condition, significantly more errors in the delayed copy condition (post hoc contrast of marginal means, $p < .001$), and the most errors for the drawing-to-name condition ($p < .001$, compared with the delayed copy condition). There was also a strong main effect of patient, $F(2, 425) = 49.5$, $p < .001$, with best performance for the least severe patient (D.S.) and worst performance for the most severe case (I.F.). This effect of severity interacted significantly with task condition, $F(4, 425) = 8.5$, $p < .001$, with all patients performing relatively well in the immediate copy condition but differently from one another in the other two tasks. The model shows a qualitatively similar pattern of results.

Omissions and intrusions by domain. Our earlier consideration of naming revealed that semantic dementia patients were more likely to make errors of commission when naming animals and errors of omission when naming artifacts. The model suggests that this behavior arises when the semantic representations generated by a visual stimulus drift from the correct state into neighboring attractors, from which the system has learned to produce information that is inappropriate to this stimulus. If this explanation is correct, we should expect to see similar domain differences in the production of visual information in drawing: more omissions for artifacts and more intrusions for animals.

To test this prediction, we tabulated the proportion of features that were omitted in each model "drawing" and in each actual patient drawing, as a proportion of the total number of target features present in the item drawn. Figure 12 shows this proportion averaged separately for animal and artifact items, across 50 trials of damage at each level of severity for the model and for each patient.

Both the model and the patients show an increasing tendency to omit visual properties as knowledge degrades. A greater proportion of attributes are omitted in the drawing task than the delayed copy task at all degrees of severity. As expected, the model omitted a greater proportion of features on average for artifacts than for animals in both conditions, but the same was not true of the patients, a point that we consider in the following discussion.

We assessed the model's propensity to add inappropriate features to a drawing by taking the total number of intrusions as a proportion of the total number of nontarget features for each item. We then averaged this statistic separately for animal and artifact items across 50 damage trials at each of four levels of damage. Analogous data were also tabulated for each patient.

Nontarget features are defined as those that should not be produced in a given drawing. For any particular item, this constitutes the majority of the visual features. Consequently, the proportion of intrusions for each item is relatively small. Nevertheless, interesting patterns are observed in the data. Figure 13 shows the mean proportion of intrusions made by the model at three levels of damage and by the 3 patients for animal and artifact items. For both domains, both the model and the patient data show a general rise in the proportion of intrusions with lesion severity. Domain differences are apparent in both the delayed copy and drawing conditions, with the model and the patients committing fewer intrusions for artifacts than for animals.

The patient results were analyzed using an ANOVA in which each drawing constitutes a separate observation, with proportion of

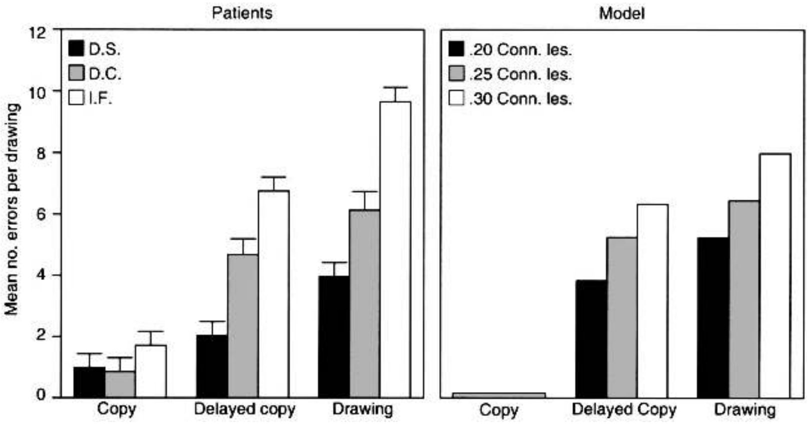

Figure 11. Average number of errors by task and severity for the patients (D.S., D.C., and I.F.; left) and the model (right). Error bars indicate standard error of the mean. Conn. les. = connections lesioned.

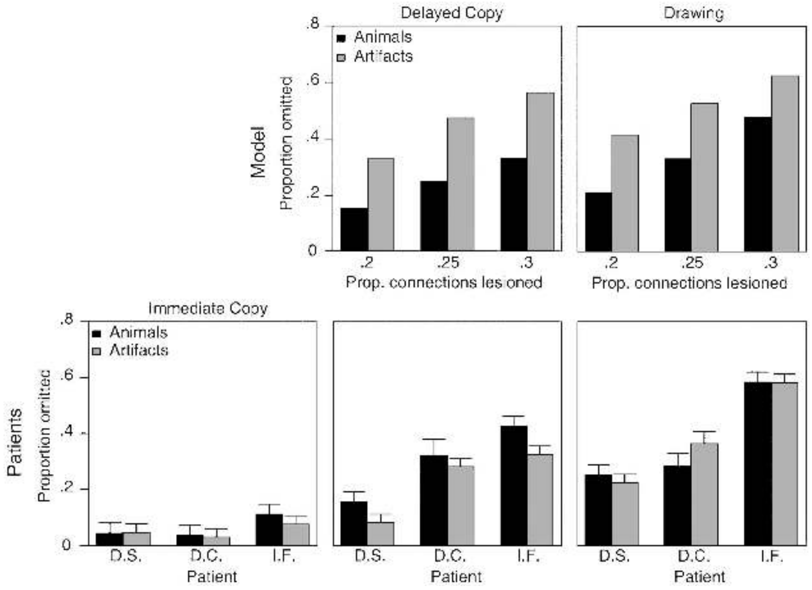

Figure 12. Proportion of features omitted per drawing in the model (top row) and for each patient (D.S., D.C., and I.F.; bottom row) plotted against lesion severity for animal and artifact domains in each task. Error bars indicate standard error of the mean. Prop. = proportion.

intrusions as the dependent measure and with patient, domain, and task as fixed, between-case factors. A strong main effect of task was apparent, $F(2, 415) = 55.0, p < .001$, with contrasts showing that both delayed copy and drawing-to-name yielded a greater proportion of intrusions than immediate copy ($p < .001$, in both cases), but with no significant difference between them (*ns*). There was also a reliable main effect of patient severity, $F(2, 415) = 31.0, p < .001$, with contrasts revealing that I.F. made more errors than D.C. ($p < .002$) and D.C. made more errors than D.S. ($p < .002$). A significant interaction between patient and task was observed, $F(4, 415) = 7.0, p < .001$, which is not surprising given that all patients were virtually at ceiling in the immediately copy condition. The effect of severity is evident only in the drawing and delayed copy condition. There was also a strong main effect of

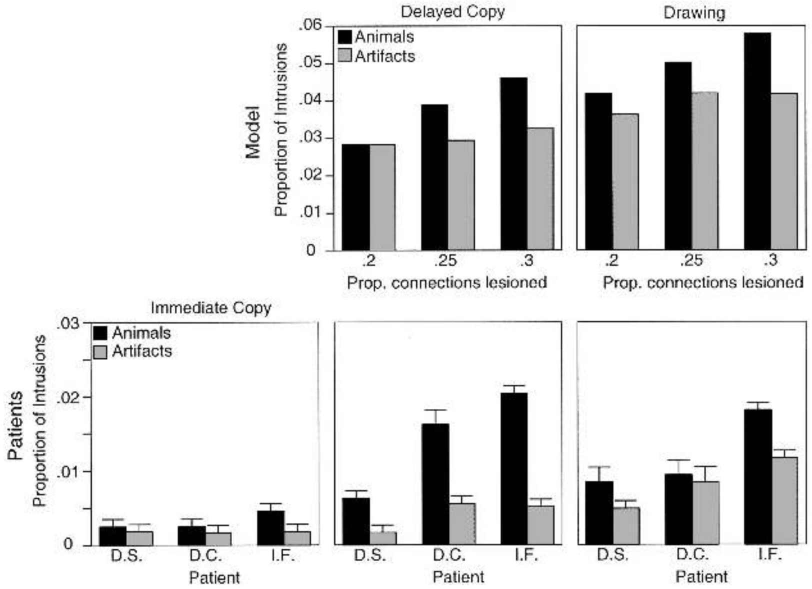

Figure 13. Proportion of intrusions made per drawing in the model (top row) and for each patient (D.S., D.C., and I.F.; bottom row) plotted against lesion severity for animal and artifact domains in each task. Error bars indicate standard error of the mean. Prop. = proportion.

domain, $F(1, 415) = 56.0$, $p < .001$, with patients making a higher proportion of intrusion errors for animals than for artifacts.

Error patterns by feature type. We have suggested that the semantic system has difficulty producing information about idiosyncratic or differentiating features under damage but that knowledge about properties shared across items with similar semantic representations is likely to be preserved. To address this claim in the context of drawing, we investigated the likelihood with which features were omitted or incorrectly activated in the model, given their propensity to be shared by other items in the same domain, shared by items in the same category, or idiosyncratic to the item.

On the basis of the visual feature norms described in the *Assessing Verbal and Visual Structure in the Environment* section, above, each visual attribute for every item was classified as follows: (a) as shared-across-domain if the property was true of more than half of the items in the same category and by more than half of the items in the contrasting category from the same domain (e.g., if the property was shared by more than half of the birds and more than half of the mammals), (b) as shared-by-category if the property was true of more than half of the items in the same category but not more than half of the items in the contrasting category from the same domain (e.g., if the property was shared by birds but not by mammals), and (c) as distinctive if the property was neither shared-across-domain nor shared-by-category.

In the model, each visual feature unit was classified the same way. This classification is item-specific—both in the model and in the world. A given feature might count as shared-by-category for some items and as idiosyncratic for others. For example, a property such as *wings* is shared across the category of birds, but it is idiosyncratic to particular mammals (such as the bat). Attributes that are shared-by-domain or shared-by-category are not necessarily true of every item in the corresponding domain or category. For example, an attribute such as *has legs* is generally true of most animals and, hence, would be considered shared-by-domain for individual animals. Nevertheless, there are some animals (such as the seal) that do not have legs. Such irregular properties are also to be found both in the model's training patterns and in the visual features that appear in the control drawings.

We calculated the proportion of the shared-by-domain, shared-by-category, and idiosyncratic target features omitted by the model and by each patient for every item. In the model, these figures were averaged over 50 damage trials at each level of lesion severity.

The means across all items tested for the patients and the model are shown in Figure 14. Here, each level of severity is plotted as a separate bar, and feature type is plotted along the abscissa. The most immediately apparent effect is that of feature type: In both the drawing and delayed copy tasks, both the model and the patients almost never omitted shared-across-domain properties and were much more likely to omit distinctive properties. As seen previously, more features were omitted in the drawing-to-name task than in the delayed copy task for all feature types at all levels of severity, by both the model and the patients.

The patient effects were again tested with a repeated-measures ANOVA in which, as previously, each picture was treated as a separate case. Feature type (shared-across-domain, shared-by-category, or distinctive) was treated as a within-case factor, and the proportion of features omitted for each type was the dependent measure. Patient, task, and domain were treated as fixed, independent between-case factors. The within-case main effect of feature type was reliable, $F(2, 432) = 203.0$, $p < .001$, with contrasts showing that participants omitted a smaller proportion of shared-by-domain features than shared-by-category features ($p < .001$)

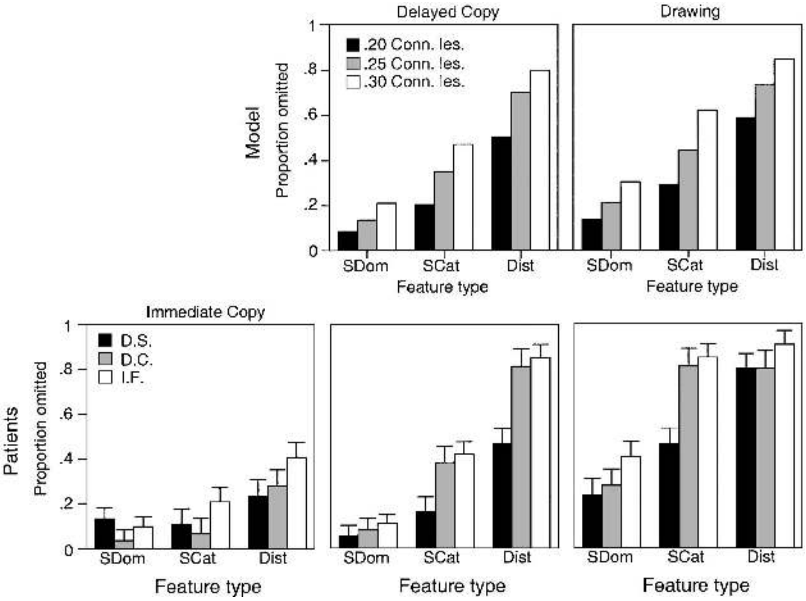

Figure 14. Proportion of shared-across-domain (SDom), shared-by-category (SCat), or distinctive (Dist) features omitted from drawings produced by each patient (D.S., D.C., and I.F.; bottom row) and by the model (top row) under increasingly severe simulated lesions. Error bars indicate standard error of the mean. Conn. les. = connections lesioned.

and a smaller proportion of shared-by-category than distinctive features ($p < .001$). Feature type also interacted reliably with patient, $F(4, 432) = 3.0, p < .02$; task, $F(4, 432) = 16.0, p < .001$; and domain, $F(2, 432) = 4.0, p < .04$. The first two of these interactions are easy to interpret from the plot of the factor means. The difference between shared-by-domain, shared-by-category, and distinctive features was greater for more severe patients, who omitted a larger proportion of features overall. Few features of any kind were omitted in the immediate copy condition. The interaction between feature type and domain is less straightforward to interpret. Patients appear to have omitted equivalent proportions of shared-by-domain properties (i.e., almost none) in both animal and artifact domains, but they have a slightly greater proportion of shared-by-category and distinctive properties for animals relative to artifacts. No third-order interaction was significant.

Tests of between-subjects factors showed reliable main effects of task, $F(2, 222) = 35.0, p < .001$, and patient, $F(2, 222) = 14.0$, $p < .001$, as expected. Task also interacted reliably with the other factors, which again reflects the patients' relatively good performance in the immediate copy condition. The effect of domain was not significant.

Finally, we assessed the likelihood of intrusions for the three different feature types. By definition, features that are shared-by-domain constitute target features for more than half of the items in a given domain, and hence, there are comparatively few shared-by-domain features that can intrude. In contrast, idiosyncratic features are shared by few items in a given domain; hence, for a given stimulus, there are many potential idiosyncratic features that could be wrongly included. For this reason, we normalized the data by calculating the total number of intruding features per item as a proportion of the total number of nontarget features for shared-by-domain, shared-by-category, and distinctive properties separately. Again, the model data were averaged across 50 damage trials at each level of severity.

The mean proportion of intrusions for each feature type in the model is shown in the top of Figure 15 and is averaged across all items and damage trials. Different degrees of severity are plotted as separate bars. For the model, shared-by-domain features are the most likely to be incorrectly added to a drawing, distinctive features are least likely to be incorrectly added, and shared-by-category features are somewhere in between, at all levels of severity.

The same effects are also apparent in the patient data on the bottom of Figure 15. The patient outcomes were assessed using a repeated-measures ANOVA with feature type as the within-case factor; proportion of intrusions as the dependent measure; and patient, task, and domain as fixed, independent between-case factors. Tests of the within-case factor revealed a strong main effect of feature type, $F(2, 66) = 28.0, p < .001$, with contrasts showing that shared-by-domain features were more often incorrectly added than shared-by-category features ($p < .001$) and that shared-by-category features were more often intruded than were distinctive features ($p < .001$). This effect did not interact reliably with any other factor except task, $F(4, 66) = 3.0, p < .05$.

It is interesting to note that domain did not yield a statistically reliable effect on intrusions, $F(1, 33) = 0.1$, *ns*. Our previous analysis of domain differences showed that the patients' drawings of animals contained a greater proportion of intrusions when feature type was not considered in the analysis. The current analysis shows that when feature type is included in the ANOVA model this effect is eliminated. In other words, the observed difference between domains in the first analysis was not due to semantic domain, per se. Rather, shared properties are incorrectly added to drawings in both animal and artifact domains. Animals

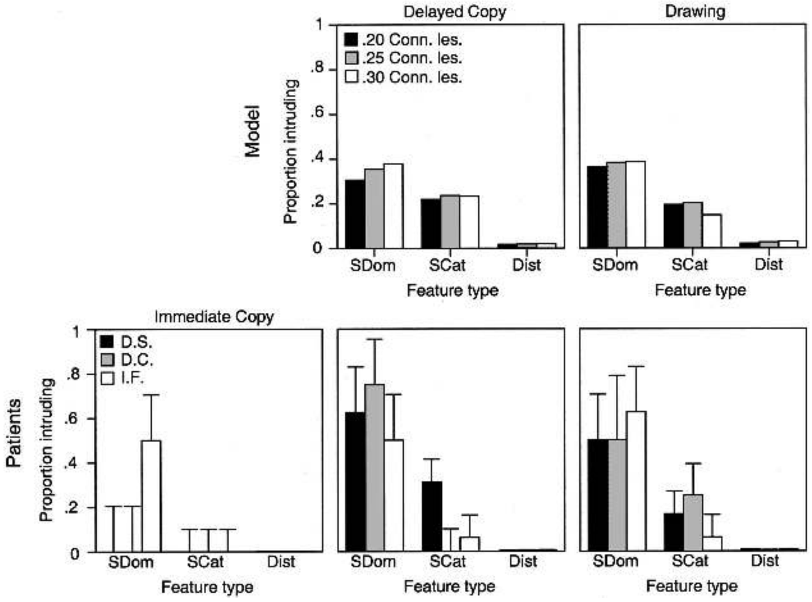

Figure 15. Proportion of intruded features in the model (top row) and for each patient (D.S., D.C., and I.F.; bottom row) for features that were shared-by-domain (SDom), shared-by-category (SCat), or distinctive (Dist). Error bars indicate standard error of the mean. Conn. les. = connections lesioned.

appear to invite more intrusions overall because there are more shared features per item in the *animal* domain, as shown in our analysis of control drawings in the *Assessing Verbal and Visual Structure in the Environment* section, above, and shared features are more likely to yield intrusions. No other between-case factors yielded reliable effects, except for task, $F(2, 33) = 5.0$, $p < .02$.

The data show that in both semantic domains, distinctive properties are likely to be omitted, and shared properties are likely to intrude in drawings of objects that do not participate in the regularities of the domain. Apparent domain differences in rates of omission and intrusion result from different tendencies for properties to be shared or distinctive in different semantic domains. In the model, these differences are sufficiently strong that we observed more intrusions for animals and more omissions for artifacts. In the patient data, the latter pattern was not observed, which possibly indicates that the proportion of distinctive visual features for the artifact items in the model was too high relative to animal items (in comparison to the true distribution in the environment). More important, the simulation demonstrates that the ultimate origin of intrusions and omissions in drawing is the same in the model and in the patients: Both derive from the propensity for a property to be shared among semantic neighbors.

Discussion

The architecture of the model has allowed us to simulate a range of semantic tasks of the sort commonly used to assess semantic memory, and the model's behavior under increasingly severe simulated lesions clearly provides a good qualitative match to the behavior of patients with semantic dementia. The model also makes several interesting predictions that have been borne out in the patient studies. Specifically, we observed (a) a greater proportion of omission errors for naming of artifacts and a greater proportion of commission errors for naming of animals, (b) better sorting at the specific relative to the general level for the category of fruits, (c) improved word-picture matching when distractor pictures are semantically distal to targets, (d) a greater likelihood of omitting idiosyncratic relative to shared visual properties of objects in drawing, and (e) a greater likelihood of incorrectly adding shared relative to idiosyncratic properties of objects in drawing. Thus, we suggest that the theory embodied by the model provides a useful way of thinking about semantic task performance and its disruption under general semantic impairment.

We have emphasized that the model's behavior under damage depends to a great extent on the structure of the semantic representations that mediate between visual and verbal representations and on the learned mappings between these representations and the expression and reception of visual and verbal information. Both of these factors depend ultimately on the structure of the visual and verbal training patterns provided to the model, which in this study incorporated aspects of structure apparent from visual and verbal attribute norms. The close match between model and patient performance across the spectrum of disease severity suggests that the representations and processes in the model may provide a good analog to those in the human semantic system and, hence, that human semantic representations may be acquired through learning about the similarity structure of the environment as experienced across different modalities, just as were the model's.

Finally, the model allows us to see that the patterns of impaired performance across naming, sorting, word–picture matching, and drawing tasks may all result from the same underlying factor, specifically, the dynamics of processing in a distributed and recurrent system as knowledge degrades. Verbal descriptors and visual features common to items that span a relatively broad and contiguous region of the model's representation space are more robust to semantic impairment, because the semantic system can generate accurate information about such properties whenever its representation state falls within this region. By contrast, distinctive properties of individual items are not shared by their neighbors; hence, as the model's representations degrade, the model grows increasingly unable to produce information about the distinguishing visual and verbal properties of individual objects. When this happens, recurrent interactions with representations in the periphery and within semantics itself can cause the system's representations to drift or become unstable. Small amounts of drift may lead the network into an inappropriate proximal attractor, from which it cannot produce information specific to a stimulus item, and the network may in fact produce incorrect responses appropriate to a semantically related object. However, properties that apply across a broad region of the representation space are robust even to relatively large amounts of damage, because the system's internal representations must be severely distorted before they drift out of the region to which such properties apply (Rogers & McClelland, in press).

This dynamic can account for all of the phenomena we have observed in the experiments described in this article, including (a) an inability to produce distinguishing information about objects, observed both in the increasing proportion of omission errors in naming, and in the omission of distinctive visual features in drawing; (b) a tendency to commit semantic errors in naming and intrusion errors in drawing, which both occur when the system's internal state is "captured" by an incorrect but proximal attractor; (c) robust preservation of information that differentiates broad semantic domains, observed in the increasing proportion of superordinate errors in naming, the relative preservation of sorting into general but not specific categories, and better accuracy in word-picture matching when distractors are semantically distal to targets; (d) a greater tendency to make errors of commission for animals relative to artifacts, observed in drawing and in naming, which derives from the structure of semantic representations that emerge in the model; and (e) worse performance on tasks that involve words as stimuli relative to those that involve pictures, which derives from the more systematic mapping between visual representations of objects and semantics.

A concise way of summarizing these observations is to submit that semantic dementia patients undergo an increasing overregularization of their conceptual knowledge. As damage accumulates and the system becomes increasingly unable to retrieve idiosyncratic and distinguishing information about objects, attractor dynamics cause the representations of less typical items to migrate toward the center of mass in the immediate neighborhood, which effectively renders them more typical. The system becomes unable to maintain distinctions between closely related items and misattributes the common properties of similar objects to related items that do not participate in the regularities of the domain. Initially, this collapse affects items within small, well-separated clusters (such as the category of birds). As damage mounts and the system is unable to maintain access to the properties that differentiate these clusters from neighboring clusters, these attractors collapse into even more general states, which reflect the central tendency of an even broader set of items. The result is that we witness a fine-to-coarse deterioration of concept knowledge, both in the system's ability to discriminate objects and in the set of properties attributed to individual items in different modes of expression.

Moreover, the magnitude of these effects depend on the density of the learned representations in different regions of the space. Domains with a high degree of structure, in which attractor states are packed into well-separated clusters, show a greater degree of overregularization—a greater likelihood of labeling items with general names, or with the names of familiar and typical neighbors, and a greater likelihood of incorrectly "adding in" common or typical properties to irregular items. Domains with a lesser degree of structure show a progressive loss of knowledge for individual properties, with a lesser tendency to overextend or misattribute names and properties to incorrect items.

In passing, we note that these factors—the density of the immediate neighborhood in the representation space and the degree to which neighbors consistently map to a common output—have also been useful for understanding overregularization errors by patients with semantic dementia in the pronunciation of written words (Patterson & Hodges, 1992; Plaut et al., 1996), in the formation of past-tense English verbs (Joanisse & Seidenberg, 1999; Patterson, Lambon Ralph, Hodges, & McClelland, 2001), and in lexical and object decision (Rogers et al., in press b; Rogers, Lambon Ralph, Hodges, & Patterson, in press a). Our account of the data from semantic tasks is closely related to connectionist accounts of the analogous phenomena in these domains. Indeed, we believe that a strength of the framework we have described is that it makes apparent the underlying similarity between patterns of impairment in conceptual knowledge and in these other domains of performance.

General Discussion

A century ago, Wernicke (1900, as cited in G. H. Eggert, 1977) put forward a theory of semantic memory that enabled him to make sense of the range of neuropsychological syndromes with which he was acquainted in his clinical practice. We have offered a parallel distributed processing implementation of this theory in the form of a neural network that acquires the ability to perform model analogs of semantic tasks through domain-general learning mechanisms. Under this theory, perceptual representations more-or-less directly encode modality-specific similarity structure in the environment. By virtue of learning the mappings between perceptual representations in different modalities, plus the further interaction of these with representations of words that refer to or describe such objects, the semantic system acquires abstract, distributed representations that encode the semantic similarity relations among different items. On this view, it is no coincidence that lesions to the anterior temporal cortex bilaterally result in the kind of general semantic impairment witnessed in semantic dementia. By virtue of their dense interconnections with association cortices in the more posterior part of the temporal lobes, these regions receive input from all sensory modalities (Gloor, 1997). It seems reasonable to suppose that they form the neural substrate within which amodal semantic representations emerge. These representations, in turn, subserve a key function of semantic memory in the intact system, namely, the generalization of stored information to novel items in the world and of newly acquired information to familiar items (see Rogers & McClelland, in press, for discussion). We have seen that when the inputs and outputs of our model capture aspects of the similarity structure of the environment, namely, similarities apparent in drawings of objects and in the words and phrases people use to describe these objects, the model provides an intuitive means of understanding patterns of impaired semantic task performance in semantic dementia.

Implications for Theories of Category-Specific Deficits

We have been silent on one contentious issue of import to theories of semantic memory, the extent to which the semantic system is organized by modality or semantic domain. Our reticence is partly due to the particular body of empirical data on which we have focused in this article. As noted in the introduction, semantic dementia provides the best evidence that there exists in the brain a single, amodal semantic store. Patients with semantic dementia are impaired on semantic tasks regardless of the modality of testing (Hodges et al., 1995; Bozeat et al., 2000; Hodges, Bozeat, Lambon Ralph, Patterson, & Spatt, 2000) and typically do not show preservation of knowledge for one domain relative to another (Lambon Ralph et al., 2001). It was our goal to understand how such global and amodal semantic deficits might arise as a consequence of the progressive deterioration of the anterior temporal cortex. Hence, we have not built into the model anything more than was necessary to explain the relevant phenomena.

However, other neuropsychological syndromes would seem to challenge the view of an amodal, homogeneous semantic store. Reports of patients with apparent category-specific deficits, modality-specific deficits, or Category × Modality interactions have led many researchers to suggest that semantic knowledge is mediated by an array of independent category- and modality-specific modules (e.g., Coltheart, Inglis, Michie, Bates, & Budd, 1998; Warrington & McCarthy, 1987). How might our theory be reconciled with these other cases? A comprehensive answer to this question is beyond the scope of this discussion. However, there are some aspects of the current work that have implications for the study of putative category-specific deficits and are worth noting.

First, we have identified three factors that (in addition to psycholinguistic factors such as familiarity and word frequency) may affect the likelihood that a given property will be retrieved in the context of a given semantic task:

1. The density of the semantic neighborhood, that is, the number of immediately proximal semantic representations. Errors of commission are more likely to occur in densely populated regions of the space.

2. The regularity of the property, that is, the degree to which the property is consistently shared among the item's semantic neighbors. If the property is not true of the test item, but is true of most of its neighbors, it is more likely to be incorrectly attributed to the test item.

3. The breadth of the semantic representation space spanned by the property. Properties that tend to be true of a broad set of semantically related items (like the shared properties of animals) are more robust to damage than properties that are true of a relatively narrow set of items (such as the properties shared by all canaries, but not other kinds of birds).

Animal and artifact domains likely differ on all three of these factors. We have seen that animals tend to share a greater number of properties with their semantic neighbors than do artifacts. In our model, artifact representations are more sparsely distributed across a broader region of the space. These factors lead to different patterns of errors in animal and artifact domains of the kind we have witnessed in the model and in the patient data. Hence, they must be added to the long list of potential confounding factors in experiments that purport to reveal true category-specific deficits.

Second, it is interesting to note that different semantic tasks revealed different aspects of structure in the patients' impaired performance. For example, the drawing tasks reported in this article reveal a somewhat richer structure to the pattern of impairment than has been heretofore elicited by such tasks as naming and word-to-picture matching. Specifically, we were able to identify from these data those bits of information that are lost to semantic dementia and also the ways in which regular properties are inappropriately added to items to which they do not belong. Similarly, our comparison of sorting with words or with pictures indicated that the overall level of performance can vary depending on the modality of testing. These observations suggest that the particular patterns observed in the data depend (perhaps to a greater degree than previously suspected) on the particular testing paradigm one adopts.

Third, our model implements a single, amodal and homogeneous system to mediate the interactions among perceptual representations in different modalities. In this sense, it is a unitary semantic system. However, the maintenance of stable semantic representations in our model depends to some extent on preserved connectivity between the semantic system and the perceptual representations with which it is connected. For example, if we were to lesion the connections between the semantic and visual layers on our model, we would not expect the model to perform perfectly even on purely verbal tasks such as naming to description. Because the entire model is interactive, disruptions in visuosemantic processing may have consequences for the system's ability to hold on to its semantic representations; as a result, the system may be impaired at semantic tasks that do not directly involve vision (Farah & McClelland, 1991; Humphreys & Forde, 2001).

Finally, because different perceptual modalities may capture different kinds of similarity relations among a group of items, we might expect different kinds of deficits to emerge in the system depending on which perceptual–semantic connections are disrupted. In our simple model, we have only implemented two perceptual modalities. However, we might suppose that the representations subserving our ability to act on objects capture a degree of richness or similarity structure among artifacts that are not mirrored in visual or verbal representations. Objects that afford similar actions may induce similar representations in areas of cortex that subserve action, and this structure may also constrain the similarity relations acquired by the semantic system as it learns the mappings between object appearances and appropriate actions. We might also assume that artifacts and living things differ in the amount of structure they share across the actions with which they are associated (Moss, Tyler, Durrant-Peatfield, & Bunn, 1998). Just as living things share a high degree of visual structure, whereas artifacts do not, artifacts may share a higher degree of structure across action representations than do living things (Plaut, 2002).

Lesions to the connections between semantics and either visual or action areas could result in different kinds of category-specific semantic deficits (a view that is similar in some respects to that described in Warrington & Shallice, 1984). Damage to the connections between semantics and action representations may lead the system to confuse artifacts, because such objects share structure in the action modality. By contrast, damage to the connections between vision and semantics may lead the network to confuse various animals with one another, because of the high degree of visual structure that is apparent in that domain. Deficits particularly affecting language may manifest when the links between verbal and semantic areas are disrupted, and generalized semantics deficits of the kind we have described in this article may arise from damage to the semantic units themselves (Rogers & Plaut, 2002). Of course, it remains to be determined whether such an account can explain the range of data reported in the literature; this is a course we will pursue in future work.

Relationship to Other Theories

Our network maps between surface forms by using distributed semantic representations with the following characteristics:

1. The conceptual representations are acquired by the network during learning and are not assigned by the computational modeler.

2. The learned representations are not feature based but are instantiated as points in a high-dimensional space.

3. There is no functional specialization of the units (e.g., perceptual vs. functional representations) within the semantic layer.

4. Modality-specific surface representations provide input to and encode output from semantics.

Other researchers have described connectionist models that adopt some but not all of these properties. Neuropsychological models based on Farah and McClelland's (1991) influential framework incorporate abstract semantic representations that map between visual and verbal representations but use hard-wired representations specified by the experimenter and assume functional specialization within semantics (e.g., Devlin et al., 1998; Lambon Ralph et al., 2001; Lambon Ralph & Howard, 2000). The interactive activation model of visual object recognition described by Humphreys and colleagues (e.g., Humphreys & Forde, 2001; Humphreys, Lamote, & Lloyd-Jones, 1995) uses a similar three-layer architecture, with visual representations of objects engaging semantic representations that in turn activate lexico-phonological representations of words. In this case, the model uses prespecified localist representations at each level. We believe that this body of work has established the utility of implementing theories of semantic representation in an explicit computational framework but that the use of prespecified semantic representations raises important questions about knowledge acquisition. The specification of representations by fiat allows the investigator to explore how different choices of representation influence the behavior of the model, but it also compromises any appeal to external validity. The theorist may demonstrate that, with a certain choice of representation, the model provides a good match to the data; without an account of where the useful structure came from in the first place, the choice of representation is constrained only by the data to be explained, and the theory has in some sense assumed what it is trying to explain. The capacity of connectionist networks to acquire abstract, distributed concept representations has been explored by several researchers (Elman, 1990; Hinton, 1986; Miikkulainen & Dyer, 1991; Rumelhart & Todd, 1993; Schyns, 1991), but this work has tended to be somewhat too abstract to provide a basis for understanding empirical data from neuropsychological studies.

Alternative approaches have attempted to address this issue by using models that learn the mappings among vectors of semantic attributes, which are derived from attribute norms as in our study. For example, Tyler et al. (2000) have described an autoassociator network that, like our model, derives a semantic space across an intermediate hidden layer and uses training patterns that incorporate aspects of structure apparent in verbal attribute norms. Tyler et al. have used the model to make predictions about patterns of impairment across different semantic domains in disturbed semantic cognition. However, in this case, the inputs and outputs incorporated in the model are construed as vectors of semantic features, and there is no distinction made between the information provided to semantics through vision from that provided through language or through other modes of perception. The same is true of the network described by McRae et al. (1997), which learns mappings among a large set of semantic feature vectors derived from an impressive corpus of verbal attribute-listing norms. McRae et al. used their network to simulate the pattern of semantic priming found in normal participants for varying types of semantic attributes. Both cases demonstrate that useful information about conceptual structure can be gleaned by considering the similarities yielded by feature-norming studies, and both bolster the argument that representational structure in semantics may derive from the attribute structure of the environment. However, the ultimate promise of this idea remains untested in both models, because the attributes from which semantic knowledge is comprised are divorced from the perceptual representations and processes that mediate our experience of the environment. That is, the attributes themselves are construed as constituents of semantic representations, but neither model suggests how these constituents might be derived from visual appearances, verbal statements, and other information available from the environment through perception. Moreover, because such models do not implement perceptual inputs or outputs, they raise questions about many of the tasks we have described in this article, which seem to require the activation of surface representations (e.g., naming, sorting, drawing, etc.).

We view our model as a useful synthesis of these different approaches. Like semantic feature based theories, our theory suggests that representational structure in semantics depends on the perceived structure of the environment and provides a means of assessing the external validity of any particular assumption about the nature of this structure. However, in concert with models that assume more abstract semantic representations, the representations that emerge in our model do not code explicit semantic content. Instead, they are structured in ways that facilitate the system's ability to generate appropriate responses when given perceptual inputs. This approach is most similar to recent work described by Plaut to explain patterns of category- and modality-specific semantic deficits (e.g., Plaut, 2002) and to the approach laid out by Rogers and McClelland (in press) in their general theory of semantic cognition. In neither of these cases, however, were model-training patterns derived from normative data such as the propositional norms and drawing features used in the present study.

The use of high-dimensional spaces to capture semantic representations is not specific to computational models. Statistical analyses such as principal component analysis and multidimensional scaling, used in this study and elsewhere (e.g., Garrard et al., 2001; Medin, Lynch, & Coley, 1997), are two better known examples. Other techniques such as latent semantic analysis (LSA; Landauer & Dumais, 1997) and hyperspace analogue to language (HAL; Burgess & Lund, 1997) are able to extract semantic representations relatively efficiently from large corpora of text (e.g., encyclopedias). Like our network, these techniques extract high-order co-occurrence statistics across stimulus events in the environment. However, to our knowledge, there has been no attempt to link these processes to neuroanatomical or computational factors in the brain. Also, LSA and HAL are entirely reliant on verbal input. This has two implications. First, like a number of the computational models noted previously, it is impossible to simulate directly the behavior of normal participants or neurologically impaired patients because there is no implemented link between the high-dimensional semantic representations and surface forms. Indeed, if one is to accept the (rather abstract) notion that concepts are points in a high-dimensional semantic space that cannot be probed directly, then it seems imperative to understand both the semantic space and its connection with receptive and expressive domains, which can be studied directly. Second, our interactions with the world are obviously not limited to the verbal modality; however, in these theories there is no influence of nonverbal experience on the semantic representations. Like the proponents of LSA and HAL, we assume that verbal experience is one contributor to our conceptual knowledge. We have, however, demonstrated that samples of the verbal domain—in this case, a feature-listing database—tend to obscure information that is readily available in other modalities (such as the visual similarities existing between fruits and small, manipulable artifacts); furthermore, we believe that preverbal learning contributes a great deal to early knowledge acquisition (e.g., Mandler, 2000; Mareschal, 2000). It is our working hypothesis that all perceptual modalities contribute to our conceptual knowledge, although the contribution of each may vary considerably.

From a neuropsychological point of view, we are not the first to suggest that semantic memory is supported by a unitary, amodal system. The organized unitary content hypothesis (OUCH; Caramazza, Hillis, Rapp, & Romani, 1990) assumes that concepts are represented in some form of space such that similar concepts are close neighbors. The same conceptual representations are accessed from different input modalities (e.g., for comprehending spoken words, pictures, objects, etc.) and drive expressive abilities such as speaking and writing. That is, semantic representations are assumed to be amodal. OUCH also assumes nonequivalence of the relationship or mapping between various surface representations and amodal semantic representations. In this view, the picture–object to semantic mapping benefits from a quasi-systematic mapping (encapsulated by the two assumptions termed *the assumption of privileged access* and *the assumption of privileged relationships*; Caramazza et al., 1990). By encapsulating these and other ideas into an implemented computational model, we are able to be much more explicit about these issues. We can explain how the representations are acquired and how they are engaged in particular tasks; we can investigate the nature of the otherwise hidden semantic system and study its relationship with surface representations; and we can be explicit about the behavioral consequences of damage to the semantic system.

Finally, we opened this article by reviewing a classical neurological view of semantic memory endorsed by Wernicke (1900, as cited in G. H. Eggert, 1977) and other neurologists at the end of the 19th century, which is captured in a formal way by the computational model. We also noted that proposals similar to this have reappeared in more contemporary accounts (e.g., Allport, 1985). Perhaps the best known example is the work of Damasio and colleagues (H. Damasio, Grabowski, Tranel, & Hichwa, 1996; Tranel, Damasio, & Damasio, 1997), who have suggested that areas in the temporal cortex act as "convergence zones" for information projecting to and from the sensory association areas. This theory capitalizes on one of the insights described so elegantly in Wernicke's writings: Semantic knowledge may be construed as a process that mediates the interactions among content-bearing perceptual representations, rather than as a repository of propositional facts about objects. This idea is echoed in the work of many contemporary researchers (Chao, Haxby, & Martin, 1999; Kellenbach et al., 2001; Mummery, Patterson, Hodges, & Price, 1998; Pulvermueller, 1999). However, Damasio's convergence zones are not assumed to encode semantic representations themselves, but they act as a kind of relay-station through which information in different sensory-motor domains can be linked. It is the latter aspect that changes in our account. Our assumption is that the anterior regions of the temporal lobes, like the hidden layer in the computational model, actually derive amodal semantic representations that encode the semantic similarity relations among objects regardless of their surface similarities (see also McClelland & Rogers, 2003). As we have seen, the extraction of similarity structure across multiple modalities can lead to the emergence of structure that is not apparent in any modality individually. We take this re-representational capacity to be one of the fundamental functions of the semantic system.

References

Allport, D. A. (1985). Distributed memory, modular systems and dysphasia. In S. K. Newman & R. Epstein (Eds.), *Current perspectives in dysphasia* (pp. 207–244). Edinburgh, Scotland: Churchill Livingstone.

Bozeat, S., Lambon Ralph, M. A., Graham, K. S., Patterson, K., Wilkin, H., Rowland, J., et al. (2003). A duck with four legs: Investigating the structure of conceptual knowledge using picture drawing in semantic dementia. *Cognitive Neuropsychology, 20,* 27–47.

Bozeat, S., Lambon Ralph, M. A., Patterson, K., Garrard, P., & Hodges, J. R. (2000). Nonverbal semantic impairment in semantic dementia. *Neuropsychologia, 38,* 1207–1215.

Brown, R. (1958). How shall a thing be called? *Psychological Review, 65,* 14–21.

Burgess, C., & Lund, K. (1997). Modelling parsing constraints with high-dimensional context space. *Language and Cognitive Processes, 12,* 177–210.

Caramazza, A., Hillis, A. E., Rapp, B. C., & Romani, C. (1990). The multiple semantics hypothesis: Multiple confusions? *Cognitive Neuropsychology, 7,* 161–189.

Chao, L. L., Haxby, J. V., & Martin, A. (1999). Attribute-based neural substrates in temporal cortex for perceiving and knowing about objects. *Nature Neuroscience, 2,* 913–919.

Coltheart, M., Inglis, L., Michie, P., Bates, A., & Budd, B. (1998). A semantic subsystem of visual attributes. *Neurocase, 4,* 353–370.

Cree, G., McRae, K., & McNorgan, C. (1999). An attractor model of lexical conceptual processing: Simulating semantic priming. *Cognitive Science, 23,* 371–414.

Damasio, A. R. (1989). The brain binds entities and events by multiregional activation from convergence zones. *Neural Computation, 1,* 123–132.

Damasio, H., Grabowski, T. J., Tranel, D., & Hichwa, R. D. (1996, April 11). A neural basis for lexical retrieval. *Nature, 380,* 499–505.

Devlin, J. T., Gonnerman, L. M., Andersen, E. S., & Seidenberg, M. S. (1998). Category-specific semantic deficits in focal and widespread brain damage: A computational account. *Journal of Cognitive Neuroscience, 10,* 77–94.

Done, D. J., & Gale, T. M. (1997). Attribute verification in dementia of Alzheimer type: Evidence for the preservation of distributed concept knowledge. *Cognitive Neuropsychology, 14,* 547–571.

Eggert, G. H. (1977). *Wernicke's works on aphasia: A sourcebook and review* (Vol. 1). The Hague, the Netherlands: Mouton.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science, 14,* 179–211.

Everitt, B. (1974). *Cluster analysis.* London: Heinemann Educational Books.

Farah, M. J., & McClelland, J. L. (1991). A computational model of semantic memory impairment: Modality-specificity and emergent category-specificity. *Journal of Experimental Psychology: General, 120,* 339–357.

Funnell, E. (1996). Response biases in oral reading: An account of the co-occurrence of surface dyslexia and semantic dementia. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology, 49*(A), 417–446.

Gainotti, G., Silveri, M. C., Daniele, A., & Giustolisi, L. (1995). Neuroanatomical correlates of category-specific semantic disorders: A critical survey. *Memory, 3,* 247–265.

Garrard, P., Lambon Ralph, M. A., Hodges, J. R., & Patterson, K. (2001). Prototypicality, distinctiveness and intercorrelation: Analyses of the semantic attributes of living and nonliving concepts. *Journal of Cognitive Neuroscience, 18,* 125–174.

Gloor, P. (1997). *The temporal lobe and limbic system.* New York: Oxford University Press.

Graham, K. S., Lambon Ralph, M. A., & Hodges, J. R. (1997). Determining the impact of autobiographical experience on "meaning": New insights from investigating sports-related vocabulary and knowledge in two cases with semantic dementia. *Cognitive Neuropsychology, 14,* 801–837.

Grey, H., & Bannister, L. H. (1995). *Grey's anatomy* (38th ed.). New York: Churchill Livingstone.

Hinton, G. E. (1986). Learning distributed representations of concepts. In *Proceedings of the 8th Annual Conference of the Cognitive Science Society* (pp. 1–12). Hillsdale, NJ: Erlbaum. (Reprinted in *Parallel Distributed Processing: Implications for Psychology and Neurobiology,* pp. 9–45, by R. G. Morris, Ed., 1989, Oxford, England: Oxford University Press)

Hodges, J. R., Bozeat, S., Lambon Ralph, M. A., Patterson, K., & Spatt, J. (2000). The role of conceptual knowledge in object use: Evidence from semantic dementia. *Brain, 123,* 1913–1925.

Hodges, J. R., Graham, N., & Patterson, K. (1995). Charting the progression in semantic dementia: Implications for the organisation of semantic memory. *Memory, 3,* 463–495.

Hodges, J. R., Patterson, K., Oxbury, S., & Funnell, E. (1992). Semantic dementia: Progressive fluent aphasia with temporal lobe atrophy. *Brain, 115,* 1783–1806.

Hodges, J. R., Salmon, D. P., & Butters, N. (1992). Semantic memory impairment in Alzheimer's disease: Failure of access or degraded knowledge? *Neuropsychologia, 114,* 1547–1558.

Humphreys, G. W., & Forde, E. M. (2001). Hierarchies, similarity, and interactivity in object-recognition: On the multiplicity of "category-specific" deficits in neuropsychological populations. *Behavioral and Brain Sciences, 24,* 453–509.

Humphreys, G. W., Lamote, C., & Lloyd-Jones, T. J. (1995). An interactive activation approach to object processing: Effects of structural similarity, name frequency, and task in normality and pathology. *Memory, 3,* 535–586.

Joanisse, M. F., & Seidenberg, M. S. (1999). Impairments in verb morphology after brain injury: A connectionist model. *Proceedings of the National Academy of Sciences, USA, 96,* 7592–7597.

Jolicoeur, P., Gluck, M., & Kosslyn, S. (1984). Pictures and names: Making the connection. *Cognitive Psychology, 19,* 31–53.

Kellenbach, M., Brett, M., & Patterson, K. (2001). Large, colorful or noisy? Attribute- and modality-specific activations during retrieval of perceptual attribute knowledge. *Cognitive, Affective and Behavioral Neuroscience, 1,* 207–221.

Lambon Ralph, M. A., Ellis, A. W., & Franklin, S. (1995). Semantic loss without surface dyslexia. *Neurocase, 1,* 363–369.

Lambon Ralph, M., Graham, K., Ellis, A., & Hodges, J. R. (1998). Naming in semantic dementia—what matters? *Neuropsychologia, 36,* 125–142.

Lambon Ralph, M., Graham, K., Patterson, K., & Hodges, J. R. (1999). Is a picture worth a thousand words? Evidence from concept definitions by patients with semantic dementia. *Brain and Language, 70,* 309–335.

Lambon Ralph, M. A., & Howard, D. (2000). Gogi-aphasia or semantic dementia? Simulating and assessing poor verbal comprehension in a case of progressive fluent aphasia. *Cognitive Neuropsychology, 17,* 437–465.

Lambon Ralph, M. A., Howard, D., Nightingale, G., & Ellis, A. W. (1998). Are living and non-living category-specific deficits causally linked to impaired perceptual or associative knowledge? Evidence from a category-specific double dissociation. *Neurocase, 4,* 311–338.

Lambon Ralph, M. A., McClelland, J., Patterson, K., Galton, C. J., & Hodges, J. (2001). No right to speak? The relationship between object naming and semantic impairment: Neuropsychological evidence and a computational model. *Journal of Cognitive Neuroscience, 13,* 341–356.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104,* 211–240.

Mandler, J. M. (2000). Perceptual and conceptual processes in infancy. *Journal of Cognition and Development, 1,* 3–36.

Mareschal, D. (2000). Infant object knowledge: Current trends and controversies. *Trends in Cognitive Science, 4,* 408–416.

Martin, A., & Chao, L. L. (2001). Semantic memory in the brain: Structure and processes. *Current Opinion in Neurobiology, 11,* 194–201.

McClelland, J. L., & Rogers, T. T. (2003). The parallel distributed processing approach to semantic cognition. *Nature Reviews Neuroscience, 4,* 310–322.

McClelland, J. L., Rumelhart, D. E., & the PDP Research Group. (1986). *Parallel distributed processing: Explorations in the microstructure of cognition: Psychological and biological models* (Vol. 2). Cambridge, MA: MIT Press.

McGuire, S., & Plaut, D. C. (1997). Systematicity and specialization in semantics: A computational account of optic aphasia. In *Proceedings of the 19th Annual Conference of the Cognitive Science Society* (pp. 502–507). Hillsdale, NJ: Erlbaum.

McRae, K., & Cree, G. (2002). Factors underlying category-specific deficits. In E. M. E. Forde & G. Humphreys (Eds.), *Category specificity in brain and mind* (pp. 211–249). Hove, England: Psychology Press.

McRae, K., De Sa, V., & Seidenberg, M. (1997). On the nature and scope of featural representations of word meaning. *Journal of Experimental Psychology: General, 126,* 99–130.

Medin, D. L., Lynch, E. B., & Coley, J. D. (1997). Categorization and reasoning among tree experts: Do all roads lead to Rome? *Cognitive Psychology, 32,* 49–96.

Miikkulainen, R., & Dyer, M. G. (1991). Natural language processing with modular PDP networks and distributed lexicon. *Cognitive Science, 15,* 343–399.

Moss, H. E., Tyler, L. K., & Devlin, J. T. (2002). The emergence of category-specific deficits in a distributed semantic system. In E. M. E. Forde & G. W. Humphreys (Eds.), *Category specificity in brain and mind* (pp. 115–148). Hove, England: Psychology Press.

Moss, H. E., Tyler, L. K., Durrant-Peatfield, M., & Bunn, E. M. (1998). Two eyes of a see-through: Impaired and intact semantic knowledge in a case of selective deficit for living things. *Neurocase, 4,* 291–310.

Mummery, C. J., Patterson, K., Hodges, J. R., & Price, C. J. (1998). Functional neuroanatomy of the semantic system: Divisible by what? *Journal of Cognitive Neuroscience, 10,* 766–777.

Mummery, C. J., Patterson, K., Price, C. J., Ashburner, J., Frackowiak, R. S. J., & Hodges, J. (2000). A voxel-based morphometry study of semantic dementia: Relationship between temporal lobe atrophy and semantic memory. *Annals of Neurology, 47,* 36–45.

Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review, 92,* 289–316.

Patterson, K., & Hodges, J. R. (1992). Deterioration of word meaning: Implications for reading. *Neuropsychologia, 30,* 1025–1040.

Patterson, K., & Hodges, J. (2000). Semantic dementia: One window on the structure and organisation of semantic memory. In F. Boller & J. Grafman (Eds.), *Handbook of neuropsychology: Memory and its disorders* (Vol. 2, 2nd ed., pp. 313–333). Amsterdam: Elsevier Science.

Patterson, K., Lambon Ralph, M., Hodges, J. R., & McClelland, J. L. (2001). Deficits in irregular past-tense verb morphology associated with degraded semantic knowledge. *Neuropsychologia, 39,* 709–724.

Plaut, D. C. (1995). Double dissociation without modularity: Evidence from connectionist neuropsychology. *Journal of Clinical and Experimental Neuropsychology, 17,* 291–321.

Plaut, D. C. (2002). Graded modality-specific specialization in semantics: A computational account of optic aphasia. *Cognitive Neuropsychology, 19,* 603–639.

Plaut, D. C., McClelland, J. L., Seidenberg, M. S., & Patterson, K. (1996). Understanding normal and impaired word reading: Computational principles in quasi-regular domains. *Psychological Review, 103,* 56–115.

Pulvermueller, F. (1999). Words in the brain's language. *Behavioural and Brain Sciences, 22,* 253–336.

Rogers, T. T., Lambon Ralph, M. A., Hodges, J. R., & Patterson, K. (in press a). Natural selection: The impact of semantic impairment on lexical and object decision. *Cognitive Neuropsychology.*

Rogers, T. T., Lambon Ralph, M. A., Hodges, J. R., & Patterson, K. (in press b). Object recognition under semantic impairment: The effects of conceptual regularities on perceptual decisions. *Language and Cognitive Processes.*

Rogers, T. T., & McClelland, J. L. (in press). *Semantic cognition: A parallel distributed processing approach.* Boston: MIT Press.

Rogers, T. T., & Plaut, D. C. (2002). Connectionist perspectives on category specific deficits. In E. Forde & G. Humphreys (Eds.), *Category specificity in brain and mind* (pp. 251). East Sussex, England: Psychology Press.

Rosch, E., Mervis, C. B., Gray, W., Johnson, D., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology, 8,* 382–439.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning internal representations by error propagation. In D. E. Rumelhart, J. L. McClelland, & the PDP Research Group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition: Foundations* (Vol. 1, pp. 318–362). Cambridge, MA: MIT Press.

Rumelhart, D. E., McClelland, J. L., & the PDP Research Group. (1986). *Parallel distributed processing: Explorations in the microstructure of cognition: Foundations* (Vol. 1). Cambridge, MA: MIT Press.

Rumelhart, D. E., & Todd, P. M. (1993). Learning and connectionist representations. In D. E. Meyer & S. Kornblum (Eds.), *Attention and Performance XIV: Synergies in experimental psychology, artificial intelligence, and cognitive neuroscience* (pp. 3–30). Cambridge, MA: MIT Press.

Schwartz, M. F., Marin, O. S. M., & Saffran, E. M. (1979). Dissociations of language function in dementia: A case study. *Brain and Language, 7,* 277–306.

Schyns, P. G. (1991). A modular neural network model of concept acquisition. *Cognitive Science, 15,* 461–508.

Sloman, S. A., & Rips, L. J. (1998). Similarity as an explanatory construct. *Cognition, 65,* 87–101.

Snowden, J. S., Goulding, P. J., & Neary, D. (1989). Semantic dementia: a form of circumscribed temporal atrophy. *Behavioural Neurology, 2,* 167–182.

Snowden, J. S., Neary, D., & Mann, D. M. A. (1996). *Frontotemporal lobar degeneration: Frontotemporal dementia, progressive aphasia, semantic dementia.* New York: Churchill Livingstone.

Tranel, D., Damasio, H., & Damasio, A. (1997). A neural basis for the retrieval of conceptual knowledge. *Neuropsychologia, 35,* 1319–1327.

Tyler, L., Moss, H. E., Durrant-Peatfield, M. R., & Levy, J. P. (2000). Conceptual structure and the structure of concepts: A distributed account of category-specific deficits. *Brain and Language, 75,* 195–231.

Warrington, E. K. (1975). Selective impairment of semantic memory. *Quarterly Journal of Experimental Psychology, 27,* 635–657.

Warrington, E. K., & McCarthy, R. (1987). Categories of knowledge: Further fractionation and an attempted integration. *Brain, 110,* 1273–1296.

Warrington, E. K., & Shallice, T. (1984). Category specific semantic impairments. *Brain, 107,* 829–853.

Received March 21, 2001
Revision received March 20, 2002
Accepted June 23, 2003 ■