Connectionist Perspectives on Category-Specific Deficits

Timothy T. Rogers
Department of Psychology and the
Center for the Neural Basis of Cognition
Carnegie Mellon University

David C. Plaut
Departments of Psychology and Computer Science
and the Center for the Neural Basis of Cognition
Carnegie Mellon University

To appear in E. Forde and G. Humphreys, *Category-Specificity in Brain and Mind*. Psychology Press.

Theories of semantic memory tend between two poles. At one extreme is a view often associated with the connectionist enterprise: that the semantic system is a unitary, homogeneous mass, without functional or neuroanatomic specialisation, that capitalises on statistical regularities in the environment in learning about and processing semantic information. On this account, double dissociations of semantic memory are explained in terms of the processing mechanisms characteristic of neural networks, the statistical structure of the environment, and various psycholinguistic factors such as familiarity, frequency, and visual complexity. At the other extreme is a view positing that semantic memory is parcelled functionally and neuroanatomically into a set of discrete processing modules, each tied to a particular modality and/or semantic domain. Under this hypothesis, double dissociations of semantic memory arise from damage to one or another of these modules, or the connections between them.

In this chapter, we will argue that neither extreme position is likely to prove satisfying. Theories that eschew any form of neuroanatomic specialisation are unlikely to capture the variety of dissociations reported in the literature, while extreme modular views lack explanatory power. Accordingly, most computational models of semantic memory place themselves somewhere in the middle by adopting at least some form of neuroanatomic specialisation.

A number of interesting questions arise from this stance. What kind of specialisation exists? How and why does it occur? How is it related to the structure of the environment and mechanisms of learning in the brain? Here we find a much greater degree of variability across theories and models. Some assume only the grossest forms of neuroanatomic specialisation, whereas others posit many fine-grained distinctions. Some theories suggest that hard boundaries exist between anatomic regions with specific functions, while others adopt more graded forms of specialisation.

Though connectionist models are sometimes caricatured as homogeneous blobs without form or specialised function, in practice they offer a useful means of exploring the space between these opposing views. All connectionist models incorporate some degree of built-in architectural specialisation, in their organisation into groups of units, their patterns of connectivity, their unit parameters and learning rules. Most also adapt to the statistical structure of their virtual environments, acquiring through experience the ability to perform model analogues of cognitive tasks. Thus the theorist is at liberty to build as much or as little "anatomical" structure into a model's architecture as necessary. Explorations of the computational properties of such systems can then clarify the extent to which such assumptions are warranted by the data.

Throughout this chapter, we will focus on aspects of the connectionist approach that render these models well-suited to addressing neuropsychological data. Unlike more traditional box-and-arrow-drawings, connectionist models allow the theorist to specify in explicit, computational terms the internal structure of representations in different areas of the system (Allport, 1985). Computer simulations have shown that representational structure has important consequences for the behaviour of neural systems under damage, in a variety of domains. In some cases, these effects can lead to apparent double dissociations even in a homogeneous network with no assumed neuroanatomic specialisation (Bullinaria & Chater, 1995; Mayall & Humphreys, 1996; Devlin, Gonnerman, Andersen, & Seidenberg, 1998; Moss, Tyler, Durrant-Peatfield, & Bunn, 1998). In others, double-dissociations may arise from damage to anatomically distinct areas that are in no way specialised to subserve the cognitive functions dissociated (Plaut & Shallice, 1993; Plaut, 1995). Such findings call into question the conclusion frequently drawn from case studies forming double dissociations, that the dissociated functions must be subserved by modules that may be damaged independently. An investigation of the computational properties of connectionist networks in pathology can help the theorist to understand when these conclusions are justified, and under what conditions a double dissociation might be observed in a homogeneous system, or as a result of damage to anatomically distinct ar-

The research was supported by an NIMH FIRST award (MH55628) to the second author and by NIMH Program Project Grant MH47566 (J. McClelland, PI). Correspondence regarding this article may be sent either to Tim Rogers (trogers@cnbc.cmu.edu) or to David Plaut (plaut@cmu.edu), Mellon Institute 115–CNBC, Carnegie Mellon University, 4400 Fifth Avenue, Pittsburgh PA 15213–2683.

eas that nevertheless do not constitute independent neural modules specialised to subserve the dissociated cognitive faculties.

To illustrate how these properties of connectionist nets can help us to understand patterns of impairment in the domain of semantic cognition, we will begin by describing an influential model of category specific deficits put forward by Farah and McClelland (1991). The model is a simple implementation of the *sensory-functional* (SF) hypothesis: that semantic information about the perceptual and functional properties of objects are stored in anatomically distinct areas of cortex; and that apparent category-specific deficits arise because different semantic domains rely to a greater or lesser degree on sensory or functional information in their representation (Warrington & Shallice, 1984). The Farah-McClelland (FM) model builds in what is essentially an anatomical segregation between areas that represent sensory and functional semantic information. It also incorporates learning mechanisms that serve to associate these semantic features with one another, and with more peripheral visual and verbal representations. The marriage of given anatomical specialisation with domain-general learning mechanisms permits the model to account for a broad variety of data, and probably constitutes the best theoretical account of category-specific deficits to date.

Despite its considerable appeal, the FM model has recently come under criticism from some quarters, for a couple of reasons. First, developments in the case literature suggest that the model as formulated cannot capture the full range of patient data. Second, efforts to measure empirically the degree to which various categories rely on "sensory" and "functional" information in their definitions have lead some to question the validity of such a distinction. In section 2, we will discuss some alternatives to the sensory-functional hypothesis that have been put forward recently, focusing in particular on two opposing hypotheses. The *domain-specific knowledge* hypothesis postulates that knowledge of different semantic domains is subserved by anatomically distinct cortical modules, dedicated at birth, that have developed over ontogeny under evolutionary pressures (Caramazza & Shelton, 1998). We argue that the domain specific knowledge hypothesis raises far more questions than it answers, and in fact offers no leverage on the problems encountered by the FM model. In contrast, *unitary semantics* hypotheses posit that some category-specific deficits may be explained without any reference to the anatomical organisation of cortex, but rather in terms of learned sensitivity of the system to the statistical properties of the environment (Moss et al., 1998; Devlin et al., 1998; McRae, Sa, & Seidenberg, 1997; Hillis, Rapp, Romani, & Caramazza, 1990; Hillis, Rapp, & Caramazza, 1995; Tippett, McAuliffe, & Farah, 1995).

We will argue that each of these extreme positions has problems that undermine its adequacy. The thesis that semantic memory is subserved by multiple, anatomically distinct cortical modules offers no leverage on the empirical challenges faced by the sensory-functional hypothesis, and lacks explanatory power. However, the antithesis to this view, expressed in the unitary-semantics hypothesis, is unlikely to explain the full range of patient deficits reported in the literature.

Where, then, is an appropriate synthesis? In section 3 we consider an approach that has been successful in accounting for data in other domains of semantic cognition—namely, semantic dementia—and which we believe holds promise for understanding category-specific impairments as well (Rogers, Lambon-Ralph, Patterson, McClelland, & Hodges, 1999). Like the FM model, this theory builds in anatomical differences already known to exist in the brain. However, differential performance in living and nonliving domains is understood with reference to the similarity structure of representations in different surface modalities. We will discuss how the same properties that lead to structured deterioration in semantic dementia might be extended to account for seeming category-specific deficits as well, in this framework.

The Farah-McClelland model

The origin of recent interest in category-specific semantic impairments is usually attributed to two papers published by Warrington and her collaborators in the mid eighties (Warrington & McCarthy, 1983; Warrington & Shallice, 1984). The first of these was a case study of VER, a patient with extensive left hemisphere damage due to stroke, who presented with a severe global dysphasia. While VER was unable to produce even simple propositional statements, Warrington and McCarthy were able to show that she had some spared verbal comprehension, through the use of a word-to-picture matching task. This sparing appeared to strongly favour particular semantic categories: VER's performance was much better for flowers, animals, and foods than for man-made objects. Because man-made objects are generally much more familiar than, for example, flowers and animals (Warrington & McCarthy, 1983), the dissociation could not be explained on the basis of familiarity alone. Thus Warrington and McCarthy speculated that VER showed an impairment of semantic knowledge about nonliving things.

This conclusion was reinforced by a second study of four patients recovering from herpes encephalitis, who appeared to have the reverse dissociation: greater difficulty identifying living relative to nonliving things (Warrington & Shallice, 1984). All four exhibited worse performance for living relative to nonliving things in a match-to-sample task like the one used with VER. For the two patients with sufficiently spared expressive speech (JBR and SBY), the asymmetry was also apparent for picture naming, description, definition, and word-to-picture matching tasks. In these two cases, the difference was quite dramatic. For example, patient JBR correctly identified only 6 % of the living things on which he was tested, but was able to identify 90 % of the nonliving

objects. The difference between living and nonliving objects persisted strongly when stimuli were controlled for familiarity and word frequency.

Warrington and Shallice (1984) further tested JBR on his ability to identify 12 items from each of 26 categories selected from the Battig and Montague category norms (Battig & Montague, 1969). JBR identified significantly fewer items than expected on the basis of item frequency alone for 12 of the 26 categories; of these, 7 were either living things or foods. By contrast, of the 14 categories for which JBR was at or above expected performance, only 2 could be considered categories of living things (animals and body parts).

Together, these early papers formed the two sides of a double dissociation of semantic knowledge for living and nonliving things. Since their publication, many reports of selective deficits for knowledge of living things have appeared in the literature (Saffran & Schwartz, 1994; De Renzi & Lucchelli, 1994; Hart & Gordon, 1992; Farah, McMullen, & Meyer, 1991; Silveri & Gainotti, 1988). A smaller but still considerable number of reports of selective impairment for nonliving things have also been published (Behrmann & Lieberthal, 1989; Hillis et al., 1990; Sacchett & Humphreys, 1992). There have been some efforts to explain such findings with an appeal to uncontrolled stimulus factors such as visual complexity, word frequency, and familiarity. For example, Funnell and Sheridan (1992) showed that performance on naming tasks can vary dramatically with such factors as familiarity and frequency; and Stewart, Parkin, and Hunkin (1992) reported a patient with an impairment for naming living relative to nonliving objects, which disappeared when nuisance factors were controlled. However, several cases are now on record describing patients who continue to show unequal performance for living and nonliving things, even when stimuli have been carefully controlled, or when the effects of confounding variables have been regressed out (Kurbat & Farah, 1998; Kurbat, 1997; Farah et al., 1991; Hillis & Caramazza, 1991).

For both empirical and theoretical reasons, Warrington and Shallice (1984) did not interpret their results as implying that different cortical areas are responsible for representing knowledge about semantically distinct domains. First, the pattern of spared and impaired categories across their patients did not respect rigid semantic boundaries. While JBR was generally worse at naming living things, he also showed impaired performance for several categories of nonliving objects: metals, types of cloth, musical instruments, and precious stones. Also, his ability to name body parts (arguably living things) was relatively intact. The mirror-image of this pattern—impaired knowledge of nonliving things in general, with the conjoint sparing of metals, cloth, musical instruments, and precious stones—was later reported by Warrington and McCarthy (1987) in patient YOT. Thus, the data were not consistent with the hypothesis that knowledge of living things is subserved by one cortical region, and knowledge of nonliving things is subserved by another.

Instead, Warrington and Shallice (1984) proposed an anatomical division of labour along an independently motivated anatomical division: perception and action. They suggested that living things are primarily differentiated on the basis of their perceptual properties, whereas artifacts are more often differentiated on the basis of their function. If knowledge of functional and perceptual attributes are stored in anatomically distinct areas, damage to one region might result in differential impairments for knowledge of living relative to nonliving things, or vice versa. This theory provided an elegant explanation for the conjoint disturbance of (for example) musical instruments and living things, under the assumption that musical instruments, like living things, are differentiated primarily on the basis of perceptual properties. Similarly, one might expect that body parts are distinguished not on the basis of what they look like, but by their functional properties; hence the reason body-part and artifact knowledge were spared together in JBR, and impaired together in YOT. Caramazza and Shelton (1998) have recently termed Warrington and Shallice's theory the *sensory-functional* hypothesis.

Overview of the Model

An influential computational implementation of the sensory-functional hypothesis was put forward by Farah and McClelland (1991). In addition to demonstrating that the theory was indeed tractable, computer simulations with the model showed that it also had some counterintuitive implications. The model is illustrated in Figure 1. Each layer consists of an assembly of simple neuron-like processing units, connected as shown, whose activity may range between ± 1. The units are linked to one another by means of weighted connections, which can take any positive or negative value, and which determine the extent to which one unit's activity can influence another's. Associated with each pair of connected units are two such connections: one which permits activation to flow from the first unit to the second, and one which permits activation to flow in the other direction. Because activation may propagate in either direction between any pair of connected units, the network is said to be *recurrent*.

Representations of objects in the model take the form of distributed patterns of activity across groups of units. The units themselves 1can be thought of as each responding to some aspect of the entity represented by the whole pattern, though these aspects need not be nameable features or correspond in any simple way to intuitions about the featural decomposition of the concept. In the *semantic* layers, some units may respond to objects with some particular visual property, while others may respond to aspects of the object's functional role. In the *visual* layer, patterns of activity correspond to more peripheral visual representations; while patterns of activity in the *verbal* layer form representations of words.

The presentation of a stimulus to the model causes

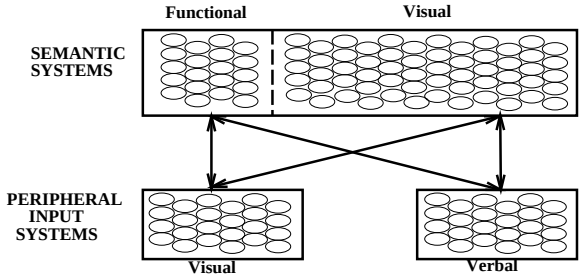

Figure 1. The architecture of the Farah and McClelland model. Redrawn with alterations from Farah and McClelland, Figure 1, p. 343. Permission pending.

an initial pattern of activity across units in one of the peripheral layers, with some of the units activated and some not. During processing, each unit's state is determined by calculating the sum of its inputs weighted by the strength of the connection between units. This sum is passed through a nonlinear squashing function bounded at -1 and 1, and the unit's activation state is then updated. Thus, a unit's activation at any point in time is determined jointly by the activation states of the other units in the network to which it is connected, and the magnitude of the weights between them.

The initial stimulus pattern presented to the network begins to change as each unit receives input from the other connected units. This dynamic flow of activation proceeds until the unit states stop changing, at which point the network is said to have settled into a steady state or *attractor*. The location of such stable configurations depends upon the connection weight matrix. The role of learning in this model is to configure the weights in such a way that, when the network is presented with a particular word or picture as input, it will settle into a stable state in which the correct pattern of activity is observed across units in the visual, verbal, and semantic layers.

Farah and McClelland created representations for ten "living" and ten "nonliving" objects, by generating random patterns of -1 and +1 across all three layers of units in the model. Each unique pattern corresponded to a representation of an individual item. Representations of living and nonliving things differed only the the proportion of active semantic units in the *functional* and *perceptual* pools. These were set to match the observed ratio of perceptual to functional features of objects in dictionary definitions (see below). Living things in the model were represented with an average of 16.1 visual and 2.1 functional units active; whereas nonliving things were represented with an average of 9.4 visual and 6.7 functional units active. All patterns had some units active in both semantic pools. The *verbal* and *visual* representations were random patterns generated in the same way for living and nonliving items.

To find a configuration of weights that would allow the network to perform correctly, the authors used an error-correcting learning algorithm called the *delta rule* (McClelland & Rumelhart, 1985; Rumelhart, Hinton, & McClelland, 1986). On each trial, an item was selected at random, and either its verbal or its visual representation was presented to the model. The network was allowed to settle for a fixed period of time, at which point the actual unit states were compared to the desired states. This discrepancy is referred to as the model's *error*. Under the delta rule, error is calculated for all units in the model. The weights received by each unit in turn are adjusted by a small amount to reduce that unit's error. After several iterations, the discrepancy between observed and desired states across all units is virtually eliminated, and the trained model generates the correct semantic, verbal, and visual patterns when presented with either a word or a picture as input.

Of interest was the model's behaviour when its semantic units were damaged. Under the sensory-functional hypothesis, units representing the functional-semantic aspects of an item can be damaged independently of the units representing the item's perceptual-semantic properties. How does the model's performance deteriorate with increasing damage to each of these pools of units?

To simulate neural trauma in the network, Farah and McClelland simply deleted some proportion of the units in either the perceptual semantic pool or the functional semantic pool. They then tested the network's ability to perform model analogues of picture naming and match-to-sample tasks. In the former, the model was presented with the picture of an object (by applying a pattern of activity to the visual units), and allowed to settle to a steady state. The resulting pattern of activity across the word units could then be read off, and compared to all the patterns in the training corpus. The model's response was considered correct if the pattern of activity across word units more similar to the correct pattern than to any other pattern. The same procedure was employed in the match-to-sample task, using a word as input and examining patterns of activity across visual units to determine the response.

Two aspects of their results are of interest. First, the model showed a clear double dissociation in its ability to name living and nonliving things. When visual semantic units were destroyed, the model exhibited a greater naming impairment for living relative to nonliving objects. The opposite was true when functional units were destroyed. Second, and more interesting, in neither case was the model completely unimpaired in the "spared" domain. Though the model was worse at naming living things when perceptual semantic features are destroyed, it was also impaired at naming nonliving things. Living things rely more heavily on perceptual semantic features in the model, but such features inform the representation of both living and nonliving objects to some degree. As this knowledge deteriorates in the model, it tends to affect naming performance for both domains, albeit to differing degrees. The same graded impairments are also witnessed in the patient data—profound impairments in one domain are almost without exception accompanied

by mild impairments in the relatively spared domain.

Farah and McClelland (1991) also examined the network's ability to retrieve functional and perceptual semantic information when given a picture or a word as input. Considering only the perceptual or the functional unit pools, they compared the pattern of activity in the damaged network when it had settled to the correct pattern, for each object. The network was considered to have spared knowledge of the perceptual properties of an item if the observed pattern of activity across perceptual semantic units was closest to the correct pattern; and spared knowledge of functional properties if the observed pattern across functional semantic units was closest to the correct pattern.

The simulations showed that the loss of semantic features in one modality had important consequences for the model's ability to retrieve properties in the spared modality. When perceptual semantic features were lost, the model had a tendency to generate an incorrect pattern of activity across functional semantic units, especially for living things. The reason is that the reciprocal connections among semantic features lead the network to rely on activity in perceptual semantic units to help produce the appropriate patterns across functional units. When this activation is reduced or disrupted as a result of damage, these lateral connections can interfere with the model's ability to find the correct states even in the spared units. Thus, the loss of perceptual semantic knowledge that occurs with trauma to the cortical areas subserving such knowledge can precipitate a disruption of semantic knowledge in the functional modality, especially for categories that rely to a large extent on perceptual information in their representation. Of course, the reverse is true when functional semantic features are damaged.

The FM model also accounts for a variety of related neuropsychological findings. For example, McCarthy and Warrington (1988) described a patient with a seeming category-specific deficit for living things, but only when tested verbally. Farah and McClelland (1991) explained this pattern of performance by positing a lesioning of the connections between verbal representations and visual semantic units. When the model was lesioned in a similar fashion, it was impaired at generating the correct semantic representations from a verbal input, especially for living things; but unimpaired at finding the correct semantic pattern from visual input. The opposite pattern—good performance when tested verbally, but poor performance on visual tests of category knowledge—has been observed in visual agnosics, who often show poorer performance when tested with pictures of living things (Dixon, 1999; Carbonnel, Charnallet, David, & Pellat, 1997). Riddoch and Humphreys (1987) and others (Arguin, Bub, & Dudek, 1996; Lecours, Arguin, Bub, Caille, & Fontaine, 1999; Dixon, Bub, & Arguin, 1997) have suggested that the data across such patients are best explained by supposing that the connections between visual and semantic representations have been disrupted. Although to our knowledge it has not been explicitly demonstrated, such a lesion in the Farah and McClelland model would be expected lead to greater impairment of living relative to nonliving things, for the reasons we have discussed.

Additional Empirical Issues

A number of other studies, many involving functional neuroimaging, have supported the general thesis that knowledge of functional and perceptual attributes are mediated by distinct brain regions. For example, Martin, Haxby, Lalonde, Wiggs, and Ungerleider (1995) used PET to show that discrete cortical regions were differentially active when subjects were shown a black-and-white drawing of an object, and required to name either its colour, or a characteristic action associated with it. The colour-naming condition produced increased activation in the ventral temporal lobes, whereas the action-naming condition produced enhanced activation of the left, posterior middle temporal gyrus. Mummery, Patterson, Hodges, and Price (1998) reported similar findings in a match-to-sample task in which subjects were required to select the object that matched the sample either in its colour, or in it's typical location. Matching on the basis of locale lead to increased activation of the posterior temporal-occipital-parietal junction, superior to the "action-naming" area identified by Martin et al. (1995). Matching on the basis of colour activated left ventral temporal cortex, just as did colour naming in the Martin et al. (1995) study.

These findings correspond well with converging evidence that the middle temporal gyrus is more strongly activated in semantic tasks involving tools, relative to those involving animals; and that the reverse is true for the posterior ventral temporal cortex (Chao, Haxby, & Martin, 1999; Moore & Price, 1999; Perani et al., 1999; Perani, Cappa, Bettinardi, & Bressi, 1995; Damasio, Grabowski, Tranel, & Hichwa, 1996). Chao et al. (1999) point out that the middle temporal gyrus is proximal to areas of cortex that process information about nonbiological motion (Zeki & al., 1991). Hence, they suggest that this area may code information about object use-associated motion, and consequently may be more important in representing artifacts than animals.

Outside the temporal cortex, a few studies have reported that prefrontal motor areas may be engaged in semantic tasks involving tools, relative to those involving animals (Spitzer et al., 1998; Spitzer, Kwong, Kennedy, & Rosen, 1995; Perani et al., 1995; Martin, Gagnon, Schwartz, Dell, & Saffran, 1996). Other studies have failed to find such a relationship (e.g. Moore & Price, 1999; Mummery et al., 1998); however, these imaging results are consistent with findings reported by Gainotti, Silveri, Daniele, and Giustolisi (1995), who surveyed the reported anatomical locus of damage in several patients with category specific deficits. These authors found that patients with impaired knowledge of living things were more likely to have damage to the medial temporal cor-

tex, proximal to areas that process colour information; whereas those with impaired knowledge of artifacts were more likely to have fronto-parietal lesions, proximal to motor planning areas. Thus, while there are a variety of studies that are generally consistent with the sensory-functional hypothesis, there remains much work to be done in this area.

Data from the domain of neuropsychology have provided mixed support for the sensory-functional hypothesis. The hypothesis would seem to predict that patients with impaired knowledge of living things should also show worse performance on tasks tapping their knowledge of the perceptual (relative to functional) attributes of living things. In the FM model, this prediction is borne out: when perceptual semantic units are damaged, the model has a harder time finding the correct pattern of activity across these units than across the functional semantic units, especially for the domain of living things. And, indeed, several studies have found similar results in patients (Bunn, Tyler, & Moss, submitted; Forde, Francis, Riddoch, Rumiati, & Humphreys, 1997; Powell & Davidoff, 1995; Basso, Capitani, & M., 1988; Sartori & Job, 1988; Silveri & Gainotti, 1988; De Renzi & Lucchelli, 1994; Gainotti & Silveri, 1996)

The interpretation of these results, however, is not straightforward. Caramazza and Shelton (1998) have argued that none of the studies described above employed adequate stimulus controls. In some cases, the items measuring functional knowledge were easier than the items assessing perceptual knowledge. In other studies, stimulus items were not controlled for familiarity, frequency, visual complexity, age of acquisition, and other psycholinguistic factors. These shortcomings have lead Caramazza and Shelton (1998) to make the strong claim that, in fact, there is no compelling evidence that perceptual knowledge and knowledge of living things are conjointly impaired in pathology. There are now several cases on record using strictly controlled stimulus materials, showing patients with apparent category-specific deficits whose knowledge of the perceptual and functional attributes of objects in the impaired domain is equally poor (Caramazza & Shelton, 1998; Samson, Pillon, & De Wilde, 1998; Lambon Ralph, Howard, Nightingale, & Ellis, 1998; Laiacona, Barbarotto, & Capitani, 1993).

Moreover, Lambon Ralph et al. (1998) have reported a semantic dementia patient (IW) with impaired knowledge of the perceptual properties of objects, but without a corresponding category-specific impairment. On tests of naming from definition, IW was fairly accurate for both living and nonliving things when the definitions included functional/associative properties, but was near chance for both domains when the definitions included only perceptual properties.

The sensory-functional hypothesis has attracted further criticism on the grounds that some of the patterns identified early on by Warrington and Shallice have not held up as the case literature has expanded. For example, though many studies have found inanimate but "sensory" categories, such as fabrics and musical instruments, to be spared or impaired along with categories of living things (Basso et al., 1988; Silveri & Gainotti, 1988; Sheridan & Humphreys, 1993; De Renzi & Lucchelli, 1994), it is no longer clear that such patterns are consistent. Other researchers claim to have found patients with semantic deficits restricted to more specific semantic categories, such as animals (Hillis & Caramazza, 1991), body parts (Shelton, Fouch, & Caramazza, 1998), and fruits and vegetables (Farah & Wallace, 1992; Hart, Berndt, & Caramazza, 1985). Again most of this work is difficult to interpret, as appropriate controls have rarely been performed; and too often, conclusions about a patient's "semantic" impairments are drawn solely from picture naming data, with no attempt to determine whether the observed deficits are specific to language (Caramazza & Shelton, 1998). Even so, at the very least, it seems safe to conclude that there exists considerable variability across patients in the particular categories of knowledge they retain.

Simplifications adopted by the FM model. These developments have generally been interpreted as problematic for the sensory-functional hypothesis. While they do appear to violate some predictions made by the FM model, it is not clear to us whether the model's shortcomings in this respect are due to theoretically interesting flaws, or to uninteresting simplifying assumptions made by Farah and McClelland in their implementation. There are three respects in which the FM model is too underspecified to yield clear predictions about the conjoint disruption of particular semantic categories on the one hand, and modalities of semantic information on the other.

First, Farah and McClelland used random patterns of activity to represent items in the visual, verbal, and semantic layers of the network. This is certainly an oversimplification of affairs, and the contribution of this choice to the network's behaviour should not be underestimated. It is now well known that the similarity structure of representations has a strong impact on the pattern of errors made by connectionist networks following damage (see, e.g., Plaut, McClelland, Seidenberg, & Patterson, 1996; Plaut & Shallice, 1993). Items that have similar internal representations are more likely to be confused with one another when the network is damaged. Structured representations can also lead the network's behaviour to be more robust under damage, and can result in nonlinear decrements in performance, depending on the extent to which categories of objects are comprised of bundles of mutually reinforcing features (Plaut, 1995; McRae et al., 1997; Moss et al., 1998). Such effects form the basis for several quite different accounts of category-specific deficits, which we will discuss in the next section. It is not clear how the patterns of deterioration observed in the FM model would change if such structure were incorporated into the network's representations.

Second, Farah and McClelland used a closest-match

paradigm to decide whether internal representation states were correct under damage. A more strict criterion for correct performance would presumably result in a greater number of errors, and it is possible that under these conditions, damage to perceptual units (for example) could lead the network to be equally impaired at retrieving visual and functional semantic information, primarily for categories of living things.

Finally, the model does not make explicit the relationship between patterns of activity across sensory and functional units, and the system's ability to make judgments about the presence or absence of particular semantic features in an attribute listing or verification task. The semantic units in the FM model are not meant to stand for explicit knowledge about the presence or absence of particular object attributes. Thus, there is no way to know, from the disturbed pattern of activity across semantic units, just what the model "knows" and what it doesn't.

There may be some confusion about this last point. Farah and McClelland did indeed consider explicit object properties, as they appear in dictionary definitions, in order to provide an empirical basis for selecting the ratio of sensory to functional features when constructing semantic representations of living and nonliving objects in the model. They had subjects read dictionary definitions of objects in various categories, and underline all the words describing either the object's sensory properties, or its functional properties. From these data, they calculated the average ratio of sensory to functional features for definitions of living and nonliving things, and employed these ratios in generating the random strings comprising the semantic representations used in the network. However, they state that the sensory and functional semantic units in the network are not meant to correspond to general intuitions about the featural decomposition of objects; that is, these units are not meant to stand for explicit object attributes. Instead, the authors assume that this measure proveds a valid indication of the extent to which various objects rely on sensory and functional information in their definitions.

Whether or not this assumption is warranted is another point of contention. Caramazza and Shelton (1998) take issue with the general claim that sensory information is more important for representations of living things, while functional information is more important for representations of nonliving things. To support their point, they replicated Farah and McClelland's dictionary study, with a slight variation in the instructions. Instead of underlining either sensory or functional properties of objects, subjects were told to underline words describing either sensory or *nonsensory* properties of objects. Under these conditions, the ratio of sensory to nonsensory properties did not differ significantly between categories of living and nonliving things.

Nevertheless, other studies designed to assess the attribute structure of objects in different categories have also found differences between categories of living and nonliving things, in their reliance on sensory or functional properties (McRae et al., 1997; Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976). Recently, Garrard, Lambon Ralph, Hodges, and Patterson (in pressa) had subjects list the properties of 64 items in living and nonliving domains, in response to prompting questions designed to extract both sensory and functional information. They classified each feature according to the following criteria: as *sensory* if they could be appreciated in some sensory modality (e.g. "an eagle is large" or "a saw is sharp"); as *functional* if they described an action, activity, or use of an item (e.g. "a cat can catch mice," or "an owl can fly,"); as *encyclopaedic* if they described some other associative relationship (e.g. "a tiger is found in India," "a toaster is kept in the kitchen"); and as *categorical* if they named a category to which the object belonged (e.g. "a dog is an animal"). Their findings confirmed the conclusions from both Farah and McClelland (1991) and Caramazza and Shelton (1998). Considering only the sensory and functional features, living things had a greater proportion of sensory features listed per item than did nonliving things. Pooling functional and encyclopaedic features together, this difference between domains disappeared, because subjects tended to list a greater number of encyclopaedic features per item in the domain of living things.

More relevant to the sensory-functional hypothesis, Garrard et al. (in pressa) found that living things tended to be differentiated primarily on the basis of their perceptual features, whereas artifacts were differentiated from one another primarily on the basis of their function. This result speaks more directly to Warrington and Shallice's initial proposal, that living things are differentiated on the basis of their appearance, whereas artifacts are distinguished primarily according to their use (Warrington & Shallice, 1984).

As yet there is little consensus regarding the best way to measure the attribute structure of categories, and their reliance on sensory or functional information. Clearly, though, the Farah and McClelland model is underspecified in this respect. The model assumes anatomic segregation between areas representing the functional and perceptual properties of objects, but does not make explicit how these representations support the recollection of particular sensory or functional properties, or indeed how and why such structure arises in the first place.

Summary.

Despite its shortcomings, the sensory-functional hypothesis and its incarnation in the FM model make sense of a broad variety of phenomena in neuropsychology. Whether or not a more detailed implementation of the sensory-functional hypothesis will be able to accommodate the anomalies in the patient data described above is an empirical question. In Section 3, we will discuss one approach that appears promising. First, though, we turn to two alternatives to the sensory-functional hypothesis that have recently been put forward, exemplifying polarised reactions to these developments.

Alternative Theories of Category-Specific Deficits

In the years since the publication of the FM model, there have been two major influences on theories of semantic memory and category-specific deficits. The first is the remarkable expansion of the relevant case literature. Since Warrington and McCarthy's (1983) work in the mid eighties, there have been at least 97 case studies of patients with purported category-specific deficits (Lambon Ralph, personal communication). As we have already intimated, this wealth of data has been both a help and a hindrance. Certainly the accumulation of information has reduced the need for theoretical speculation; but the considerable variability in the test items, methods, and controls adopted, have made it near impossible to compare results across studies. The field awaits a critical review and synthesis of this material, but in the mean time, there has been a preponderance of new ideas about how best to capture central tendencies in the group data on the one hand, and the range of effects across all patients on the other.

The second important development in the past decade has been an increasing appreciation of the counter-intuitive ways that complex systems, as embodied in connectionist networks, behave under pathology. Computer simulations are playing an increasingly important role in the explanation of a variety of cognitive phenomena, and these ideas have had a strong impact in the domain of semantic cognition—in some cases, leading theorists to dispense with traditional cognitive constructs all together.

In this section, we will consider two reactions to these opposing pressures. First, we discuss Caramazza and Shelton's (1998) thesis semantic categories constrain the functional neuroanatomy of the brain. Second, we address the antithesis, put forward in different forms by a variety of theorists, that semantic knowledge is subserved by a unitary and anatomically homogeneous neural system.

Thesis: The Domain-Specific Knowledge Hypothesis

One natural reaction to the literature on category-specific deficits is to suggest that knowledge of different semantic domains is mediated by different cortical systems, which may be damaged independently from one another. This is the stance taken recently by Caramazza and Shelton (1998), in response to the perceived incapacity of the sensory-functional hypothesis to account for variability in the case literature. They write (p. 9):

> The hypothesis we wish to entertain is that evolutionary pressures have resulted in specialised mechanisms for perceptually *and* conceptually distinguishing animate and inanimate kinds, leading to a *categorical* organisation of this knowledge in the brain. We will call this hypothesis the *domain-specific knowledge* hypothesis.

The arguments these authors marshal to support their thesis rest primarily on a critical analysis of the sensory-functional hypothesis, much of which we have already discussed. In their view, the empirical phenomena described in Section 2 are grounds for rejecting the sensory-functional framework all together. According to Caramazza and Shelton:

1. The SF hypothesis does not predict the occurrence of category-specific deficits with equal impairment of sensory and functional features, though there is good evidence that such patients exist.

2. The SF hypothesis predicts that impaired knowledge of the sensory or functional properties of objects should always be accompanied by a category-specific deficits, but there are counter-examples to this prediction.

3. The SF hypothesis suggests that patients with deficits for living things should have greater difficulty retrieving the sensory (relative to the functional) properties of objects, while the reverse should be true of patients with deficits for nonliving things. There is no adequately-controlled study documenting a patient with a category-specific impairment and the conjoint impairment of knowledge for the corresponding (sensory or functional) modality. Studies that claim to have shown such an association either did not control for the difficulty of the judgment, or for other confounding factors.

4. Under the SF hypothesis, patterns of spared and impaired categories across patients should be consistent. The patient record shows instead that this is not the case. There is no good evidence that "perceptually" defined categories such as cloth, minerals, and musical instruments, are jointly spared or impaired with knowledge of living things.

5. The SF hypothesis leaves no room for the impairment of categories more specific than the global categories *living* and *nonliving*, though such patients have now been reported. There exists good evidence that the categories *foods/plants*, *animals*, and *other things* may be selectively and independently impaired in neuropathology.

Of course, these conclusions depend upon one's reading of a complicated literature. But accepting for the moment this interpretation of the data, it behooves us to consider the ability of the domain-specific knowledge hypothesis to address the shortcomings of the FM model. It is not clear that we gain any leverage on these problems by invoking domain-specific semantic modules. According to Caramazza and Shelton (1998, p. 9), "[one] expectation derived from this hypothesis is that (everything else being equal) category-specific deficits should result in comparable impairments for the visual and functional attributes of a concept." As we have noted, a handful of patients have indeed presented with apparent category-specific deficits, and equal impairment of sensory and functional attributes. However, it is not clear how the domain-specific-knowledge hypoth-

esis might account for dissociations that do fall along sensory/functional lines—for example, patients with impaired knowledge for the sensory or functional attributes of objects, regardless of category, or patients who appear to have conjoint modality- and category-specific deficits. Caramazza and Shelton do not deny that such cases exist—indeed, their rejection of the SF hypothesis rests in part on the existence of these patients. But the theory offers no means of anticipating or interpreting these findings.

Also unspecified under the domain-specific-knowledge hypothesis is an explanation of why category-specific deficits so rarely confine themselves to even very broad semantic domains. As we have noted, patients with so-called category-specific deficits are almost never completely unimpaired in the relatively spared domain. Warrington and Shallice's conclusion that "perceptually" defined categories, such as minerals and musical instruments, are consistently spared or impaired along with categories of living things may have been premature; however, there is little doubt that patients with deficits for living things in general can also have great difficulty with particular categories of nonliving things. One of the strong points of the FM model is that it makes explicit how such graded impairments might arise.

These limitations are arguably to be expected from any immature theory. However, it is difficult to imagine how the theory might be developed without running into major difficulties. Consider Hart and Gordon's (1992) study of KR, a patient who presented with a severe anomia specific to animals. KR showed perfect performance verifying the functional and perceptual properties of artifacts, as well as the functional properties of animals, but had difficulty verifying the perceptual properties of animals. Furthermore, KR's impairment was only apparent when tested verbally. For example, she was able to discriminate appropriately from inappropriately coloured pictures of animals, but performed poorly on the same test when the colours were given verbally. Caramazza and Shelton conclude that KR shows an impairment for the sensory properties of animals specific to the verbal modality. It is not obvious how such a pattern would occur under the domain-specific knowledge hypothesis. It would not do to suggest that KR had an impairment specific to the "animal" area of the semantic system, for such an impairment would affect both the sensory and functional properties of animals in all testing modalities. Nor would it do to posit a lesion specific to language areas, for such damage should affect knowledge of the sensory and functional properties of both living and nonliving categories. Indeed, it seems the only way to explain this data in a manner consistent with the domain-specific knowledge hypothesis is to posit the existence of dedicated neural circuits, not only for the representation of animals, but also for the representation of the perceptual properties of animals as expressed verbally.

This train of thought leads us toward a model in which semantic knowledge is subserved by a large number of independent, highly specialised neural modules, each tied to a particular semantic domain, type of information, and modality (see, for example, Coltheart, Inglis, Michie, Bates, & Budd, 1998). The problems with such a model are obvious. As the number of supposed innately-specified modules increases, it becomes increasingly difficult to understand why some combinations of deficit are observed frequently (such as the conjoint impairment of animals and foods) while others are observed rarely or not at all (e.g. the conjoint impairment of artifacts and foods). It also becomes more difficult to accept that such modules have developed in response to evolutionary pressures—or at least, more difficult to construct, after the fact, an evolutionary rationale for the existence of such highly specialised modules.

Furthermore, the theory does not help us to understand other disturbances of semantic cognition. Another strength of the FM model is that it ties together data from patients with category-specific dysphasias and visual agnosias. The domain-specific-knowledge hypothesis does not explain why visual agnosics are often worse at identifying living things, even when the stimulus items are controlled for confounding visual and psycholinguistic factors (e.g., Dixon, 1999; Arguin, Bub, & Dudek, in press). Nor does it explain the occurrence of generalised semantic deficits which do not favour some categories over others, such as are observed in semantic dementia (Snowden, Goulding, & Neary, 1989; Hodges, Patterson, Oxbury, & Funnell, 1992; Hodges, Graham, & Patterson, 1995).

In addition to these empirical challenges, the theory seems to us to lack any real explanatory power. If, as Caramazza and Shelton (1998) claim, the data support the conclusion that categories of living things, foods, and artifacts are the only ones that may be selectively impaired, what does it add to propose that the brain must have evolved special modules for processing information about each of these domains? To their credit, they acknowledge that, "...unless we can independently motivate the assumption of categorical organisation of conceptual knowledge, we would have merely assumed what we are trying to explain—an infelicitous circularity" (p. 18).

We, of course, agree. In search of the desired motivation, the authors adopt an evolutionary perspective, arguing that there is a high fitness value to the evolutionary adaptations that would allow an organism to discriminate between predators and prey, and to identify food sources. Under this view, it makes sense that the only three innately specified semantic domains are those of foods, animals, and "other things." Tellingly, Shelton and Caramazza have already amended this claim to admit category-specific deficits for finer-grained but "evolutionarily motivated" categories such as body parts (Shelton et al., 1998); and have even suggested that impairments specific to "non-evolutionary" categories may

be observed as a consequence of mechanisms similar to those adopted by the SF hypothesis (Shelton & Caramazza, 1999). Thus, it is not clear exactly where the line the authors wish to draw should go.

However, taking the domain-specific knowledge hypothesis at face value, speculation about which semantic domains are innate is necessarily post-hoc; and it is difficult to seriously conclude that such activity provides independent motivation for the theory. Post-hoc evolutionary arguments have been made, with varying degrees of success, to support a host of differing claims about which semantic distinctions are innately given, and which are learned (Pinker, 1994, 1997; Carey & Spelke, 1994; Wellman & Gelman, 1997; Springer & Keil, 1991). In the absence of converging empirical evidence to support them, such claims amount to little more than restatements of the data.

Furthermore, there are good reasons why the categories of foods, animals, and artifacts might be special, quite apart from their fitness value. Statistical analyses of attribute-listing studies have shown that, simply on the basis of their propensity to share properties, objects cluster naturally into these global categories. Garrard et al. (in pressa) entered 64 items from an attribute-listing study into a hierarchical clustering algorithm, based on the vector of properties attributed to each object by subjects in the study. The algorithm divided the various objects into three broad clusters,corresponding to the categories of animals, foods, and artifacts. It is at least possible, then, that the statistical structure of the environment contributes to the differentiation of objects into global categories. We will return to this idea in Section 3.

In summary, the expansion of the case literature in recent years, and the accompanying increase in the variability of deficits reported across patients, have lead some theorists to reject the sensory-functional framework outright. As an alternative, Caramazza and Shelton (1998) propose that evolutionary pressures have lead to the development (across phylogeny) of separate neural modules dedicated to storing semantic information (both sensory and functional) about the categories of animals, foods, and artifacts. We have argued that such an account is likely to fail for several reasons. First, it appears equally vulnerable to the empirical criticisms levelled against the sensory-functional hypothesis. To explain the range of deficits reported in the literature, one must posit multiple independent neural modules, each restricted to storing a particular kind of information about a single semantic domain, in a single modality. However, if this view is correct, it is not clear why some combinations of deficit occur frequently, while others are observed rarely or not at all. Second, the theory does not account well for other disorders of semantic cognition, such as visual agnosia and semantic dementia. Third, it provides no explanation of the graded nature of category-specific deficits, or the sloppy boundaries between impaired and spared domains. Finally, the theory has little explanatory power, and seems to us to be little more than a redescription of the data.

Antithesis: Unitary Semantics Hypotheses

As the patient record has expanded, so has our understanding of the counterintuitive ways that complex neural systems may behave under damage. Connectionist models have demonstrated that the behaviour of such systems can vary dramatically depending upon the theorist's assumptions about how internal representations are structured. Traditional neuropsychological approaches often do not take into account representational structure (Allport, 1985), attributing cognitive deficits in neuropathology to the all-or-none impairment of broad areas of cortex, whose finer details remain unspecified. Connectionist models have allowed the theorist to explore the inner workings of such black boxes, and as a consequence, there has grown an increasing appreciation of the extent to which representational structure may contribute to the ordered breakdown of cognitive function, even in anatomically homogeneous systems. This progress has lead other researchers to explore the possibility that category-specific deficits may be explained without reference to the functional and anatomical specialisation of the cortex. Following Caramazza, Hillis, Rapp, and Romani (1990), we will refer to such approaches as *unitary semantics* hypotheses.

The central challenge of a unitary semantics hypothesis is to explain how double dissociations of semantic memory can occur without invoking any neuroanatomic specialisation in the semantic system. In other domains of cognition, such as word reading, it has been demonstrated that a network's ability to perform correctly under damage may differ for various stimuli, depending upon their propensity to participate in regular or systematic mappings. Items that conform to systematic input-output mappings, such as the "regular" items in an orthography-to-phonology translation, are often easier for a network to learn, and may be more robust to small perturbations in the weights. Items that do not adopt such regularities, such as "irregular" words, can be more difficult to learn, and more vulnerable to damage. Several researchers (Moss et al., 1998; Devlin et al., 1998; Tippett et al., 1995; McRae et al., 1997) have suggested that these properties of neural networks may give rise to category-specific deficits, under the assumption that the domains of living and nonliving things share differing degrees of structure.

To understand how such an account might work, consider a model of naming put forward by Devlin et al. (1998), illustrated in Figure 2. In this network, word sounds are represented by random patterns of activity across the layer labelled *phonology*, whereas semantic representations are reflected by patterns of activity across the *semantic* units. Each layer is recurrently connected to a hidden set of *clean-up* units: additional units that permit the network to form attractors corresponding to representations of an object's name and identity (Hinton & Shallice, 1991). The layers are also fully interconnected

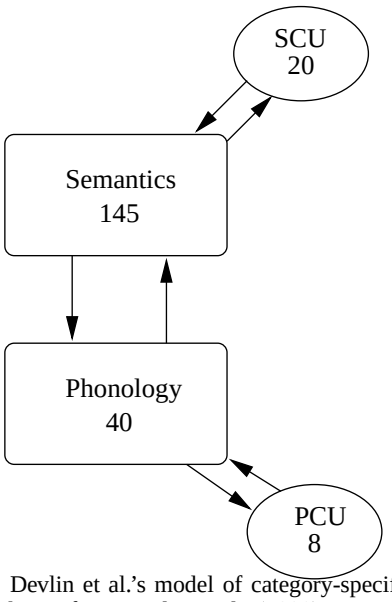

Figure 2. Devlin et al.'s model of category-specific deficits in DAT. Redrawn from Devlin et al. (1998), Figure 2, p. 82. Permission pending.

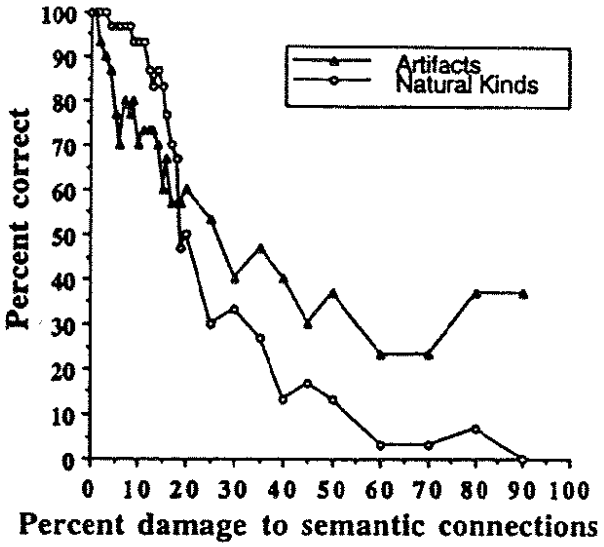

Figure 3. Simulation of picture naming in Devlin et al.'s model, under increasing amounts of damage. With small amounts of damage, the greater degree of structure among living things benefits the network's performance for these items, resulting in a modest category-specific deficit favouring living things. With greater amounts of damage, the network shows a 'critical mass' effect for living things: it's ability to name them degrades increasingly rapidly, so that in later stages the network shows a category-specific deficit favouring nonliving things. Reprinted from Devlin et al. (1998), Figure 6, p. 88. Permission pending.

with one another.

Devlin et al. (1998) constructed semantic and phonological representations for each of 60 objects, from various living and nonliving categories. Phonological patterns were simply random binary vectors across the 40 phonological units. Like the FM model, semantic representations were composed of units encoding either sensory or functional properties of objects. Unlike the FM model, however, Devlin et al. built varying degrees of similarity into their semantic representations, depending on each object's category. Exemplars of categories in the domain of living things were more likely to share features with one another, whereas nonliving things were more likely to be composed of idiosyncratic features. Also, living things had a higher proportion of correlated features (across items) in their representations than did nonliving things. Thus, the semantic representations for two kinds of bird were more similar to one another than to the representation of, for example *car*; and across living things, there was a higher incidence of features that consistently occurred together (such as *wings, feathers, beak*). [1]

The network was trained using the delta-rule, just as described for the FM model. In its trained state, the presentation of a semantic representation would lead the network to generate the appropriate pattern of activity across the phonology units; while the presentation of an item's name would lead the network to generate the correct identity representation across the semantic units. Of particular interest was the network's behaviour when connections to the semantic units (both sensory and functional) were damaged indiscriminately. Would the incorporation of structure into the semantic representations cause the network to show doubly dissociated category-specific deficits in naming with increasing amounts of damage?

Devlin et al. (1998) ran several simulations of damage, each time removing some proportion of weights in the network and stepping through the set of training stimuli to assess performance. The results of one such run are illustrated in Figure 3. With small amounts of damage, the network frequently had greater difficulty naming nonliving objects. However, as damage accumulated, the network showed a "critical mass" effect in the domain of living things. Its ability to name these objects declined sharply, so that with greater amounts of damage, the network showed the reverse dissociation: greater difficulty naming living relative to nonliving things. Devlin et al. attributed this result to the structural differences among semantic representations of living and nonliving things in their model. Groups of objects that share structure, in the form of overlapping, intercorrelated features, are robust to small amounts of damage, because the network can "fill in" missing features on the basis of its knowledge about how properties co-occur with one another, as

[1] These statistical properties of the training set were based on the results of an attribute listing task showing that there were a greater number of shared features the domain of living relative to nonliving things, and more correlations among features for categories of living relative to nonliving things.

encoded by the interconnecting weights. Because such bundles of mutually reinforcing features are assumed to occur more frequently for categories of living things, this property is reflected in a category-specific deficit favouring living things under small amounts of damage. However, as damage increases, shared structure becomes a liability. Once whole groups of mutually reinforcing features are lost, the network is unable to correctly fill in the gaps. In fact, it may make incorrect inferences on the basis of its remaining weights, and as a consequence, its ability to name degrades rapidly. By contrast, objects represented primarily by idiosyncratic properties, which tend to be nonliving things, are relatively immune to these forces. Hence, the ability to name such objects degrades more linearly with increasing damage.

Closely related accounts have been put forward by a number of other theorists (Moss et al., 1998; McRae et al., 1997; Tyler & Moss, 1998). Though these hypotheses share the same general form, they vary in the degree and kind of structure assumed to exist across categories. For example, Tyler and Moss (1997) suggest that there exist differences between animals and artifacts in their form-function correlations. Perceptual properties that are shared in the animal domain (e.g. eyes, legs) also tend to be correlated with functional attributes (e.g. seeing, walking). In the artifact domain, the idiosyncratic visual features tend to be correlated with functional attributes (e.g. *has a blade* co-occurs with *can cut*). As a consequence, their model makes exactly the opposite prediction about how semantic knowledge should degrade with generalised damage: under small amounts of deterioration, living things ought to be impaired relative to artifacts, while the reverse should be true for extensive lesions (Moss et al., 1998) .

We have focused on the Devlin et al. (1998) model, because it offers the opportunity to discuss the limitations as well as the strengths of this approach. We have two principal reservations regarding Devlin and colleagues' work. The first stems from the manner in which the network's behaviour under damage was assessed, and calls into question the authors' interpretation of their model's impaired performance. The second applies more generally to the capacity of unitary-semantics hypotheses to account for the range of patient data.

In Figure 3, we showed the results from one among 50 simulations of damage in Devlin et al.'s study. On this particular trial, the network showed the desired pattern of performance. On 12 of the 50 trials, however, the network did not show this behaviour. In 11 trials, it exhibited worse performance for living things throughout the progression of deterioration; and in 1 trial, it actually showed the reverse pattern from that expected: better performance on nonliving things with small amounts of damage, and better performance on living things with large amounts of damage. Thus, under damage, the network displayed a range of behaviours, only some of which support the conclusions drawn by the authors. Put another way, had Devlin and colleagues elected to focus on a different set of damage trials, they may have drawn quite different conclusions. Because the network's pattern of performance varies from trial to trial, the interpretation of its behaviour is subject to confirmation bias: it is easy to draw attention only to those trials that support one's hypothesis, and to explain away those that do not. The problem is exacerbated by the fact that the model's propensity to show a particular distribution of behaviours under damage can vary depending upon parameter choices not constrained by the theory, such as the number of training patterns, the number of hidden units, and the number of features per item in the training corpus (Perry, 1999). Thus, though the model showed the predicted pattern of behaviour on the majority of the damage trials reported by Devlin et al., it may not have done so under a different choice of model parameters.

Of course, this dilemma is not specific to the Devlin et al. (1998) model, but is endemic to any connectionist model of a neuropsychological syndrome wherein damage is administered to the network at random. Under this circumstance, the model's behaviour will always vary from one instance of damage to the next. If the model only occasionally shows the effect of interest, to what extent does it provide an adequate explanation of the phenomenon?

One response to the situation is to point out that, if the model's behaviour varies as a function of damage, so do the patients'. Though difficult to prove empirically, it is conceivable that two patients with lesions of comparable magnitude to the same cortical areas may nevertheless exhibit markedly different cognitive deficits. Perhaps, then, the variability of the model's behaviour across different instances of damage is more of an asset than a liability, since it allows one to model the distribution of possible deficits across a set of patients with qualitatively similar lesions. Under this view, a single instance of damage to the model is analogous to a single case study in the literature. Thus, to explain a particular patient's data, one need simply demonstrate that the model is occasionally capable of showing the predicted pattern of breakdown—that is, that the case to be explained falls somewhere within the range of behaviours produced by the network across different instances of damage. This is the strategy adopted by Devlin et al. (1998), who write, "the variability in the effects of damage on performance in the modelling results is consistent with the variability observed among AD subjects, and helps to explain some seeming inconsistencies in the behavioural literature" (p. 87; also see Mayall and Humphreys (1996), and Joanisse and Seidenberg (1999)).

However, this approach is problematic. Connectionist networks are, of necessity, orders of magnitude smaller than the actual systems they are intended to model. Consequently, random damage to a given anatomical region is likely to yield much more variable behaviour from trial to trial in the model than in the actual system. If the model's behaviour under damage is more variable than the actual system's, it is more likely to occasionally ex-

hibit patterns of deficit under pathology that do not arise from general properties of the network, but from a kind of sampling error. Given the small scale of the network, there is always the possibility on a given trial that, just by chance, the weights will be altered in such a way that the network displays a pattern of behaviour supporting almost any hypothesis. Thus, for any single instance of damage, it is not clear to what extent the network's behaviour results from theoretically interesting properties, or from the whims of chance on the given trial (see Plaut, 1995, for further discussion).

A better strategy, on our view, is to examine the network's behaviour averaged across many different instances of damage. Under this approach, a single case is modelled not by a single administration of damage to the network, but by damaging the network several times and assessing its average behaviour. This method has several advantages. First, it eliminates the influence of confirmation biases on the interpretation of network performance, by providing an objective measure of performance that does not allow the theorist to pick and choose which trials to include. Second, it greatly reduces the likelihood that the observed patterns of deficit result from sampling error in selecting the weights or units to be lesioned. Third, it requires the theorist to provide an actual explanation of why patients differ, instead of simply attributing patient variability to chance. That is, in order to account for data across a variety of patients, the theorist must identify those parameters of the model that lead it, on average, to behave like patient X under one choice of values, and like patient Y under a different choice. When the effects of such parameters are understood in the model, the theorist is a step closer toward understanding how they may operate in the actual system. While Devlin et al. (1998) demonstrate that their model is capable of showing the predicted behaviour, and they provide a good explanation for why this might be, they leave us wondering why, if their theory is correct, the model does not show the predicted behaviour on 22 percent of the trials; and why the model does not show the effect at all when its performance is averaged across all 50 damage trials. A more satisfactory account would shed light on the factors that lead the network to behave sometimes one way, sometimes another, and would relate these back to assumed properties of the actual system.

Though our strategy has the benefit of providing a clear criterion for interpreting the network's impaired performance, it also has the consequence of reducing the range of behaviours that a network can "explain." In examining only the central tendencies in the model's behaviour for a given type of lesion, it becomes more difficult to account for extreme cases in the literature. It may be fairly easy to show that such cases fall within the range of behaviours produced by a network under damage; but considerably more difficult to construct a network that, on average, behaves as a single extreme case.

For this reason, it seems unlikely to us that unitary-semantics hypotheses of the kind we have discussed will be capable of explaining the full range of reported category-specific patients. All of the unitary-semantics theories to which we have made reference predict that performance on semantic tasks should, on average, decline as illustrated in Figure 3. Each of these models might, by chance, show almost any pattern of breakdown on a single trial of damage. However, barring the consideration of individual damage trials, they all make clear predictions about the range of deficits that ought to be observed in the literature. Patients with dissociations of knowledge favouring items in the unstructured domain (whichever it may be under a given theory) should never be at ceiling for such items, but must always show at least a mild impairment. Similarly, patients with preserved knowledge of the "structured" domain must never show chance performance in the unstructured domain, but should always have some spared knowledge for such items. However, the existence of extreme cases in the literature (for example, JBR and SBY; see Warrington & Shallice, 1984) contradict these predictions. Simply attributing these cases to chance hardly constitutes a satisfactory explanation, but we see no other way for unitary semantics theories to accommodate them.

There are also empirical findings which, *prima facie*, seem incompatible with a unitary semantics hypothesis. As we have already noted, functional neuroimaging studies suggest that different cortical areas may be differentially involved in processing information about different kinds of objects (Martin et al., 1995), or in processing different kinds of semantic information about objects (Mummery et al., 1998); and these results are consistent with differences in the neuropathology that accompanies impaired knowledge of living or nonliving things (e.g. Gainotti et al., 1995; Damasio et al., 1996). Though these studies are by no means conclusive, they do not fit well into the unitary-semantics framework.

Of course, there is no reason why the principles illustrated in unitary-semantics models may not act in concert with principles of anatomic specialisation to provide a full account of the data. In fact, Devlin et al. (1998) replicate Farah and McClelland's results in their own model, by administering focal damage to either the functional- or sensory-semantic units in their architecture. They suggest that some severe cases of category-specific deficits may result from such anatomically localised lesions to sensory or motor areas as indicated by the sensory-functional hypothesis, while other more graded impairments may arise from general damage to the entire system, under the assumption that semantic representations share structure. Aside from our concerns about how the model was tested, it seems possible that some variation on this approach may ultimately prove fruitful, as we suggest in the next section.

Even from this stance, however, there remain many important questions to be answered. There still exists little consensus regarding the extent to which representations of various kinds of objects share structure, or which kinds of structure contribute to the preservation of se-

mantic knowledge in the face of damage (e.g. Garrard et al., in pressa; Moss et al., 1998). As we have seen, different positions on this issue can lead to radically different predictions about the patterns of deficits that should be observed in the patient data, even if one ignores extreme cases.

There is as yet little empirical data available to illuminate any potential relationship between the overall extent of cortical damage, and the relative preservation of knowledge for animal and artifact categories. Gonnerman, Andersen, Devlin, Kempler, and Seidenberg (1997) reported cross-sectional data from 15 DAT patients with impaired semantic memory, supporting Devlin et al.'s theory. Patients with mild cognitive impairment, as assessed by overall naming performance, showed a slight preservation of knowledge for the names of living relative to nonliving things, whereas patients with a greater degree of dysfunction showed quite pronounced category-specific deficits favouring artifacts. However, Garrard, Patterson, and Hodges (1998) failed to replicate these results in another cross-sectional study of DAT, using a different measure of overall cognitive impairment. Moreover, cases of semantic dementia—a progressive syndrome in which semantic knowledge undergoes a profound degradation—typically do not show the consistent preservation of one semantic domain relative to another (Hodges et al., 1995; Lambon Ralph, Graham, Patterson, & Hodges, 1999).

In summary, connectionist instantiations of unitary-semantics hypotheses have demonstrated that representational structure can have profound consequences for the pattern of decline witnessed in different semantic domains. In some cases, such structure can lead to mild double-dissociations in a network under increasing amounts of damage applied to the same locus. However, the range of effects such networks can show has probably been overestimated. Barring the consideration of individual, idiosyncratic instances of damage, it is unlikely that a unitary-semantics theory can produce the degree and variety of category-specific deficits that have been observed in patients.

Synthesis: A Promising Approach

On one hand, models that invoke independent and innate neural processing modules to explain category specific deficits are too fragmented and underspecified to provide a satisfying explanation of the data. On the other, the limited ability of homogeneous connectionist models to explain extreme double dissociations suggest that some anatomical differentiation must be invoked to accommodate the range of patient data. A fruitful middle ground may be found in models that incorporate principles of anatomic specialisation and representational structure to explain the data. It is too early to determine whether the principles that emerge from further investigation in this direction will provide a full account of category-specific deficits. However, recent work in other domains of semantic cognition has been successful in explaining a variety of related phenomena. In the last section of this chapter, we consider a model of semantic dementia put forward by Rogers et al. (1999), which suggests a promising direction for future research.

Semantic dementia refers to the progressive deterioration of semantic memory that is often observed as a consequence of the cortical atrophy that accompanies Pick's disease (Snowden et al., 1989). Patients suffering from the disorder exhibit a marked anomia and a profound difficulty with semantic tasks such as word to picture matching, word and picture sorting, attribute listing, definition, and the Pyramids and Palm Trees test (Howard & Patterson, 1992; Hodges et al., 1995). Other cognitive faculties, however, seem remarkably spared. Patients with semantic dementia make an interesting contrast to those with herpes encephalitis (the vast majority of category-specific cases), as they do not show relative sparing of some semantic domains over others. For example, although patient IW (described above) was apparently worse at retrieving the sensory relative to the functional properties of objects, her ability to name, draw, and define living and nonliving objects was equally impaired (Lambon Ralph et al., 1998); and this appears to be true of semantic dementia patients generally (Hodges et al., 1995).

Although they do not show category-specific deficits in their overall correct performance, semantic dementia patients do show different patterns of incorrect responding when naming animals compared to artifacts. Rogers et al. (1999) interpreted these differences in the context of a connectionist model similar in many respects to those discussed above. The model was proposed to explain the generalised deterioration of semantic memory without preferential sparing of one domain over another; but like semantic dementia patients, it showed different patterns of errors for living and nonliving things. The principles that lead to these differences in the model may help us understand how category-specific deficits can arise from neuroanatomic differences already known to exist in the brain.

The Rogers et al. model is illustrated in Figure 4. Like the FM model, it uses semantics to map between verbal and visual representations, whose structure are determined by the physical properties of the environment. In learning to do so, the network acquires internal representations of objects that reflect their semantic relations (Hinton, 1986; Rumelhart & Todd, 1993).

Each unit in the *Visual* layer responds to some aspect of an object's appearance, such as its shape, texture, colour, or shading. Each unit in the *Verbal* layer represents a propositional statement describing the object; for example, its name, or other explicit attributes that can be expressed verbally. When a picture is presented to the network, units in the visual layer are clamped to the corresponding pattern, and the network's job is to activate the propositional features that apply to the item depicted. Conversely, when the network is presented with a name

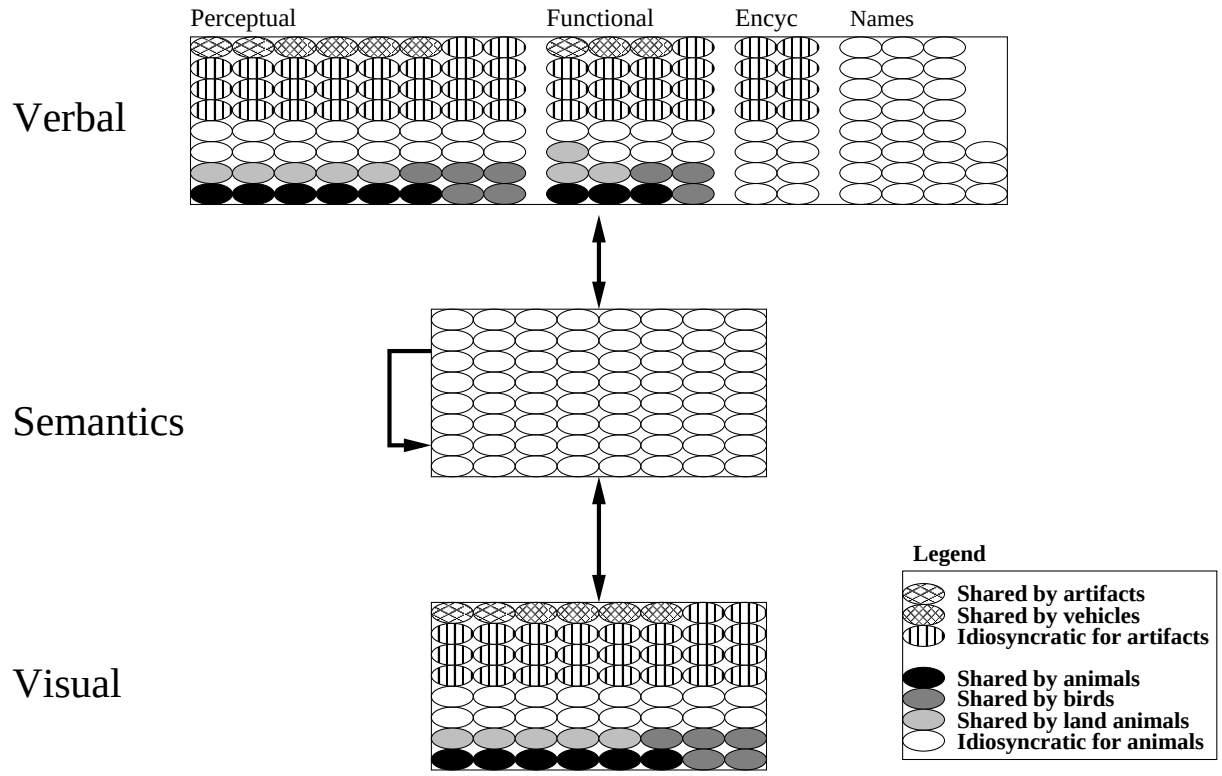

Figure 4. A connectionist model implementing the interaction of high-level verbal and visual representations via semantics. *Visual* and *Verbal* representations are directly activated by input from perception, and hence are bound to the structure of the environment. *Semantic* representations are not assigned, but emerge in the network as a consequence of the learning algorithm, and the structure of the surface representations. The illustration shows the number of distinctive features, and the number of features shared across domain (living thing / artifact) and superordinate category. The top units are local representations of propositional features, such as "can fly" and "has a long neck."

or a verbal description of an object, the corresponding propositional features in the *Verbal* layer are clamped, and the network must correctly activate the appropriate visual units, as well as any propositional units not specified in the input.

Like Devlin et al. (1998), the authors assumed that objects in the world share different degrees of visual and propositional structure, depending upon their category membership. On the basis of similarity measures derived from the featural decomposition of drawings, they constructed visual representations for 32 items drawn from two living (birds and land animals) and two nonliving (vehicles and household objects) categories. Following the results of Garrard et al.'s (in pressb) attribute-listing study, they also constructed propositional representations, intended to capture regularities in the verbal descriptions that people apply to various objects. Items with similar verbal descriptions were represented by similar patterns of activity across the *Verbal* units. These propositions might describe any of the sensory, functional, or encyclopaedic properties of an object. Note, though, that a propositional feature describing a visual property is quite distinct from the visual property itself in this model. Instances of living things had a greater tendency to share both visual and propositional features than did instances of nonliving things.

Unlike either Devlin et al. (1998) or Farah and McClelland (1991), Rogers et al. (1999) did not construct semantic representations for the objects in their corpus. Instead, they trained the network with a variant of the backpropagation learning algorithm suited to recurrent networks. Like the delta-rule described earlier, backpropagation uses the discrepancy between observed and desired activations to adjust weights in the network so that its performance improves. However, backpropagation also allows error derivatives to be passed backward through intermediate units whose representations are unspecified. For example, the error signal from *Verbal* units in the Rogers et al. model could be used to adjust the weights connecting *Visual* and *Semantic* units, without having to specify a particular mediating pattern of activity in the *Semantic* layer. Thus, the network was able to acquire the mappings between verbal and visual patterns without having assigned intermediate representations (Rumelhart & Todd, 1993). In so doing, it came to represent each object in its environment with a stable pattern of activity across its hidden unit. Because these patterns were discovered by the learning algorithm, they may be considered *learned internal representations* (Rumelhart, McClelland, & PDP

Research Group, 1986).

An interesting property of the internal representations that evolve in backpropagation nets is that they capture the similarity structure across their inputs and outputs (Hinton, 1986). In the Rogers et al. model, semantically related items come to be represented by similar patterns of activity across the hidden units, by virtue of the tendency for objects in the same category to share visual and propositional features. Because this tendency is much stronger among animals than among artifacts, the network's representations of animals are much more rigidly structured. The steady states that represent individual birds, for example, are very similar to one another; whereas the attractors that correspond to individual vehicles are much more widely dispersed in representation space. Furthermore, representations of birds and land animals, though fairly distinct, are still more similar to one another than to the various vehicles and household objects. Thus, the network comes to acquire representations that capture the degree of semantic relatedness among objects, by virtue of the backpropagation learning algorithm operating on patterns that share structure.

To understand how the model's behaviour deteriorates in pathology, Rogers et al. (1999) lesioned the network by removing an increasing proportion of all its weights. They then assessed its performance on a model analogue of a picture naming task, in which the damaged network was presented with a visual representation as input, and was allowed to settle to a steady state. To determine the model's response, they simply selected the name unit most strongly activated above its midpoint (0.5). The authors damaged the network several times, and tested its ability to name all 32 items in its vocabulary. The responses were coded as *correct* if the damaged network gave the same response as the undamaged network; as *superordinate errors* when the damaged network gave a correct but more general response than the undamaged network; as *semantic errors* when the damaged network gave an incorrect response from the same superordinate category as the correct response; and as *no response* when the network was unable to activate any name unit above 0.5.

The most interesting result for our purposes is that the network showed a different pattern of responses for animals and artifacts. It made a greater proportion of "No responses" at all levels of impairment in response to artifacts, but a greater proportion of "Superordinate" and "Semantic" errors when naming animals. This is just the pattern of errors found in picture naming with semantic dementia patients (Rogers et al., 1999). That is, both the model and the patients are more likely to give an incorrect response for living things, but less likely to give a response at all for nonliving things.

We can understand this behaviour by considering the similarity structure of the network's internal representations. Recall that the attractors corresponding to individual land animals are quite similar to one another, whereas the attractors corresponding to various vehicles are more widely distributed in representation space. As connections in the network are lesioned, this attractor structure degrades, and the steady states in the space may drift or disappear. Because the representations for various land animals are all similar to one another, such drift is likely to land the network in the incorrect attractor. For example, with damage, the attractor state for *pig* may become unstable. If this happens, "pig" stimuli can get drawn into nearby attractors that have not yet degraded, or into spurious attractors that have formed as a result of damage. Because of the learned similarities between land-animals, such proximal attractors are likely to correspond to the representations of semantically related items, such as *dog*. Accordingly, the network will attribute to the pig the properties it knows to be true of dogs. In some cases—namely, for those properties common to dogs and pigs—the network can still make correct inferences about the object. For example, because the name "Animal" applies both to pigs and to dogs, the network will correctly verify that the pig is an animal, even if it falls into the attractor for *dog*. However, properties that serve to differentiate dogs from pigs may be lost or misattributed when the network's weights are perturbed; for example, the network may attribute the property *is furry* to the pig when it falls into the *dog* attractor. Under this view, errors of commission (such as calling the pig a "Dog") occur more frequently in highly structured domains, because there is a greater likelihood that the network will fall into a neighbouring attractor when damaged.

By contrast, in the domain of nonliving things, there are few proximal attractors into which a given item can fall when its own representation becomes unstable. For example, the model's representation of *spinning wheel* can drift quite far before getting captured by another nonliving object representation. As it wanders into uninformative areas of the space, the network will become increasingly unable to make any inference about the properties of the spinning wheel. As long as it avoids falling to the wrong attractor state, however, the network will not attribute inappropriate properties to the item. Thus, in unstructured domains, the model will increasingly omit properties, but will relatively rarely make errors of commission.

Though intended to explain the generalised impairments observed in semantic dementia, this work has implications for a theory of category-specific deficits as well. The Rogers et al. framework builds in anatomical distinctions between semantic association cortex, and the areas that subserve sensorimotor representations in various modalities. The model implements visual and verbal surface modalities, but these are not the only kinds of information available from the environment. Other surface properties of objects, such as their feel, their taste, and the motor actions they afford, might also reasonably be expected to inform deep semantic representations. In particular, the various ways in which we engage an object in behaviour likely play an important role in our under-

standing of the object's identity—especially for artifacts.

The representations that subserve our ability to act on objects— which we assume to reside in an area of cortex separate from semantic, visual, and verbal areas—may share a degree of structure not mirrored in visual or verbal representations. Objects that afford similar actions, such as a typewriter and a piano, may induce similar representations across areas of cortex that subserve action. This structure may serve to inform the semantic similarities between objects just as visual and verbal representations are assumed to do in the Rogers et al. model. Furthermore, we might expect artifacts and living things to differ in the amount of structure they share across the actions with which they are associated (Moss et al., 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, 1998).

This speculation leads us to a general framework in which semantic representations mediate activity among surface visual, verbal, and action representations. Under this view, lesions to the connections between semantics and either of visual or action areas could result in different category-specific deficits. Damage to the connections between semantics and action representations may lead the network to confuse various artifacts with one another, 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 living things with one another, because of the high degree of visual structure shared in that domain. Deficits specific to language might manifest when the links between verbal and semantic areas are damaged, while the deficits apparent in semantic dementia could arise from generalised deterioration of the semantic units themselves. Thus, like the FM model, the theory has the potential to draw together phenomena from various neuropsychological syndromes. It is also consistent with the data from neuroimaging and neuropathology findings we have discussed.

Conclusion

Though these ideas are speculative at best, they illustrate how known properties of recurrent connectionist networks might be extended to account for category-specific deficits in a manner consistent with the sensory-functional hypothesis, without introducing anatomical distinctions that have not been shown to exist in the brain. Whether or not these principles can accommodate the entire range of patient behaviour is an empirical question. Nevertheless, it seems likely that no story will be complete without appealing both to the structure of representations in the semantic system, and the neuroanatomic architecture of cortex.

References

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

Arguin, M., Bub, D., & Dudek, G. (1996). Shape integration for visual object recognition and its implication in category-specific visual agnosia. *Visual Cognition, 3*(3), 221-275.

Arguin, M., Bub, D., & Dudek, G. (in press). Shape integration for visual object recognition and its implication in category specific visual agnosia. *Visual Cognition.*

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

Battig, W. F., & Montague, W. E. (1969). Category norms for verbal items in 56 categories: A replication and extension of the connecticut category norms. *Journal of Experimental Psychology, 80.*

Behrmann, M., & Lieberthal, T. (1989). Category-specific treatment of a lexical semantic deficit: A single case study of global aphasia. *British Journal of Communication Disorders, 24,* 281-299.

Bullinaria, J. A., & Chater, N. (1995). Connectionist modelling: Implications for cognitive neuropsychology. *Language and Cognitive Processes, 10,* 227-264.

Bunn, E. M., Tyler, L. K., & Moss, H. E. (submitted). *Category-specific semantic deficits: The role of familiarity and property type re-examined.* (Submitted to Neuropsychologia, March 1997)

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

Caramazza, A., & Shelton, J. R. (1998). Domain-specific knowledge systems in the brain: The animate- inanimate distinction. *Journal of Cognitive Neuroscience, 10*(1), 1-34.

Carbonnel, S., Charnallet, A., David, D., & Pellat, J. (1997). One or several semantic system(s)? maybe none: Evidence from a case study of modality and category-specific 'semantic' impairment. *Cortex, 33*(3), 391-417.

Carey, S., & Spelke, E. (1994). Mapping the mind: Domain specificity in cognition and culture. In L. A. Hirschfeld (Ed.), (p. 169-200). New York, NY: Cambridge University Press.

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*(10), 913-919.

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

Damasio, H., Grabowski, T. J., Tranel, D., & Hichwa, R. D. (1996). A neural basis for lexical retrieval. *Nature, 380*(6574), 499-505.

De Renzi, E., & Lucchelli, F. (1994). Are semantic systems separately represented in the brain? The case of living category impairment. *Cortex, 30*(1), 3-25.

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*(1), 77-94.

Dixon, M. (1999). Tool and bird exemplar identification in a patient with category-specific visual agnosia. *Brain and Cognition, 40*(1), 97-100.

Dixon, M., Bub, D., & Arguin, M. (1997). The interaction of object form and object meaning in the identification performance of a patient with category- specific visual agnosia. *Cognitive Neuropsychology*, *14*(8), 1085-1130.

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*(4), 339-357.

Farah, M. J., McMullen, P. A., & Meyer, M. M. (1991). Can recognition of living things be selectively impaired? *Neuropsychologia*, *29*(2), 185-193.

Farah, M. J., & Wallace, M. A. (1992). Semantically-bounded anomia: Implications for the neural implementation of naming. *Neuropsychologia*, *30*, 609-621.

Forde, E. M. E., Francis, D., Riddoch, M. J., Rumiati, R. I., & Humphreys, G. W. (1997). On the links between visual knowledge and naming: A single case study of a patient with a category-specific impairment for living things. *Cognitive Neuropsychology*, *14*(3), 403-458.

Funnell, E., & Sheridan, J. (1992). Categories of knowledge? Unfamiliar aspects of living and nonliving things. *Cognitive Neuropsychology*, *9*(2), 135-153.

Gainotti, G., & Silveri, M. C. (1996). Cognitive and anatomical locus of lesion in a patient with with a category-specific semantic impairment for living beings. *Cognitive Neuropsychology*, *13*, 357-389.

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. (in pressa). Prototypicality, distinctiveness and intercorrelation: Analyses of the semantic attributes of living and nonliving concepts. *Cognitive Neuroscience*.

Garrard, P., Lambon Ralph, M. A., Hodges, J. R., & Patterson, K. (in pressb). *Prototypicality, distinctiveness and intercorrelation: Analyses of the semantic attributes of living and nonliving concepts.* (Manuscript submitted for publication)

Garrard, P., Patterson, K., & Hodges, J. R. (1998). Category-specific semantic loss in dementia of Alzheimer's type: functional-anatomical correlations from cross-sectional analyses. *Brain*, *121*, 633-646.

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. S., & Caramazza, A. (1985). Category-specific naming deficit following cerebral infarction. *Nature*, *316*, 439-440.

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

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

Hillis, A. E., Rapp, B., & Caramazza, A. (1995). Constraining claims about theories of semantic memory: More on unitary versus multiple semantics. *Cognitive Neuropsychology*, *12*(2), 175-186.

Hillis, A. E., Rapp, B., Romani, C., & Caramazza, A. (1990). Selective impairments of semantics in lexical processing. *Cognitive Neuropsychology*, *7*, 191-243.

Hinton, G. E. (1986). Learning distributed representations of concepts. In *Proceedings of the 8th Annual Conference of the Cognitive Science Society* (p. 1-12). Hillsdale, NJ: Erlbaum.

Hinton, G. E., & Shallice, T. (1991). Lesioning an attractor network: Investigations of acquired dyslexia. *Psychological Review*, *98*(1), 74-95.

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.

Howard, D., & Patterson, K. (1992). *Pyramids and palm trees: A test of semantic access from pictures and words.* Bury St. Edmunds, Suffolk, U.K.: Thames Valley.

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

Kurbat, M. A. (1997). Can the recognition of living things really be selectively impaired? *Neuropsychologia*, *35*(6), 813-827.

Kurbat, M. A., & Farah, M. J. (1998). Is the category-specific deficit for living things spurious? *Journal of Cognitive Neuroscience*, *10*(3), 0898-929.

Laiacona, M., Barbarotto, R., & Capitani, E. (1993). Perceptual and associative knowledge in category specific impairment of semantic memory: a study of two cases. *Cortex*, *29*, 727-740.

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*(3), 309-335.

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.

Lecours, S., Arguin, M., Bub, D., Caille, S., & Fontaine, S. (1999). Semantic proximity and shape feature integration effects in visual agnosia for biological kinds. *Brain and Cognition*, *40*(1), 171-174.

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

Martin, N., Gagnon, D. A., Schwartz, M. F., Dell, G. S., & Saffran, E. M. (1996). Phonological facilitation of semantic errors in normal and aphasic speakers. *Language and Cognitive Processes*, *11*(3), 257-282.

Mayall, K., & Humphreys, G. (1996). A connectionist model of alexia: Covert recognition and case mixing effects. *British Journal of Psychology*, *87*(3), 355-402.

McCarthy, R., & Warrington, E. K. (1988). Evidence for modality-specific meaning systems in the brain. *Nature*, *334*, 428-430.

McClelland, J. L., & Rumelhart, D. E. (1985). Distributed memory and the representation of general and specific information. *Journal of Experimental Psychology: General*, *114*(2), 159-188.

McRae, K., Sa, V. R. de, & Seidenberg, M. S. (1997). On the nature and scope of featural representations of word meaning. *Journal of Experimental Psychology: General*, *126*(2), 99-130.

Moore, C. J., & Price, C. J. (1999). A functional neuroimaging study of the variables that generate category-specific object processing differences. *Brain*, *122*, 943-962.

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: Case studies in neuropsychology, neuropsychiatry, and behavioural neurology*, *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*(6), 766-777.

Perani, D., Cappa, S. F., Bettinardi, V., & Bressi, S. (1995). Different neural systems for the recognition of animals and man-made tools. *Neuroreport*, *6*(12), 1637-1641.

Perani, D., Schnur, T., Tettamanti, M., Gorno-Tempini, M., Cappa, S. F., & Fazio, F. (1999). Word and picture matching: A PET study of semantic category effects. *Neuropsychologia*, *37*(3), 293-306.

Perry, C. (1999). Testing a computational account of category-specific deficits. *Journal of Cognitive Neuroscience*, *11*(3), 312-320.

Pinker, S. (1994). *The language instinct*. New York: Morrow.

Pinker, S. (1997). Words and rules in the human brain. *Nature*, *387*(6633), 547-548.

Plaut, D. C. (1995). Double dissociation without modularity: Evidence from connectionist neuropsychology. *Journal of Clinical and Experimental Neuropsychology*, *17*(2), 291-321.

Plaut, D. C. (1998). Systematicity and specialization in semantics. In D. Heinke, G. W. Humphreys, & A. Olson (Eds.), *Connectionist models in cognitive neuroscience: Proceedings of the fifth annual neural computation and psychology workshop*. New York: Springer.

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.

Plaut, D. C., & Shallice, T. (1993). Deep dyslexia: A case study of connectionist neuropsychology. *Cognitive Neuropsychology*, *10*(5), 377-500.

Powell, J., & Davidoff, J. (1995). Selective impairments of object knowledge in a case of acquired cortical blindness. *Memory*, *3*, 435-461.

Riddoch, M. J., & Humphreys, G. W. (1987). A case of integrative visual agnosia. *Brain*, *110*, 1431-1462.

Rogers, T. T., Lambon-Ralph, M., Patterson, K., McClelland, J. L., & Hodges, J. R. (1999). A recurrent connectionist model of semantic dementia. In *Cognitive neuroscience society annual meeting program 1999.*

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., & McClelland, J. L. (1986). A general framework for parallel distributed processing. In D. E. Rumelhart, J. L. McClelland, & the PDP Research Group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition* (Vol. 1, p. 45-76). Cambridge, MA: MIT Press.

Rumelhart, D. E., McClelland, J. L., & PDP Research Group the. (1986). *Parallel distributed processing: Explorations in the microstructure of cognition. Volume 1: Foundations.* 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, artifical intelligence, and cognitive neuroscience* (p. 3-30). Cambridge, MA: MIT Press.

Sacchett, C., & Humphreys, G. W. (1992). Calling a squirrel a squirrel but a canoe a wigwam: A category-specific deficit for artefactual objects and body parts. *Cognitive Neuropsychology*, *9*(1), 73-86.

Saffran, E. M., & Schwartz, M. F. (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 aspects of cognitive processing* (p. 507-536). Hillsdale, NJ: Erlbaum.

Samson, D., Pillon, A., & De Wilde, V. (1998). Impaired knowledge of visual and non-visual attributes in a patient with a semantic impairment for living entities: A case of a true category-specific deficit. *Neurocase*, *4*(4/5), 273-290.

Sartori, G., & Job, R. (1988). The oyster with 4 legs: A neuropsychological study on the interaction of visual and semantic information. *Cognitive Neuropsychology*, *5*, 105-132.

Shelton, J. R., & Caramazza, A. (1999). Deficits in lexical and semantic processing: Implications for models of normal language. *Psychonomic Bulletin & Review*, *6*, 5-27.

Shelton, J. R., Fouch, E., & Caramazza, A. (1998). The selective sparing of body part knowledge: A case study. *Neurocase*, *4*, 319-351.

Sheridan, J., & Humphreys, G. W. (1993). A verbal-semantic category-specific recognition impairment. *Cognitive Neuropsychology*, *10*, 185-200.

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

Snowden, J. S., Goulding, P. J., & Neary, D. (1989). Semantic dementia: A form of circumscribed cerebral atrophy. *Behavioral Neurology*, *2*, 167-182.

Spitzer, M., Kischka, U., Gueckel, M. E., Friedemann amd Bellemann, Kammer, Thomas, Seyyedi, S., Weisbrod, M., Schwartz, A., & Brix, G. (1998). Functional magnetic resonance imaging of category-specific cortical activation: Evidence for semantic maps. *Cognitive Brain Research*, *6*(4), 309-319.

Spitzer, M., Kwong, K. K., Kennedy, W., & Rosen, B. R. (1995). Category-specific brain activation in fmri during picture naming. *Neuroreport: An International Journal for the Rapid Communication of Research in Neuroscience*, *6*(16), 2109-2112.

Springer, K., & Keil, F. (1991). Early differentiation of causal mechanisms appropriate to biological and nonbiological kinds. *Child Development*, *62*, 767-781.

Stewart, F., Parkin, A. J., & Hunkin, N. M. (1992). Naming impairments following recovery from herpes simplex encephalitis: Category specific? *Quarterly Journal of Experimental Psychology*, *44A*, 261-284.

Tippett, L. J., McAuliffe, S., & Farah, M. J. (1995). Preservation of categorical knowledge in Alzheimer's disease: A computational account. *Memory*, *3*, 519-553.

Tyler, L. K., & Moss, H. E. (1997). Functional properties of concepts: Studies of normal and brain-damaged patients. *Cognitive Neuropsychology*, *14*(4), 511-545.

Tyler, L. K., & Moss, H. E. (1998). Going, going, gone...? Implicit and explicit tests of conceptual knowledge in a longitudinal study of semantic dementia. *Neuropsychologia*, *36*(12), 1313-1323.

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

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.

Wellman, H. M., & Gelman, S. A. (1997). Knowledge acquisition in foundational domains. In D. Kuhn & R. Siegler (Eds.), *Cognition, perception and development* (Vol. 2, 5 ed., p. 523-573). New York: John Wiley and Sons.

Zeki, S., & al. et. (1991). A direct demonstration of functional specialization in human visual cortex. *Journal of Neuroscience*, *11*, 641-649.