Category-Specific Semantic Deficits in Focal and Widespread Brain Damage: A Computational Account

Joseph T. Devlin, Laura M. Gonnerman, Elaine S. Andersen, and Mark S. Seidenberg
University of Southern California

Abstract

■ Category-specific semantic impairments have been explained in terms of preferential damage to different types of features (e.g., perceptual vs. functional). This account is compatible with cases in which the impairments were the result of relatively focal lesions, as in herpes encephalitis. Recently, however, there have been reports of category-specific impairments associated with Alzheimer's disease, in which there is more widespread, patchy damage. We present experiments with a connectionist model that show how "category-specific" impairments can arise in cases of both localized and widespread damage; in this model, types of features are topographically organized, but specific categories are not. These effects mainly depend on differences between categories in the distribution of correlated features. The model's predictions about degree of impairment on natural kinds and artifacts over the course of semantic deterioration are shown to be consistent with existing patient data. The model shows how the probabilistic nature of damage in Alzheimer's disease interacts with the structure of semantic memory to yield different patterns of impairment between patients and categories over time. ■

INTRODUCTION

Semantic memory impairments can be caused by several types of neuropathology, including Alzheimer's disease (e.g., Nebes, 1989), herpes simplex encephalitis (e.g., Warrington & Shallice, 1984), and cerebrovascular accidents (e.g., Hart & Gordon, 1990). A goal of research on such impairments is to understand them in terms of damage to the normal semantic system. Connectionist modeling has provided a useful tool in pursuing this goal. This approach involves implementing semantic networks using the units, weights, learning algorithms, and other components of the connectionist approach (Rumelhart & McClelland, 1986). Semantic memory impairments can then be simulated by introducing anomalies in the computational model (e.g., Hinton & Shallice, 1991; Plaut & Shallice, 1993). Such simulations can provide important clues concerning the organization of semantic memory and the bases of semantic impairments. Although these models have as yet incorporated only very general neurobiological constraints, they represent a step along the path toward understanding how semantic information is represented and processed in the brain.

An example of this approach is provided by recent work on category-specific semantic impairments. Warrington and her colleagues (Warrington & McCarthy, 1983, 1987; Warrington & Shallice, 1984) observed a striking double dissociation in the residual semantic abilities of six subjects after focal brain injury. Two were impaired in their ability to recognize pictures of artifacts compared to pictures of food and animals. Four others displayed the opposite pattern: They were impaired at naming pictures of animals and food compared to pictures of artifacts. The authors interpreted this double dissociation as indicating that artifacts and natural kinds could be preferentially affected by brain damage. Certain categories proved problematical for this division, however. One patient, JBR, suffered from a natural kinds deficit but was also impaired on musical instruments, gemstones, and fabrics. He also had preserved knowledge of body parts, which might be considered a natural kinds category. Patient YOT, on the other hand, had an artifacts deficit that included body parts but not musical instruments, gemstones, or fabrics. This pattern led Warrington and McCarthy to recast the distinction between artifacts and natural kinds as one between functional and perceptual features as follows (Warrington & McCarthy, 1987, p. 1275):

> We have . . . argued for a refinement of the animate/inanimate dichotomy and suggested that a distinction cast in terms of the relevant salience of processing functional/physical (also termed sensory) attributes might account for these dissociations. Thus similar items from the broad category of

> objects are primarily comprehended in terms of their core functional significance, physical attributes being of less importance for their differentiation. . . . Food and living things by contrast are differentiated primarily in terms of their physical attributes.

This pattern of results has now been observed several times (see Saffran & Schwartz, 1994, for a review). Artifact categories such as musical instruments pattern with natural kinds categories such as animals because musical instruments, which have similar functions, are largely distinguished from one another on the basis of perceptual properties. Conversely, body parts are natural entities that may pattern with artifacts because their functions are highly relevant to distinguishing among them. Category-specific impairments can then be seen as secondary to selective damage to different types of features.[1]

Farah and McClelland (1991) developed a connectionist model based on these insights. They derived an empirical estimate of the relative frequency of the two types of features, which yielded a 3:1 ratio of visual to functional features overall. More significantly, this ratio varied from 7.7:1 for natural kinds to 1.4:1 for artifacts. Thus, both types of features are relevant to both domains, but for the natural kinds, perceptual features are much more important than functional ones, whereas for the artifacts they are more equally important. Farah and McClelland incorporated these observations about the proportions of featural types in a connectionist model that took a representation of either a picture or a word as input and generated its semantic representation as output. The trained model was then damaged in differing degrees by eliminating either visual or functional semantic features. Damage to visual features yielded a selective impairment for natural kinds, while removing functional features resulted in a deficit specific to artifacts. In both cases the severity of the deficit was correlated with an increase in semantic damage. Thus, the model produced category-specific deficits even though its semantic representation was not organized by category. "Category specificity" was seen as an emergent property of the semantic system arising from the underlying distribution of perceptual and functional semantic features.

In addition to reproducing the basic patterns of impaired behavior, the Farah and McClelland (1991) model explained the additional finding that several subjects with natural kinds deficits also exhibit impaired functional knowledge on the same items (Basso, Capitani, & Laiacona, 1988; Sartori & Job; 1988; Silveri & Gainotti, 1988). For example, Silveri and Gainotti described a patient with a severe natural kinds deficit who had difficulty naming animals from visual descriptions (e.g., "a black-and-white striped wild horse" for zebra; 11% correct) and from nonvisual descriptions (e.g., "the farm animal which bleats and supplies us with wool" for sheep; 58% correct). Farah and McClelland argued that the deficit in functional knowledge was a secondary result of damage to perceptual features. In their view, the conjunction of visual and functional features that forms an object's semantic representation constitutes a single stable pattern such that without a "critical mass" of activated features the integrity of the entire representation suffers. Their model demonstrated that damage to perceptual features had secondary effects on the activation of functional features because of the connections between them. For instance, having removed 40% of the model's visual features, they found a degraded pattern of activation over the functional features for natural kinds.

Category-Specific Impairments in Alzheimer's Disease

The work of Warrington and colleagues (Warrington & McCarthy, 1983, 1987; Warrington & Shallice, 1984) and of Farah and McClelland (1991) speaks directly to the structure of the normal semantic system, suggesting that information must be topographically organized such that focal brain damage can preferentially affect either perceptual or functional features. Given that we accept a topographically organized system, it not obvious how nonfocal neuropathologies, such as the widespread, patchy damage found in Alzheimer's disease (AD), could produce the same kinds of semantic impairments. It is therefore of considerable theoretical interest that there have been two recent reports of category-specific impairments associated with Alzheimer's disease (Mazzoni, Moretti, Lucchini, Vista, & Muratorio, 1991; Silveri, Daniele, Giustolisi, & Gainotti, 1991).

Both studies examined the behavior of subjects with probable AD and found selective deficits in natural kinds compared to artifacts. The authors noted that the AD subjects were impaired in the same domain, natural kinds, as were herpes simplex encephalitis (HSE) patients who have been reported to have category-specific impairments (e.g., Pietrini, Nertempi, Vaglia, Revello, Pinna, & Ferro-Milone, 1988), despite significant differences in the etiologies. Both Silveri et al. (1991) and Mazzoni et al. (1991) drew inferences from these deficit patterns about the localization of perceptual and functional information in the brain. Perceptual information was said to be localized in the temporo-limbic areas of the brain, and functional information in the fronto-parietal regions, a hypothesis that is receiving increasing support from both neuropsychological studies and imaging studies with normals (Damasio, Grabowski, Tranel, Hichwa, & Damasio, 1996; Martin, Haxby, Lalonde, Wiggs, & Ungerleider, 1995; Martin, Wiggs, Ungerleider, & Haxby, 1996; Perani et al. 1995). The researchers suggested that both HSE and Alzheimer's patients show natural kinds impairments because early in the course of the disease,

AD affects temporo-limbic areas in much the same way HSE does.

There are a number of reasons to be cautious about this explanation, however. Gonnerman, Andersen, Devlin, Kempler, and Seidenberg (1997) conducted studies examining category-specific impairments in two groups of AD subjects. Both studies assessed the lexical semantic knowledge of 15 mild to moderate AD subjects using black-and-white line drawings of natural kinds and artifacts, as in the Silveri et al. (1991) research. In both of Gonnerman et al.'s studies, analyses of the group data did not yield an overall deficit for natural kinds. However, two individuals did present a double dissociation involving artifact and natural kind domains. The first subject, GP, displayed a natural kinds deficit similar to those reported in Silveri et al. (1991) and Mazzoni et al. (1991) on three tasks: picture-naming, superordinate comprehension , and word-picture matching; the second subject, NB, showed the opposite pattern: selective impairment on artifacts on all three tasks while performing at ceiling on natural kinds. If early AD affects the brain in much the same way that HSE does, it is unclear why two separate groups of mild to moderate AD subjects failed to yield an overall natural kinds deficit. It is even less clear how such a hypothesis can explain the artifact deficit of NB.

Finally, it may be misleading to equate herpes simplex encephalitis and mild Alzheimer's disease in terms of their neuropathological effects. HSE acutely affects the hippocampus and proximal cortical regions (Damasio & Van Hoesen, 1985). AD, a progressive disease, presents a more complicated pattern. Although there is substantial evidence that early AD strongly affects the hippocampal formation (e.g. Hyman, Van Hoesen, Damasio, & Barnes, 1984), this early stage does not include neocortical temporal regions affected in HSE. As the disease progresses, its effects become widespread and include the association cortices, among other regions (Pearson, Esiri, Hiorns, Wilcock, & Powell, 1985). One hypothesized scenario is that AD spreads transynaptically out of the hippocampal formation via cortico-cortical fibers (see DeLacoste & White, 1993, for a review). If this is correct, the neocortical temporal regions and fronto-parietal areas would be affected simultaneously as both regions receive hippocampal afferents. Thus AD pathology is thought to affect the hippocampal formation first and then spread to association cortices. This is quite different from HSE, which simultaneously affects the hippocampus, the anterior temporal regions, and the orbito-frontal cortex. This account provides little basis for expecting early AD to preferentially affect perceptual information.

The research described below evaluates possible computational bases of category-specific impairments resulting from both focal (e.g., HSE) and widespread (AD) types of neuropathology. Our primary goal was to determine whether the similar patterns of behavioral impairment resulting from different etiologies could be explained within a single, unified semantic system. The structure of the paper is as follows. First, we describe a theory of semantic representation in which category-specific impairments can arise from both focal and diffuse types of damage. This approach was outlined by Gonnerman et al. (1997) and is developed further here. Second, we describe an implementation of the theory within a connectionist model of semantic memory. Simulations examine the effects of different types and degrees of damage within this semantic memory system. We then assess some predictions of the model concerning the interaction between type of damage, degree of damage, and semantic category with respect to case reports in the literature.

Category-Specific Impairments: Type and Degree of Damage

In connectionist models of semantic memory (e.g., Farah & McClelland, 1991; Hinton, 1981; Small, Hart, Gordon, & Hollard, 1993), concepts such as DOG or BOAT are represented as patterns of activation distributed over computational units encoding semantic primitives. These features can include perceptual (e.g., *has-fur*), functional (e.g., *guards-house*), and encyclopedic (e.g., *man's-best-friend*) information, among others; they are simplified localist representations of knowledge that is assumed to be encoded by larger pools of neurons in the brain. Such models do not entail the claim that all semantic knowledge is reducible to simple features; in fact there is good evidence from both child development and adult performance concerning the use of nonfeatural forms of knowledge (e.g., Keil, 1989; Murphy & Medin, 1985). Rather, the claim is that important aspects of semantic knowledge relevant to tasks such as word or object naming are captured by featural representations. McRae, de Sa, and Seidenberg (1997) provide evidence that early, rapid, automatic aspects of word processing make use of featural representations, whereas later, slower, more intentional aspects of processing draw on other forms of knowledge. When a word or picture is presented in a connectionist model, a pattern of activation is computed over these semantic units. Farah and McClelland's (1991) model made two additional assumptions: (1) perceptual and functional features are topographically distinct and (2) the ratios of perceptual and functional features differ in natural kinds and artifacts. Category-specific impairments due to focal lesions could then be explained in terms of damage to different types of semantic features.

Our account of category-specific impairments due to diffuse damage turns on two additional properties of semantic representation. The first is that features differ in terms of how informative they are about the concepts in which they participate. Variants of this idea have been developed by Rosch (1975), Rosch and Mervis (1975), Smith, Shoben, and Rips (1974), Tversky (1977), Warrington and Shallice (1984), and others. For our pur-

poses, we need to incorporate the minimal assumption that features differ in the degree to which they help distinguish among concepts. Thus, some features are more relevant (i.e., informative, distinguishing, defining) than others with respect to categorizing an entity as an instance of a given type. For example, the feature *has-fur* provides little identifying information at the basic level that most mammals are named. The conjunction of *has-fur* and *has-claws* provides some additional constraint but not enough to identify the animal. Knowing that the referent *has-fur* and *has-stripes,* however, greatly increases the likelihood that the entity is a tiger. *Having-stripes,* then, is more informative about the object's identity within the animal category than is *having-fur* or *having-claws.* In part this is because *fur* co-occurs with several other features in many animal concepts, while *stripes* occurs less often and is not strongly correlated with other features (e.g., it co-occurs with *fur* in TIGER but *hair* in ZEBRA). Similarly, *having-red-breast-marking* is important in distinguishing a ROBIN from a BLUEJAY, which both share *having-feathers, having-a-beak,* and *having-claws.* Such features do not provide necessary or sufficient information for identifying an object; rather, they tend to be more informative on a probabilistic basis. As Warrington and Shallice observed, artifacts tend to be distinguished in terms of functional features and natural kinds in terms of perceptual, but these generalizations are not inviolable (Malt & Johnson, 1992).

The second important aspect of semantic representation is the existence of intercorrelations among features and differences in the distribution of these intercorrelated features across natural kinds and artifacts. Many people have observed that semantic features are often correlated with one another, especially for natural kinds concepts (Keil, 1987, 1989; Malt & Smith, 1984). For example, an animal that has fur is also likely to have claws, whiskers, and a tail. Evidence that people encode such intercorrelations is provided by McRae et al. (1997). McRae et al. gathered feature norms that indicated that the features that subjects listed for natural kinds such as TIGER were more highly correlated with one another than the features listed for artifacts such as CHAIR. Moreover, for natural kinds, the extent to which words overlapped in terms of correlated features predicted the magnitude of semantic priming effects in behavioral experiments. For artifacts, in contrast, overlap in terms of simple features predicted priming effects. McRae et al. developed a connectionist model of natural kinds and artifact categories that acquired knowledge of these relationships among features and simulated the results of several experiments. Such correlated features play a central role in our account of category-specific deficits in AD; they provide the basis for catastrophic loss of categories without focal damage to perceptual or functional types of features. The models described below acquire information about the informativeness of features and the correlations among them on the basis of exposure to concepts. Acquiring this statistical knowledge can be seen as an important aspect of concept development.

These two additional aspects of semantic representation were implicit in earlier models such as Farah and McClelland's (1991), but did not figure directly in their account of the phenomena that were the focus of attention. The distributions of correlated features across natural kind and artifact categories and the fact that features differ in terms of how much they serve to distinguish one object from others derive from facts about the nature of objects in the world. Thus, we would expect these characteristics to be represented in any model that encodes the relevant sorts of featural primitives for a sufficiently large sample of objects and has the capacity to represent relations among features.

Within this framework we derive category-specific impairments in AD as follows. We assume that damage to the semantic system affects random features by reducing the degree to which they are activated. Whether features that fail to become active should be considered "lost" or merely "inaccessible" (Rapp & Caramazza, 1990; Shallice, 1988) is not relevant to this account.[2] What matters is that some features become unavailable due to damage. When these include informative features that happen to distinguish between two entities (e.g., *having-stripes* distinguishes TIGERs from LIONS), errors occur in naming these objects. The more informative a particular feature is to a given object, the more likely that object will be misnamed when the feature is damaged. Conversely, losing a feature that is uninformative has little behavioral consequence. Consider the following example. *Has-fur* is less informative than *barks* because many animals have fur but only a few animals, such as dogs, seals, and hyenas, bark. When damage makes *has-fur* unavailable, a patient might still produce the word DOG (although be unable to verify whether dogs have fur). If *barks* were unavailable, however, the remaining semantic features might afford more than one response (e.g., DOG *or* CAT), increasing the likelihood of an overt naming error. Thus the behavioral consequence of damage is significant relative to the informativeness of the lost feature.

Because features that occur across many items are less informative than those occurring in only a few items and because natural kinds tend to share more features across items, natural kinds have relatively fewer informative features than do artifacts. Although damage to these features can lead to errors on individual natural kind items, the behavioral effect of damage to informative features will be stronger for artifacts (i.e., more artifacts will be misnamed) because they have proportionally more informative features than natural kinds.

The intercorrelations among features figure in the category-specific impairments as follows. We assume that the architecture of the semantic system permits the encoding of these correlations. At low levels of damage features that participate in these correlations are more resistant to random damage than other features because

of strong collateral support within a set of intercorrelated features. In a highly interactive architecture it will be possible to complete a semantic pattern despite moderate damage affecting individual features. As damage becomes more severe, however, the feedback excitation among intercorrelated features will itself be reduced until it can no longer provide compensation. At that point, the loss of individual features within an intercorrelated set can have catastrophic effects, reducing the activation of other features in the set. This may in turn cause the other features to fall below threshold levels of activation and become inactive, yielding a cascade of lost features. Thus, an entire set of intercorrelated features can be impaired together (see Ruppin & Reggia, 1995, for a discussion of catastrophic loss in associative memory systems).

The predicted behavioral effect of this damage scenario is an inability to name all of the items that rely on the affected intercorrelated features. Natural kind categories tend to include many items with such features, creating their tendency to show "category-specific" impairments. Artifact categories tend to entail fewer of these intercorrelated features; therefore individual items will typically be impaired as features that are highly informative about their identities are lost. These predictions must be modulated by two further observations. First, items within natural kind categories differ in terms of the extent to which they entail intercorrelated features; thus not all members of a category will necessarily be equally affected (compare lemons, limes, and oranges, which share several intercorrelated features, with pineapples or tomatoes). We therefore predict a simultaneous loss of many items that share feature structure. Idiosyncratic items within such categories that do not entail the correlated features (such as tomato) should behave more like typical artifacts: They should be more susceptible to mild damage and lost on an item-by-item basis rather than en masse. Second, there will be natural kind and artifact categories that deviate from these predictions because they have distributions of features that are more typical of categories in the other domain (e.g., musical instruments resemble natural kinds more than they do other artifact categories). This is consistent with our view that semantic memory is structured in terms of distributions of types of features, not categories per se.

To summarize, nonfocal damage yields different progressions for different categories depending on their featural properties. Categories with many informative features and few intercorrelated ones, which tend to be artifacts, lose individual items as random damage affects relevant informative features. Loss of items is predicted to be roughly linear with the amount of damage. On the other hand, categories with many intercorrelated features, which tend to be natural kinds, should follow a nonlinear trajectory, with exemplars initially resistant to minor damage and then clusters of items simultaneously affected by the loss of shared sets of features. The net result is that the degree of damage interacts with the structure of semantic representations to produce a double dissociation over time. An initial, mild difficulty with artifacts compared to natural kinds is predicted to yield to much more severe, catastrophic impairment of natural kinds as damage to the semantic system increases. In the most severe phase of semantic impairment, of course, patients may become globally impaired.

Behavioral Evidence

Gonnerman et al. (1997) presented cross-sectional studies of the picture-naming performance of two groups of probable AD subjects that supported this hypothesized progression. In the first study, 15 clinically diagnosed AD subjects named 36 black-and-white line drawings of common objects, including both natural kinds and artifacts. They found that subjects with the best overall performance showed a slight disadvantage for artifacts and that most of the more impaired subjects showed more difficulty with natural kinds. To determine if this finding would replicate, a second study was conducted with a different group of 15 AD subjects who named 72 black-and-white line drawings of common objects. The results for subjects in the second study are shown in Figure 1. Three of the five subjects with the best performance had more difficulty with artifacts than with natural kinds, with the other two subjects performing equally well on items from both domains. The ten remaining subjects did poorly on natural kinds relative to artifacts, with the size of the difference increasing as overall performance declined. In general then, the data from both studies are consistent with the hypothesized progression of an early, mild difficulty with artifacts over natural kinds, which becomes a definite natural kinds deficit as semantic deterioraton increases.

THE MODEL

We developed a connectionist computational model of the proposed semantic system to explore three questions: (1) whether diffuse, progressive damage to the model's semantic system would yield the effects observed by Gonnerman et al. (1997), (2) whether the model could explain reported variation among AD subjects (e.g., why only some studies observed a natural kinds deficit for groups of AD subjects), and (3) whether the same model could account for category-specific deficits arising from both AD and focal brain damage. In addressing these questions we exploited two aspects of the modeling approach. First, if each each simulation is considered an artificial "subject," we can create and evaluate longitudinal data from many more subjects than traditional neuropsychological methods permit in a comparable time period. Second, because each simulation begins with an identical semantic system, we can control for premorbid differences across subjects and thus ob-

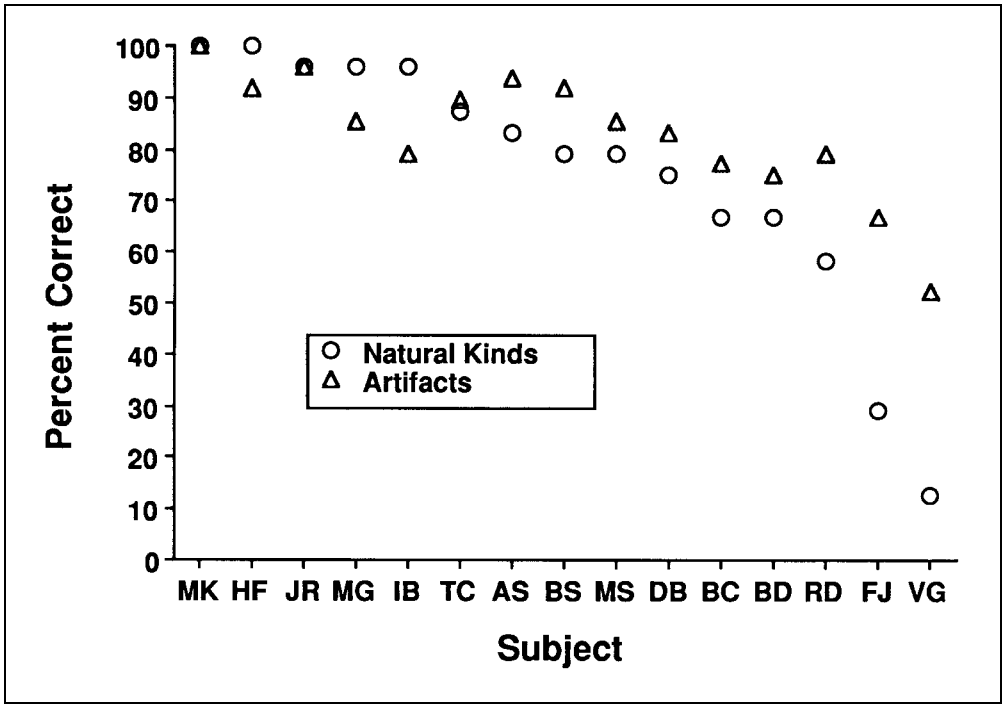

Figure 1. Performance of 15 Alzheimer's subjects on a picture-naming task. Percentage correct scores for natural kinds and artifacts are plotted for each subject, ordered according to degree of impairment on natural kinds. From Gonnerman et al. (1997).

serve the effects of different pathologies without this confound.

We implemented a hetero-associative memory network that mapped between semantic and phonological representations. The model was one of a general class of models called *attractor networks,* where patterns of activity form stable points in processing space called *attractors* (Amit, 1989; Hopfield, 1982). Each attractor develops a region surrounding it, called a *basin of attraction,* such that activity in this region will eventually settle into the attractor (see Plaut, 1991; Plaut & Shallice, 1993, for more extensive discussion). In our model the attractor states corresponded to the phonological and semantic patterns for words.

Architecture

The network consisted of four layers, a phonological layer, a semantic layer, and two hidden layers, as shown in Figure 2.[3] One hidden layer was fully connected to the phonological layer, the other, to the semantic layer; these are labeled Phonological Clean-Up units (PCU) and Semantic Clean-Up units (SCU), respectively (cf. Plaut 1991). In addition, the phonological and semantic layers were fully connected to each other, but units within a layer were not. The model contained 18040 weighted connections.

Input to the model was either a phonological or a semantic pattern of activation. To model comprehension, a phonological input was provided and the associated semantic pattern was computed. To model production, a semantic input was provided and the corresponding

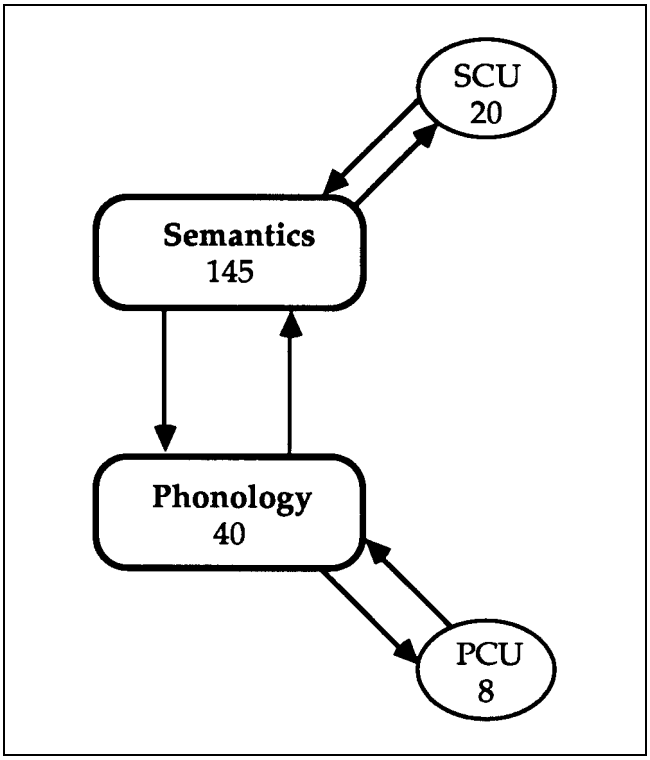

Figure 2. Model architecture: An attractor network with four layers: Semantics, Phonology, and two layers of clean-up units. The numbers in each oval indicate the number of units in that layer, and arrows indicate full connectivity between groups.

phonological output was computed. The focus of the present work was on the production task.

Representations

The training set included 60 words, half of them natural kind terms and half artifacts. Categories included animals (15 items), fruits and vegetables (15), vehicles (10), clothing (10), and tools (10). Phonological patterns were random binary vectors of length 40 with a mean of 12.9 phonological features (*SD* = 3.9) active per word. A vector served only as a unique identifier for a word. This representation captured the fact that the mapping between phonology and semantics is essentially arbitrary for monomorphemic words; however, because this code did not represent an actual phonological structure, any errors that were made were not phonologically meaningful.

The semantic representations were designed to implement properties that our theory suggests are relevant to explaining category-specific impairments: (1) the different proportions of perceptual and functional information in the natural kind and artifact domains, (2) the greater numbers of correlated property pairs in natural kinds compared to artifacts, and (3) the fact that features differ in informativeness. Semantic representations were developed on the basis of feature norms collected from 30 undergraduate subjects. Subjects were given a word such as BOOK and asked to list perceptual and functional properties of the item. This collection of feature lists was used as a guide in constructing the semantic representations, with the most frequently named features included in them.[4] Each of the 60 words was represented as a binary vector over 145 semantic units; a 1 indicated the presence of a feature and a 0, its absence. On average each word had 8.2 active semantic features with a standard deviation of 1.0. A one-way analysis of variance revealed no significant difference in the mean number of features by category, although there was a trend toward more features in natural kinds than artifacts, ($F(4, 55) = 3.54, 0.05 < p < 0.10$). Of the 145 semantic features, 88 were perceptual and 57 functional. A feature was considered perceptual if it described a visual, auditory, or tactile property of a concept (e.g., *has-wheels*). Functional features were those describing what an item does or what it is used for (e.g., *used-for-driving*). Attributes that met neither of these criteria, such as encyclopedic knowledge (e.g., *found-in-Africa*), were not included.

Each word was represented by both perceptual and functional properties. Table 1 shows the distribution of features by categories. Overall the model had a 1.5:1 ratio of perceptual to functional features; the ratio for natural kinds was 3.0:1, whereas the ratio for artifacts was only 1.4:1. Our ratios do not exactly match Farah and McClelland's (1991), but the distributions are similar in so far as natural kinds have more perceptual features than do artifacts, whereas artifacts have greater numbers of functional features than do natural kinds.

Having constructed representations for the 60 words based on these principles, we then examined the extent to which features were correlated. Such correlations must exist in the representations of concepts if they are to be encoded by the network during training. Following McRae et al. (1997), the Pearson product moment correlation was computed for all pairs of semantic features. Of the 10,440 possible correlations, only 416 were significant ($r > 0.216, p < 0.05$). Figure 3 shows the mean number of correlations by category. Natural kinds had reliably more correlated features (mean = 18.3, *SD* = 8.5) than did artifacts (mean = 13.8, *SD* = 5.8), $F(4, 55) = 5.59$,

Table 1. Feature Distributions Across Categories

	Number of words	*Total number of features/word*	*Number of perceptual features/word*	*Number of functional features/word*
Natural Kinds				
Fruits & vegetables	15	8.3	6.1	2.2
(*SD*)		(0.88)	(0.70)	(0.41)
Animals	15	8.6	6.6	2.0
(*SD*)		(1.30)	(1.12)	(0.76)
Artifacts				
Vehicles	10	7.9	4.6	3.3
(*SD*)		(0.88)	(0.70)	(0.48)
Clothing	10	7.9	4.5	3.4
(*SD*)		(1.00)	(0.97)	(0.52)
Tools	10	8.0	5.0	3.0
(*SD*)		(0.94)	(0.67)	(0.67)

$p < 0.05$. Examples of correlated property pairs are shown in Table 2. The highly interactive nature of the architecture allowed these correlations to be stored as connection strengths within the semantic system (i.e., the semantic and SCU layers of the model and the connections between them).

Earlier it was hypothesized that damaging highly informative features would affect naming behavior by making similar objects less distinguishable and thus harder to name. A simple measure of feature informativeness is given by 1 over the number of objects the feature occurs in. For example, in our corpus the feature *barks* occurred only in the word DOG and consequently *barks* was highly informative. Conversely, the feature *eaten-by-people* occurred in 18 words giving it an information score of 0.056, thus not a very informative feature.

Because artifacts have fewer intercorrelated features, their features tend to be more informative than those in natural kinds. We calculated the mean feature information score (mean = 0.298) and labeled all features with scores above the mean as "informative." Artifacts within the model had reliably more informative features per word (mean = 3.4) than natural kinds (mean = 2.4), $t = 2.53$, $p < 0.05$. Consequently random damage is more likely to strike an informative feature in an artifact object than in a natural kind object and thus increase the likelihood of misnaming that item. It is important to note that random damage will affect both artifacts and natural kinds, but with limited damage it will affect artifacts more heavily. Later, once sets of intercorrelated features are lost, many items from categories relying on those intercorrelations will be impaired.

The informativeness of features necessarily depends on both the set of words in the chosen corpus and the set of features used to represent these words. Consequently, while the values we have calculated are exact in the context of the model, they should be taken as only rough approximations of how informative individual semantic features really are. The feature representations used in the model were based on empirical norms reflecting subjects' knowledge of all concepts, not merely those included in the model, suggesting that these relationships should be preserved in larger-scale models. We should note, however, that the semantic representations maintained some of the limitations of the empirical norms from which they were derived. Subjects only listed the features they found most salient for each concept; thus, the fact that *has-an-engine* was not listed for BUS and the fact that *has-a-mouth* was only listed for HIPPO are reflected in the model's representations for these concepts. Thus, the model representations capture only the most salient features of concepts, not everything a subject might know about them.

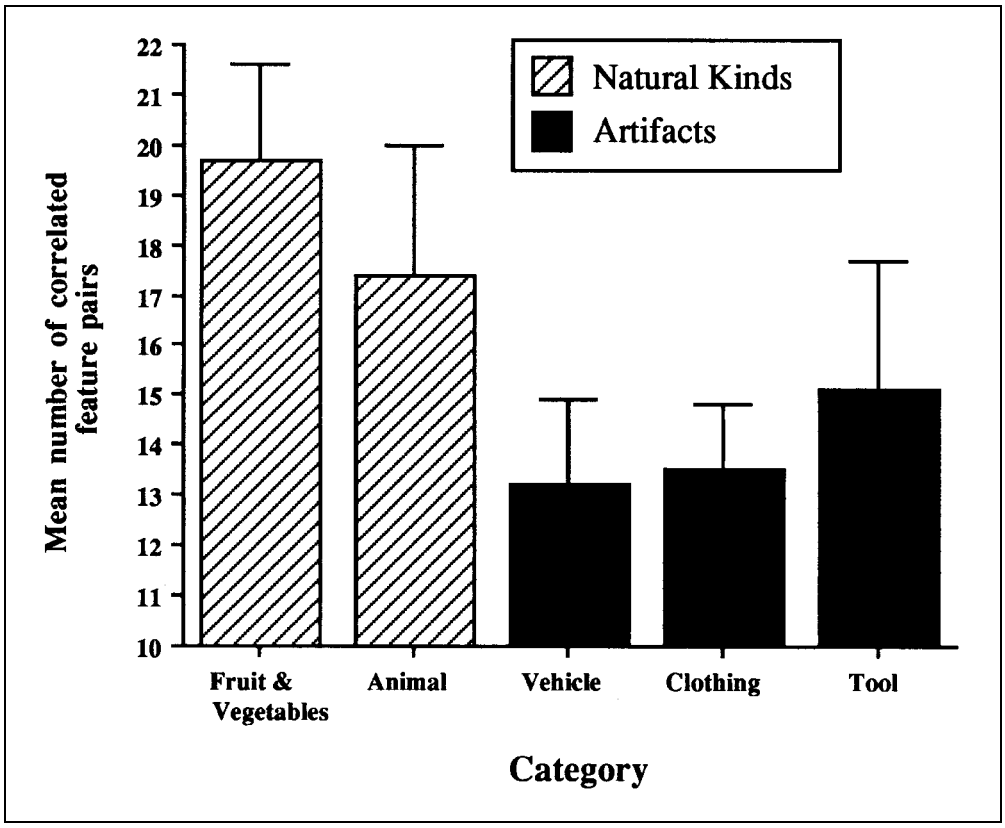

Figure 3. Mean number of correlated property pairs by category with 95% confidence error bars. Natural kinds have reliably more intercorrelations than artifacts.

Table 2. Examples of Correlations in the Semantic Representations

Correlated property pairs	*Pearson r*
High correlation	
is-sharp—used-for-cutting	1.000
has-a-handle—used-manually	0.927
is-juicy—is-sour	0.809
made-of-wool—is-soft	0.809
is-tasty—is-sweet	0.718
Medium correlation	
keeps-body-warm—made-of-cloth	0.688
used-for-transportation—has-wheels	0.528
has-buttons—has-a-zipper	0.483
eats—has-a-tail	0.417
made-of-wood—made-of-metal	0.348
Low correlation	
has-4-legs—is-gentle	0.288
has-legs—has-a-long-tail	0.287
eats—has-4-legs	0.279
is-small—has-fur	0.247
pollutes-the-air—has-an-engine	0.245

Training

The goal of the training procedure was to obtain a set of weights that allowed the model to perform two functions: It had to encode the essentially arbitrary mappings between semantics and phonology and it had to learn attractor states for each word's semantic and phonological patterns. The implications of this training are twofold. First, there was only a single phonological system for both production and comprehension, as opposed to separate lexicons for phonological input and output (cf. Hillis & Caramazza, 1991). This system, consisting of the phonology units, phonological clean-up units, and the connections between them, encoded phonological regularities and processed both phonological input and output representations. Such a system is capable of producing dissociations between errors of production and comprehension, although a discussion of this behavior is outside the scope of this paper (Devlin, in preparation). Second, individual words corresponded to attractor states in a dynamic processing space as opposed to static nodes in a mental lexicon. Thus, although features are instantiated as individual units, words are not. The meaning of a word was its pattern of activation over semantic features, whereas its phonological form was its pattern of activation over phonological features. Both of these claims, that there is only one phonological system and that words correspond to attractor states, are not consequences of the modeling but rather are substantive claims about the normal cognitive system.

The model was trained using the recurrent backpropagation algorithm with conjugate gradient descent and a line search technique to determine step sizes (cf. Hassoun, 1995). The network began in a random state with all connection strengths initialized to values in the range [−.1.0.1] with a random uniform distribution. Two exemplars per word were then presented to the net, one providing phonology as input and expecting semantics as output (P→S), the other providing semantics as input and expecting phonology as output (S→P). The 120 exemplars (60 S→P, 60 P→S) constituted one epoch. Weights were updated after each epoch. Training ended when the activation of all output units was within 0.2 of the target values. Training took 210 epochs.

During training each exemplar was active for 10 time steps. A pattern of activation was clamped over the input units for the first three time steps until activity had spread to all layers of the net. Then the input units were unclamped for seven time steps, and during this time target values existed for every unit in the input and output groups. The net had to learn to compute the appropriate output given an input and to correctly maintain an input pattern without external assistance.

To illustrate, given the P→S AIRPLANE training exemplar, a phonological input pattern was clamped over the phonology units to begin a training trial. At the next time step, activation spread to the semantic layer and to the PCU. During the third time step, it spread further to the SCU. From the fourth to the tenth steps, the phonological units were unclamped (i.e., no external input was provided) but were expected to maintain their current levels of activation, whereas the semantic units were expected to display activation of properties such as *has-wings, flies, made-of-metal,* etc. Similarly, the S→P AIRPLANE training exemplar mapped in the opposite direction by presenting the semantic features of AIRPLANE and computing the phonological pattern for the word.

EXPERIMENT 1: MODELING ALZHEIMER'S DISEASE

Connectionist models are often described as *neurally inspired* because they capture general properties of neural computation, such as distributed representation and massively parallel processing, while typically abstracting away from neurophysiological details (Rumelhart & McClelland, 1986). Most models incorporate some general biological constraints and are broadly compatible with others. The principal goal of first-generation at-

tempts to simulate effects of brain injury was to account for nontransparent aspects of the behavioral data (e.g., Hinton & Shallice, 1991; Patterson, Seidenberg, & McClelland, 1990). However, recent models of normal and disordered cognition have begun to incorporate more specific neurophysiological constraints (Cohen & Servan-Schreiber, 1992; Horn, Ruppin, Usher, & Herrmann, 1993). Our goal was to simulate the behavioral effects of two different types of brain damage using mechanisms that capture basic features of these neuropathologies. Thus, the etiology of the disease constrained the manner in which the model was damaged.

A large literature exists concerning the neuropathology of AD, with research ranging from the molecular biological level (e.g., Yankner & Mesulam, 1991) to the anatomical level (e.g., Price, Davis, Morris, & White, 1991; see Henderson & Finch, 1989, for a review). We focused on three elements: (1) regional distributions of histopathological markers involving association cortices early in the disease (Pearson et al., 1985; Rogers & Morrison, 1985), (2) cortical synaptic loss as a major pathological component of AD correlated with degree of cognitive impairment (DeKosky & Scheff, 1990; Gibson, 1983; Scheff, DeKosky, & Price, 1990; Terry et al., 1991), and (3) AD as a dynamic process of progressive degeneration. These properties all constitute general, noncontroversial, well-established aspects of AD that can be plausibly incorporated in a connectionist framework, insofar as they can be related to basic properties of such nets. The fact that AD is characterized by relatively diffuse damage to most regions of association cortex suggests that even if perceptual and functional features are localized in different parts of the cortex (e.g., temporal vs. fronto-parietal), both types of features are likely to be affected by AD. Moreover, the primary histopathological markers of AD, neuritic plaques and neurofibrillary tangles, are correlated with a loss of synaptic junctures, the primary site of neuronal communication (Masliah, Hansen, Mallory & Terry, 1991; Masliah, Terry, Mallory, Alford, & Hansen, 1990; Terry et al., 1991). Within our model, the best analogy to this pathology is provided by removing connections between units. The intent here is not to suggest that features in the model map directly onto individual neurons or that connections between features correspond to synapses. Rather, the information represented as a single semantic feature in the model is assumed to be represented in the brain by a large number of neurons with complex interactions. Thus the loss of synapses in AD would presumably have a graded effect on the ability to use the information from a single feature. Therefore removing a feature within the model is too abrupt a form of damage. Instead, removing connections causes features to gradually decay in efficacy as their input degrades. To simulate the widespread affects of AD, connections within the semantic system involving both perceptual and functional features were randomly chosen and removed. Finally, because AD is a degenerative disease, damage to the model was applied progressively. Again, this procedure is intended to capture a basic characteristic of the disease process, while abstracting away from more specific details.

In the AD simulations, damage was limited to connections within the semantic system, namely, those between the Semantic and SCU units. Because the association cortices affected by AD presumably play an integral role for both phonology and semantics, this was not an anatomical constraint. Instead, it was based on the fact that while AD consistently impairs semantic processing, there is little evidence that AD directly affects phonological processing (but see Patterson, Graham, & Hodges, 1994, and Patterson & Hodges, 1992, for evidence of indirect phonological impairment in AD). Consequently, connections directly affecting phonological computations were spared in these simulations, and those affecting both perceptual and functional semantic information were randomly chosen and removed.

Testing

Testing focused on the equivalent of a picture-naming task. Naming a black-and-white line drawing normally begins with analysis of the input by the visual system, which activates perceptual semantic features corresponding to elements present in the image (Caramazza, Hillis, Rapp, & Romani, 1990; Shallice, 1993). As these visual features increase in activation, they begin to activate features associated with properties of the object not present in the visual image. For instance, a picture of a school bus might activate the semantic features *has-wheels, is-large,* and *is-long,* which would in turn activate other properties of buses not present in the image, such as *is-yellow, has-an-engine,* and *used-for-transportation.* As the semantic pattern builds, it activates a corresponding phonological pattern. Partial activation in the phonological system may itself feed back on semantics. Once the phonological pattern is sufficiently specified, the subject can initiate production of the object's name. Damage to the semantic system interferes with this process, causing activation of incorrect semantic representations that can produce naming errors.

In our simulation, we activated an item's semantic pattern as input and clamped it for three time steps, placing the model in the trained attractor state corresponding to the word's semantics and phonology. Randomly distributed damage within the semantic system affected processing in three ways: (1) A semantic pattern was severely disrupted such that the wrong output was produced, (2) a semantic pattern included some incorrect features, or failed to activate some correct features, and yet was still close enough to the veridical pattern to produce the correct phonology, or (3) a given item was not affected at all. Although the process of picture naming in the model did not directly correspond to that of subjects, (i.e., the model began with a correct semantic

pattern and drifted away as a result of damage, whereas subjects presumably begin with only visual features and fail to generate the rest of the correct pattern), we assume that the final computed patterns are roughly equivalent.

To simulate the progressive nature of the disease, the model was lesioned cumulatively. Initially the lesions were in small increments that gradually grew larger as damage accumulated. A lesion consisted of removing a percentage of the 5800 weights between the semantic layer and the SCU. The first lesion randomly deleted 1% of the connections, or 58 weights. This procedure was repeated until 20% of the connections were gone. Lesions then occurred in 5% increments until the 50% damage level was reached. Finally, all remaining lesions were in 10% increments. After each lesion, the model was evaluated on the above task for all 60 S→P exemplars. The pattern of activation over the phonological units was considered the output vector.[5] An output was judged correct when the Euclidean distance from the output vector to the target phonology was closer than the distance between the output and the phonological pattern of any other word. One consequence of this evaluation scheme is that no matter how distorted a phonological output pattern was, it was judged one of the 60 "words." One complete simulation consisted of 30 lesions (1%, 2%, …, 20%, 25%, …, 50%, 60%, …, 90%). Each simulation was repeated 50 times.

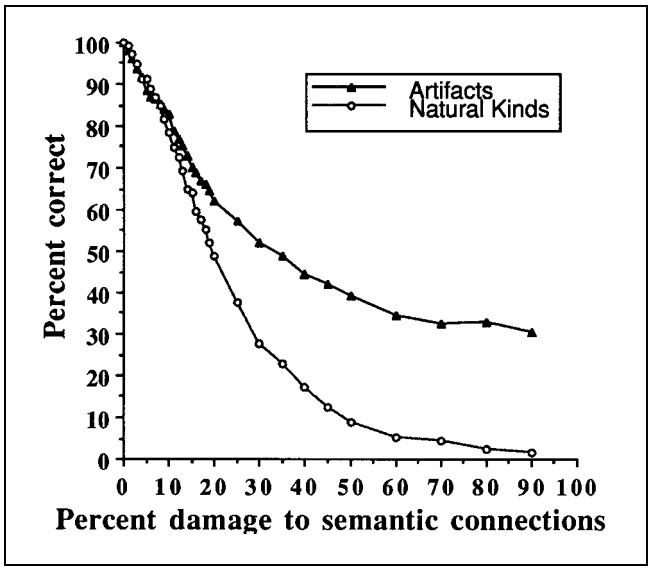

Figure 4. The model's performance averaged over all 50 trials. As damage increased a preferential impairment in natural kinds arose.

Results and Discussion

Figure 4 presents the results averaged over all 50 simulations. For the averaged scores of the entire set, the general pattern is one where increasing damage led to a preferential impairment of natural kinds over artifacts. Figure 5 displays the same data separated into the two main subpatterns of performance. In Figure 5a, the modal pattern is shown. In 38 out of 50 simulations, low levels of damage led to more errors on artifacts than natural kinds, with increasing damage producing a crossover to more errors on natural kinds than artifacts. In Figure 5b, the model's performance is shown averaged over 11 simulations that displayed only a natural kinds deficit throughout the entire progression.

Thus, in most cases (76%) the model displayed the expected progression of an initial mild selective difficulty for artifacts that crossed over into a preferential impairment of natural kinds as damage accumulated. Interestingly, 22% of the simulations resulted in more difficulty with natural kinds that persisted throughout the simulation. The remaining one simulation was both aberrant and intriguing. It had initial problems with natural kinds, which then reversed into an artifacts deficit as the damage increased, before finally turning into an equal impairment for both domains. We return to this case later.

The model exhibited considerable variability across simulations. This is evident when comparing the two patterns discussed above, but also within each pattern. For example, among the 38 trials in which the model experienced greater difficulty with artifacts before crossing over into a natural kinds impairment, both the magnitude of the initial artifacts deficit and the point of crossover varied (see Figure 6). In most of these trials the effect of low levels of damage on artifacts was only mildly greater than its affect on natural kinds (e.g., Figure 6, left). On a few trials, however, the initial artifacts impairment was quite large; one such trial is shown in Figure 6, right. Similarly, the crossover point between an artifact and natural kinds deficit ranged from 3 to 20% damage across simulations. Consequently, the strong nonlinear deterioration of natural kinds (the result of a set of intercorrelated features failing en masse) is clearly seen in individual simulations, such as in Figure 6, but less pronounced when averaged together (see Figure 5).

The variability in the effects of damage on performance in the modeling results is consistent with the variability observed among AD subjects and helps to explain some seeming inconsistencies in the behavioral literature. Recall that two studies found a category-specific deficit for natural kinds in groups of mild to moderate AD subjects (Mazzoni et al., 1991; Silveri et al., 1991), while one found no overall effect of category (Gonnerman et al., 1997). It seems clear from Figure 4 that the sampling of subjects in a group study will play a critical role in determining the observed findings. Silveri and Mazzoni's subjects could simply have tended to be further along the progression than those in the Gonnerman et al. study. Indeed, the picture-naming scores from individual subjects as displayed in Mazzoni et al.'s Figure 4b indicate that those subjects scored in the 80

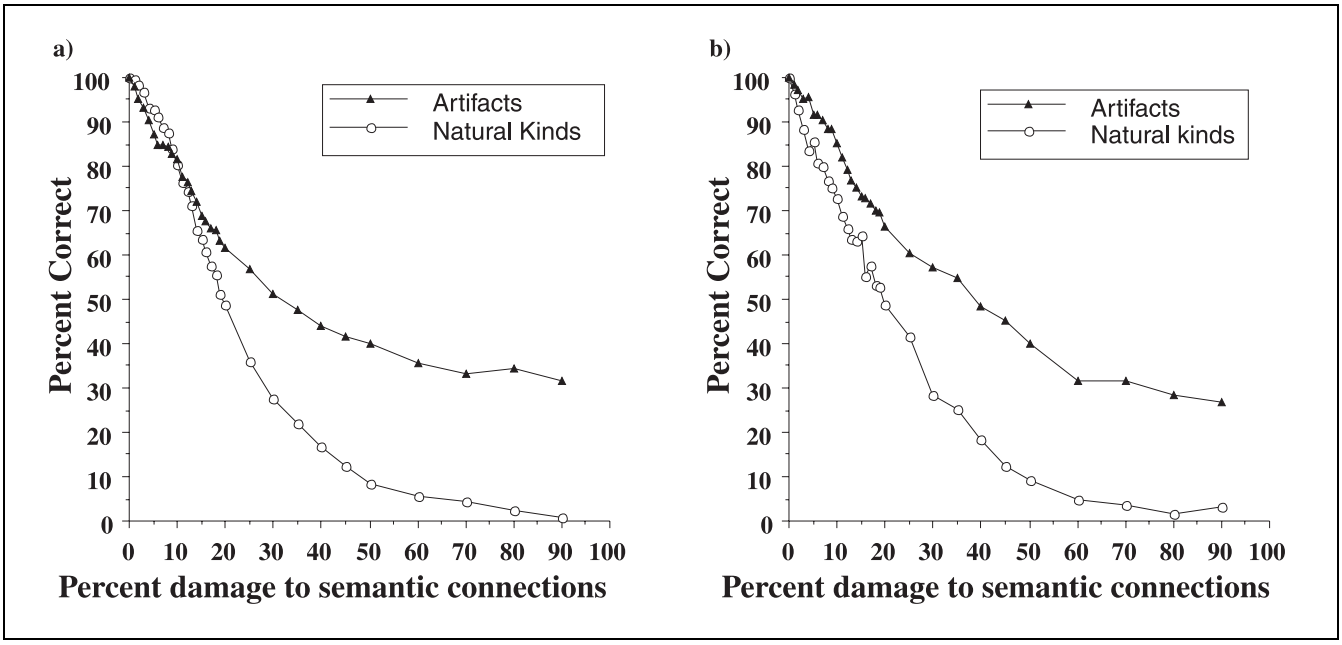

Figure 5. (a) The model's performance averaged over the 38 trials displaying the most frequently observed progression. Low levels of damage led to more errors on artifacts than natural kinds, with increasing damage producing a crossover to more errors on natural kinds than artifacts. (b) The model's performance averaged over the 11 trials where the model made consistently more errors on natural kinds.

to 100% correct range for artifacts but in the 30 to 80% correct range for natural kinds, a larger overall impairment than observed by Gonnerman et al. In fact, most of Gonnerman et al.'s subjects appeared to be relatively close to their hypothesized crossover points, whereas the two subjects who exhibited a pronounced deficit on natural kinds would be more like the subjects in the Silveri and Mazzoni studies.

The modeling data are also consistent with Gonnerman et al.'s (1997) two case studies of subjects NB and GP, who displayed a double dissociation with regard to semantic domain. The higher-functioning subject, NB,

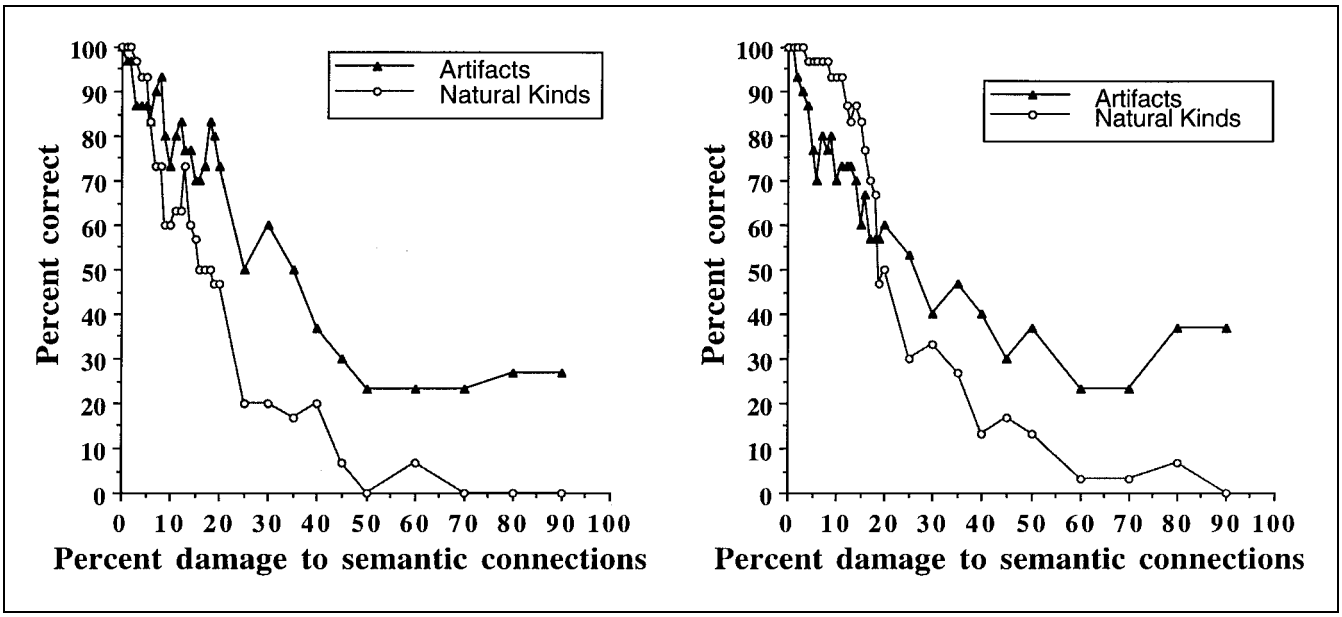

Figure 6. (right) The model's performance on simulation 22. On this trial the initial artifacts deficit is relatively clear, and the crossover point does not occur until 18% damage. (left) The model's performance on simulation 18. On this trial the initial difficulty with artifacts was quite mild, as it was in the most cases.

made fewer errors overall and had more difficulty with artifacts than with natural kinds. Indeed her errors were evenly spread across all artifact categories, as one would expect if this deficit were the result of damage striking individual distinguishing features. GP made more errors overall and displayed a preferential impairment of natural kinds, as we would expect for a subject further along in the progression. The only surprising result from these subjects was the magnitude of NB's artifact deficit. The model typically produced a much smaller artifacts deficit; however, it should be noted that the model did produce a more pronounced initial artifacts impairment on occasional trials, such as the one in Figure 6, right. Thus NB may represent a rare case in terms of the size of her impairment, but she nonetheless fits within the range of patterns observed in the model.

Finally, recall the simulation that produced an aberrant pattern (Figure 7): There was an initial natural kinds deficit followed by an artifacts deficit before global anomia. Most of the initial natural kind errors were in the fruits and vegetables category, whereas very few were in animals. The artifact errors, on the other hand, were distributed more equally across the tools, clothing, and vehicles categories. Thus, the random damage initially struck the intercorrelations underlying fruits and vegetables enough to preferentially affect that category. That particular pattern of damage, however, was not enough to cause a catastrophic loss of the intercorrelations underlying animals, and consequently animals were relatively spared. With fewer intercorrelations, items from the artifact categories were gradually lost as damage increased and distinguishing features were affected. Thus the immediate loss of many fruits and vegetables produced a relative impairment of natural kinds that gradually crossed over into an artifact deficit as individual artifacts continued to be lost. Finally, enough damage accumulated that the intercorrelations among the animals were lost, affecting many items from this category and leaving the model generally impaired.

In summary, these results parallel those observed in AD subjects, namely, that on average a small artifact deficit will precede a larger preferential impairment of natural kinds when the semantic system suffers progressive, diffuse damage. Although the majority of simulations show this pattern, the number of artifacts on which they initially err is usually small; such deficits may not reach statistical significance for individual subjects and may therefore easily escape notice. Indeed, mean performance from groups of subjects would probably reveal only the more pronounced deficit, the natural kinds impairment. These simulations illustrate how a probabilistic account of damage yields a variety of impairment patterns, including the small initial artifact deficit, no initial artifact deficit, and an initial artifact advantage in rare cases. All of these patterns have the same computational basis, however. These results highlight the importance of understanding the factors that give rise to individual differences among subjects, a point to which we return in the "General Discussion."

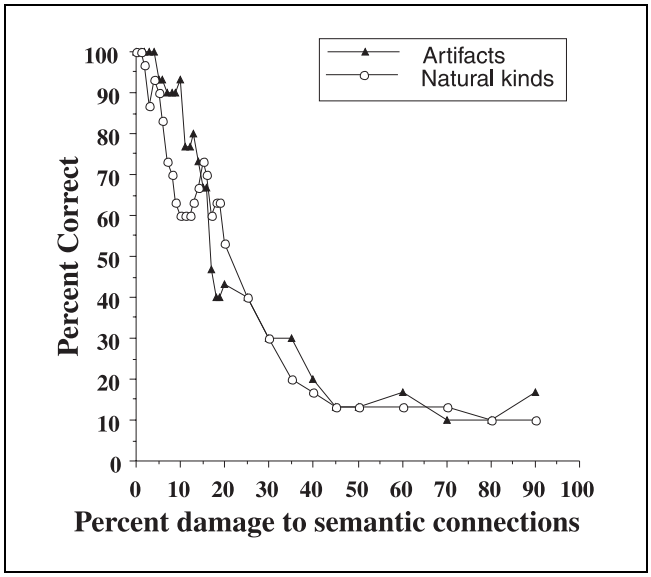

Figure 7. Performance of the model on simulation 7 where the model displayed an initial natural kinds deficit, then an artifacts deficit, and finally a global deficit.

EXPERIMENT 2: MODELING AD AS UNIT LOSS

In the first experiment AD was modeled as a progressive loss of connections within the semantic system on the assumption that this form of damage was most closely analogous to the neuropathology of AD. It might be argued, however, that motivating the type of damage to the model by analogy to the actual neuropathology was meaningless because the model is so far removed from neurobiological reality. If this is correct, one might expect other forms of damage to produce similar effects. To address this question, we conducted another experiment in which AD was modeled as a progressive loss of semantic units. If the type of damage to the semantic system is relevant, we would expect both that different types would yield different results but also that the one that matches the neuropathology more closely would provide a better account of the behavioral data.

Testing

Damage was applied progressively to semantic units. Each "lesion" consisted of removing a certain percentage of semantic features randomly selected from both perceptual and functional units. The first lesion removed 1% of the units, or 1.5 units on average, and was repeatedly applied until 10% of the units were removed. Lesions then occurred in 5% increments until the 50% damage level was reached. All subsequent lesions were in 10%

increments. One complete simulation consisted of 22 lesions (1%, 2%, …, 10%, 15%, …, 50%, 60%, …, 90%). At each level of damage the model was evaluated on its performance for all 60 S→P exemplars. Each simulation was repeated 50 times.

Results and Discussion

Figure 8 presents the overall results of the 50 simulations. The data demonstrated a small natural kinds deficit for relatively low levels of damage that then faded into a global anomia as the model became severely impaired.

The most striking aspect of the data was the variability of individual simulations. Although the overall result was a natural kinds impairment followed by a more equal distribution of errors across domains, only 15 of the 50 trials showed this pattern. An additional 9 trials followed the overall pattern found in Experiment 1 with more difficulty for artifacts at low levels of damage followed by a natural kinds deficit and finally a more global impairment. The majority of simulations did not conform to either of these patterns. Twenty-six simulations presented a variety of results including selective deficits for artifacts (4) or natural kinds (4), equal distributions of errors across domains (2), and a host of more complicated patterns that varied with the degree of damage (16).

Thus, the different types of damage introduced into the model were associated with different behavioral results, with the form of damage that more closely resembles the pathological affects of AD providing a much better fit to the subject data. Whereas the simulations in Experiment 1 yielded two main deficit patterns that could be related to behavioral data in a straightforward way, the simulations in Experiment 2 yielded a broader range of patterns, including ones not observed in any AD subjects.

EXPERIMENT 3: MODELING FOCAL BRAIN DAMAGE

Because category-specific deficits are more typically associated with focal brain damage, any explanatory model would have to demonstrate how the same behavioral deficit could arise from different pathologies. In the third experiment, therefore, we attempted to do so by replicating Farah and McClelland's (1991) findings with our model, using methods that closely approximated theirs. Farah and McClelland removed only visual or only functional semantic units, depending on the trial, based on the assumption that focal brain damage could preferentially affect either type of semantic information across subjects. In order to capture the effects of a distinct neuropathological episode as in a CVA or in herpes encephalitis, damage was introduced on a one-shot, rather than cumulative, basis.

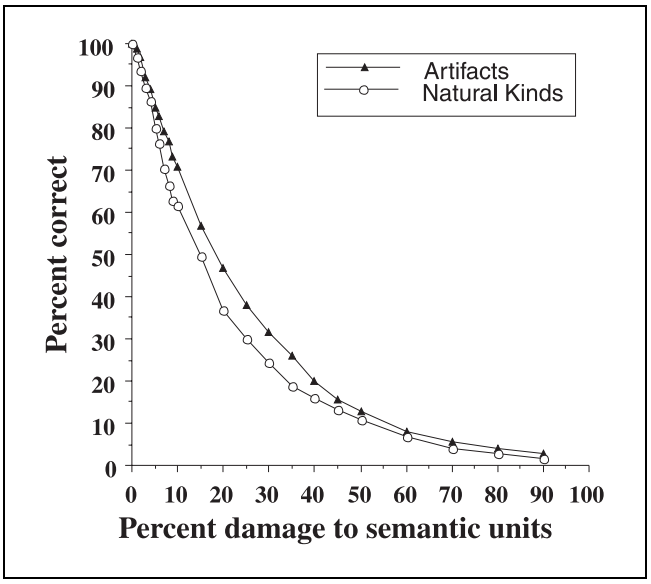

Figure 8. The model's performance over 50 trials when Alzheimer's disease is modeled as progressive loss of semantic features. The data show a natural kinds deficit for low levels of damage that becomes a more global deficit as the damage increases in severity.

Testing

Twenty types of simulations were run corresponding to 10, 20, 30, 40, 50, 60, 70, 80, 90, and 99% unit loss of either perceptual or functional semantic units. On each trial the perceptual or functional units of an intact model were subjected to either a 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, or a 0.99 chance of being removed. The same criteria from the previous experiments were used to evaluate the model's performance. Each condition was repeated 50 times.

Results and Discussion

Table 3 shows the results from these simulations. When damage was applied exclusively to perceptual units, the model displayed a selective deficit for natural kinds. Conversely, damage to functional units yielded an artifacts impairment. In both cases, as the level of damage increased, so did the severity of the impairment. These results clearly replicate the basic findings of Farah and McClelland (1991).

GENERAL DISCUSSION

In this paper we have used a connectionist model to demonstrate how category-specific deficits resulting from two types of brain injury can arise in a single semantic system without explicit category representations. Our findings indicate that the following four properties of the semantic system are relevant to category specific deficits: (1) perceptual and functional fea-

Table 3. Performance of the Model When Either Perceptual or Functional Semantic Units Are Removed

Amount of damage[a]	*Probability correct for artifacts*	*Probability correct for natural kinds*
Damage to perceptual features		
0	1.00	1.00
10	0.80	0.71
20	0.67	0.50
30	0.56	0.34
40	0.42	0.24
50	0.33	0.18
60	0.28	0.11
70	0.20	0.07
80	0.14	0.06
90	0.10	0.04
99	0.04	0.03
Damage to functional features		
0	1.00	1.00
10	0.87	0.88
20	0.73	0.76
30	0.58	0.67
40	0.47	0.56
50	0.38	0.53
60	0.32	0.46
70	0.23	0.37
80	0.16	0.36
90	0.13	0.34
99	0.07	0.30

[a] Amount of damage refers to the percentage of perceptual or functional semantic units destroyed.

tures, which are more important to natural kinds and artifacts, respectively, (2) topographic organization of the two types of features, (3) variation in the informativeness of features, and (4) more intercorrelations among features in natural kinds than artifacts. The first two properties have been used by Warrington and Shallice (1984), Farah and McClelland (1991), and others to explain how focal brain damage, such as herpes encephalitis or cerebrovascular accident, can preferentially impair perceptual or functional semantic information, yielding a natural kinds or artifacts deficit, respectively. This work extends those claims by making use of the third and fourth properties and demonstrates that in widespread, progressive damage, such as in Alzheimer's disease, the degree of damage interacts with the structure of the semantic system to yield a progression from an initial mild difficulty with artifacts to a definite natural kinds impairment. This account not only provides an integrated explanation of category-specific deficits in different types of brain damage but also helps to explain inconsistencies in reported behavioral data.

Garrard et al. (in press) have recently taken issue with our account, arguing that we "accept an anatomical account of category specificity in HSV encephalitis and MCA territory infarction, but reject it for DAT" and implying that our account is therefore unparsimonious. They suggest that the neuropathology of AD may be more focal than we have assumed, allowing all cases of category specificity to be explained in terms of focal damage. What is at issue here is a question of fact rather than parsimony. There is considerable debate about the pathogenesis of AD, as summarized in the "Introduction." Our theory is consistent with the majority of this evidence, which suggests that AD and HSE are associated with different types of neuropathology, as reflected in our simulations.

Our work is closely related to two other computational models of double dissociations, Farah and McClelland's (1991) model of category-specific semantic impairments and Plaut's (1995) model of selective impairments of either concrete or abstract nouns. In our work, as in theirs, words were represented as attractor states in a semantic space composed of perceptual and functional features, processing was highly interactive, and the behavior was an emergent property of a system not explicitly organized along category lines. The important difference between these models and ours lies in the way double dissociations were produced. In each of their models, damage localized to a single component of the semantic system yielded one-half of the double dissociation. In order to produce the other half, a different component was damaged in a separate simulation. The double dissociation relied on a comparison between these simulations. In our model, however, the double dissociation relied on the progression of damage, which affected the entire semantic system. Thus within a single simulation, both sides of the double dissociation are observed at different points in time.

Aside from explaining detailed aspects of behavioral data, our model makes a number of testable predictions. The first arises from the single aberrant simulation in Experiment 1. Recall that by chance the initial damage in that simulation was specific enough to effectively remove the intercorrelated features underlying fruits and vegetables while leaving all other categories largely unaffected, in essence creating a "pure" category-specific deficit with very narrow scope. This outcome should only occur in categories with strong intercorrelations, which typically are natural kinds categories (though not natural kinds such as body parts) but may also include

some unusual artifact categories such as musical instruments or gemstones. Interestingly, there are four such pure deficits in the literature: two were specific to animals, leaving plants and artifacts spared (Hart & Gordon, 1992; Hillis & Caramazza, 1991), and two affected plants specifically with a relative sparing of animals and artifacts (Hart, Berndt, & Caramazza, 1985; Pietrini et al., 1988). There are no reported cases of a specific artifact category being selectively affected by brain damage. Such a case could be accommodated within the current model only if it were one of the atypical artifact categories such as gemstones whose semantic features are highly intercorrelated.

A more obvious prediction relates to the behavioral progression that individual AD subjects should display. The model predicts that a majority of AD subjects will display an initial mild difficulty with artifacts over natural kinds that will eventually cross over into a natural kinds impairment as the degree of semantic impairment increases. It is not crucial to this claim that all subjects follow this hypothesized progression. Because of the variability in the disease process, not all subjects will begin at the same point or follow an identical pattern of deterioration. More specifically we would expect that (1) a large minority of cases might never display an initial artifacts deficit, (2) cases like NB with significant artifacts deficits would be rare, (3) preferential impairments to artifacts would only occur in subjects at the beginning of the progression, and (4) no subject would initially display a natural kinds deficit and then cross over into an artifacts deficit unless the initial deficit was specific to a single natural kinds category. To date our model is consistent with the existing behavioral literature, but these data are only suggestive. The modeling results make specific predictions about the interaction of degree of damage and the type of impairment that need to be evaluated in further, longitudinal AD studies.

Our model provides a detailed illustration of why it is important to examine data from individual patients. As Caramazza and his colleagues have argued (Caramazza, 1986; Caramazza & McCloskey, 1988), averaging subjects may obscure important aspects of the data. In Experiment 1, for example, averaging across simulations masked the fact that on a majority of the trials there was a small initial artifact deficit that developed into a natural kinds deficit, as well as the fact that the deterioration followed two distinct trajectories. On the other hand, it may also be necessary to examine large numbers of patients rather than relying solely on individual case studies in order to identify subtle effects, such as the initial artifact deficit, that may not be statistically reliable in any single case. Our model addresses both of these concerns: It explains the manifest variability in individual subjects while still capturing the group profile.

Although there is widespread agreement that it is necessary to examine individual patient data, there is less agreement about how to explain the variability that is observed. Differences in the patterns of impairment exhibited by patients are typically assumed to reflect different types of underlying deficits. Our modeling results, in common with the research described by Plaut (1995), strongly call into question the validity of this assumption (see also Seidenberg, 1988, and Farah, 1994). We have shown that a given type of etiology (e.g., loss of connections) can give rise to qualitatively different patterns of behavioral impairment. The fact that the different impairment progressions observed in Experiment 1 all derived from the same type of underlying deficit could not have been deduced from the behavioral data alone. These findings suggest that it is important not only to examine individual patient data but also the bases of different deficit patterns. In light of the simulations we have reported, it would be naive to assume that all differences in deficit patterns among patients necessarily reflect different types of underlying pathology. Computational models of the sort we have described provide a way to understand how probabilistic aspects of neuropathology interact to produce different behavioral deficits.

Acknowledgments

This work was supported in part by NIA Grants AG10109-01, 2T32AG00037, and AGO 5142–10 and by NIMH grants MH47566 and 01188. We thank Daniel Kempler and Victor Henderson for helpful discussions about the behavioral data and neurological implications, respectively.

Reprint requests should be sent to Mark S. Seidenberg, Neuroscience Program, University of Southern California, Los Angeles, CA 90089–2520, or via e-mail: marks@gizmo.usc.edu.

Notes

1. The word *category* has been used in this literature to refer to both narrower categories such as animals, tools, and vehicles and broader categories such as natural kinds or artifacts. We will use *category* in the former sense and *domain* in reference to the latter.
2. The cited authors have debated criteria for deciding between loss of a representation vs. loss of access to it, but these criteria do not capture the range of effects that are observed in connectionist networks. For example, damage to connections might result in a given feature never being activated even though its unit ("representation") remains intact.
3. All simulations were performed on an HP730 using the Xerion simulator developed by Tony Plate, Drew van Camp, and Geoff Hinton at the University of Toronto.
4. Our semantic representations are available upon request.
5. It is important to note that after damage more time was necessary for the model to settle into stable patterns. Whereas 10 time steps were enough to form stable patterns during training, the damaged model needed up to 38 time steps until activity settled. Consequently the phonological pattern of activation at time = 38 was used as the output vector.

REFERENCES

Amit, D. (1989). *Modeling brain function: The world of attractor neural networks.* Cambridge: Cambridge University Press.

Basso, A., Capitani, E., & Laiacona, M. (1988). Progressive language impairment without dementia: A case study with isolated category specific semantic deficit. *Journal of Neurology, Neurosurgery, and Psychiatry, 51,* 1201–1207.

Caramazza, A. (1986). On drawing inferences about the structure of normal cognitive systems from the analysis of patterns of impaired performance: The case for single-patient studies. *Brain and Cognition, 5,* 41–66.

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

Caramazza, A., & McCloskey, M. (1988). A case for single patient studies. *Cognitive Neuropsychology, 5,* 583–623.

Cohen, J., & Servan-Schreiber, D. (1992). Context, cortex, and dopamine: A connectionist approach to behavior and biology in schizophrenia. *Psychological Review, 99,* 45–77.

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

Damasio, A., & Van Hoesen, G. (1985). The limbic system and the localization of herpes simplex encephalitis. *Journal of Neurology, Neurosurgery, and Psychiatry, 48,* 297–301.

DeKosky, S., & Scheff, S. (1990). Synaptic loss in frontal cortex biopsies in Alzheimer's disease: Correlation with cognitive severity. *Annals of Neurology, 27,* 457–464.

DeLacoste, M.-C., & White, C. (1993). The role of cortical connectivity in Alzheimer's disease pathogenesis: A review and model system. *Neurobiology of Aging, 14,* 1–16.

Devlin, J. (in preparation). Doctoral dissertation, Department of Computer Science and Neuroscience Program, University of Southern California.

Farah, M. (1994). Neuropsychological inference with an interactive brain: A critique of the "locality" assumption. *Behavioral and Brain Sciences, 17,* 43–104.

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

Garrard, P., Patterson, K., Watson, P., & Hodges, J. (in press). Category-specific semantic loss in dementia of Alzheimer's type: Functional-anatomical correlations from cross-sectional analyses. *Brain.*

Gibson, P. (1983). EM study of the numbers of cortical synapses in the brains of ageing people and people with Alzheimer-type dementia. *Acta Neuropathologica, 62,* 127–133.

Gonnerman, L. M., Andersen, E. S., Devlin, J. T., Kempler, D., & Seidenberg, M. S. (1997). Double dissociation of semantic categories in Alzheimer's disease. *Brain and Language, 57,* 254–279.

Hart, J., Berndt, R., & Caramazza, A. (1985). Category specific naming deficit following cerebral infarction. *Nature, 316,* 439–440.

Hart, J., & Gordon, B. (1990). Delineation of single-word semantic comprehension deficits in aphasia, with anatomical correlation. *Annals of Neurology, 27,* 226–231.

Hart, J., & Gordon, B. (1992). Neural subsystems for object knowledge. *Nature, 359,* 60–64.

Hassoun, M. (1995). *Fundamentals of artificial neural networks.* Cambridge, MA: MIT Press.

Henderson, V., & Finch, C. (1989). The neurobiology of Alzheimer's disease. *Journal of Neurosurgery, 70,* 335–353.

Hillis, A., & Caramazza, A. (1991). Category specific naming and comprehension impairment: A double dissociation. *Brain and Language, 114,* 2081–2094.

Hinton, G. (1981). Implementing semantic networks in parallel hardware. In Hinton, G., & Anderson, J. (Eds.), *Parallel models of associative memory* (pp. 161–188). Hillsdale, NJ: Erlbaum.

Hinton, G., & Shallice, T. (1991). Lesioning an attractor network: Investigations of acquired dyslexia. *Psychological Review, 98,* 74–95.

Hopfield, J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Science, USA, 79,* 2554–2558.

Horn, D., Ruppin, E., Usher, M., & Herrmann, M. (1993). Neural-network modeling of memory deterioration in Alzheimer's disease. *Neural Computation, 5,* 736–749.

Hyman, B., Van Hoesen, G., Damasio, A., & Barnes, C. (1984). Alzheimer's disease: Cell-specific pathology isolates the hippocampal formation. *Science, 235,* 1168–1170.

Keil, F. (1987). Conceptual development and category structure. In U. Neisser (Ed.), *Concepts and conceptual development.* Cambridge, England: Cambridge University Press.

Keil, F. (1989). *Concepts, kinds, and cognitive development.* Cambridge, MA: MIT Press.

Malt, B., & Johnson, E. (1992). Do artifact concepts have cores? *Journal of Memory and Language, 31,* 195–217.

Malt, B., & Smith, E. (1984). Correlated properties in natural categories. *Journal of Verbal Learning and Verbal Behavior, 23,* 250–269.

Martin, A., Haxby, J., Lalonde, F., Wiggs, C., & Ungerleider, L. (1995). Discrete cortical regions associated with knowledge of color and knowledge of action. *Science, 270,* 102–105.

Martin, A., Wiggs, C., Ungerleider, L., & Haxby, J. (1996). Neural correlates of category-specific knowledge. *Nature, 379,* 649–652.

Masliah, E., Hansen, L., Mallory, M., & Terry, R. (1991). Immunoelectron microscopic study of synaptic pathology in Alzheimer's disease. *Acta Neuropathologica, 81,* 428–433.

Masliah, E., Terry, R., Mallory, M., Alford, M., & Hansen, L. (1990). Diffuse plaques do not accentuate synapse loss in Alzheimer's disease. *American Journal of Pathology, 137,* 1293–1297.

Mazzoni, M., Moretti. P., Lucchini, C., Vista, M., & Muratorio, A. (1991). Category-specific semantic disorders in Alzheimer's disease. *Nuova Rivista di Neurologia, 1,* 77–85.

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

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

Nebes, R. (1989). Semantic memory in Alzheimer's disease. *Psychological Bulletin, 106,* 377–394.

Patterson, K., Graham, N., & Hodges, J. (1994). The impact of semantic memory loss on phonological representations. *Journal of Cognitive Neuroscience, 6,* 57–69.

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

Patterson, K. E., Seidenberg, M. S., & McClelland, J. L. (1990). Connections and disconnections: Acquired dyslexia in a computational model of reading processes. In R. G. M. Morris (Ed.), *Parallel distributed processing: Implications for psychology and neuroscience* (pp. 131–181). London: Oxford University Press.

Pearson, R., Esiri, M., Hiorns, R., Wilcock, G., & Powell, T. (1985). Anatomical correlates in the distribution of the pathological changes in the neocortex in Alzheimer's disease. *Proceedings of the National Academy of Sciences, USA, 82,* 4531–4534.

Perani, D., Cappa, S., Bettinardi, V., Bressi, S., Gornotempini, M., Matarrese, M., & Fazio, F. (1995). Different neural systems for the recognition of animals and man-made tools. *Neuroreport, 6,* 1637–1641.

Pietrini, V., Nertempi, P., Vaglia, A., Revello, M., Pinna, V., & Ferro-Milone, F. (1988). Recovery from herpes simplex encephalitis: Selective impairment of specific semantic categories with neuroradiological correlation. *Journal of Neurology, Neurosurgery, and Psychiatry, 51,* 1284–1293.

Plaut, D. (1991). *Connectionist neuropsychology: The breakdown and recovery of behavior in lesioned attractor networks.* Unpublished doctoral dissertation. Carnegie-Mellon University.

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

Plaut, D., & Shallice, T. (1993). Deep dyslexia—a case study of connectionist neuropsychology. *Cognitive Neuropsychology, 10,* 377–500.

Price, J., Davis, P., Morris, J., & White, D. (1991). The distribution of tangles, plaques and related immunohistochemical markers in healthy aging and Alzheimer's disease. *Neurobiology of Aging, 12,* 295–312.

Rapp, B., & Caramazza, A. (1990). On the distinction between deficits of access and deficits of storage: A question of theory. *Cognitive Neuropsychology, 10,* 113–141.

Rogers, J., & Morrison, J. (1985). Quantitative morphology and regional and laminar distributions of senile plaques in Alzheimer's disease. *Journal of Neuroscience, 5,* 2801–2808.

Rosch, E. (1975). Cognitive representations of semantic categories. *Journal of Experimental Psychology: General, 104,* 192–233.

Rosch, E., & Mervis, C. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology, 7,* 573–605.

Rumelhart, D., & McClelland, J. (1986). *Parallel distributed processing: Explorations in the microstructure of cognition.* Cambridge, MA: MIT Press.

Ruppin, E., & Reggia, J. (1995). A neural model of memory impairment in diffuse cerebral atrophy. *British Journal of Psychiatry, 166,* 19–28.

Saffran, E., & Schwartz, M. (1994). Of cabbages and things: Semantic memory from a neuropsychological perspective—A tutorial review. In C. Umilta & M. Moscovitch (Eds.), *Attention and performance* XV: *Conscious and nonconscious information processing* (pp. 507–536). Cambridge, MA: MIT Press.

Sartori, G., & Job, R. (1988). The oyster with four legs: A neuropsychological study on the interaction of visual and semantic information. *Cognitive Neuropsychology, 5,* 103–132.

Scheff, S., DeKosky, S., & Price, D. (1990). Quantitative assessment of cortical synaptic density in Alzheimer's disease. *Neurobiology of Aging, 11,* 29–37.

Seidenberg, M. S. (1988). Cognitive neuropsychology and language: The state of the art. *Cognitive Neuropsychology, 5,* 403–426.

Shallice, T. (1988). *From neuropsychology to mental structure.* Cambridge: Cambridge University Press.

Shallice, T. (1993). Multiple semantics: Whose confusions? *Cognitive Neuropsychology, 10,* 251–261.

Silveri, M.-C., Daniele, A., Giustolisi, L., & Gainotti, G. (1991). Dissociation between knowledge of living and nonliving things in dementia of the Alzheimer type. *Neurology, 41,* 545–546.

Silveri, M.-C., & Gainotti, G. (1988). Interaction between vision and language in category-specific semantic impairment. *Cognitive Neuropsychology, 5,* 677–709.

Small, S., Hart, J., Gordon, B., & Hollard, A. (1993). Performance variability in a diffusely lesioned model of semantic representation for object naming. *Neurology, 43,* 404.

Smith, E., Shoben, E., & Rips, L. (1974). Structure and process in semantic memory: A featural model for semantic decision. *Psychological Review, 81,* 214–241.

Terry, R., Masliah, E., Salmon, D., Butters, N., DeTeresa, R., Hill, R., Hansen, L., & Katzman, R. (1991). Physical basis of cognitive alterations in Alzheimer's disease: Synaptic loss is the major correlate of cognitive impairment. *Annals of Neurology, 30,* 572–580.

Tversky, A. (1977). Features of similarity. *Psychological Revieu, 84,* 327–352.

Warrington, E., & McCarthy, R. (1983). Category specific access dysphasia. *Brain, 106,* 859–878.

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

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

Yankner, B., & Mesulam, M. (1991). Seminars in Medicine of the Beth Israel Hospital, Boston. Beta-amyloid and the pathogenesis of Alzheimer's disease. *New England Journal of Medicine, 325,* 1849–1857.