CORE
Metadata, citation and similar papers at core.ac.uk
Provided by Birkbeck Institutional Research Online

Could Category-Specific Semantic Deficits Reflect Differences in the Distributions of Features Within a Unified Semantic Memory?

(In *Proceedings of the 20th Annual Conference of the Cognitive Science Society*, 1998, NJ:LEA, 374-379)

Robert M. French (rfrench@ulg.ac.be)
Psychology Department, B32, University of Liège
4000 Liège, Belgium

Denis Mareschal (d.mareschal@exeter.ac.uk)[1]
Department of Psychology, Exeter University
Perry Rd., Exeter, EX4 4QG, UK

Abstract

Category-specific semantic deficits refer to the inability to name objects from a particular category while the naming of words outside that category remains relatively unimpaired. We suggest that such semantic deficits arise from the random lesioning of a unified semantic network in which internal category representations reflect the variability of the categories themselves. This is demonstrated by lesioning networks that have learned to categorize butterflies and chairs. The model shows category-specific semantic deficits of the narrower category (butterfly) with the occasional reverse semantic deficits of the relatively broader category (chair) .

Introduction

Category-specific semantic deficits refer to the inability to name objects from a particular category as a result of neurological damage. The naming of objects outside the impaired category is relatively well preserved. Perhaps the most striking category-specific semantic deficit is the dissociation found between animate vs. inanimate objects. In general, naming of inanimate objects is found to be better preserved than naming of animate objects (Warrington and Shallice, 1984; Silveri & Gainotti, 1989; Gainotti & Silveri, 1996; Sartori & job, 1988; Funnell & Sheridan, 1992; Farah, Meyer, & McMullen, 1996). However, for a small number of patients, the naming of animate objects is better preserved (Warrington & McCarthy, 1987; Hillis & Caramazza, 1991, Sacchett & Humphreys, 1992).

Warrington and Shallice (1984) have tried to explain these findings by suggesting that words for animate and inanimate objects are learned in different ways. Words for animate objects are learned primarily though association with perceptual cues because animate objects tend to be described by their surface features (e.g., color, size). However, words for inanimate objects are learned primarily through association with the object's function because inanimate (man-made) objects tend to be described by their use (e.g., a car is for driving). According to this view, the naming dissociation does not reflect a taxonomic ordering of semantic memory, but rather, the differing proportion of the type of semantic features (perceptual vs. functional) associated with a word.

Farah and McClelland (1991) explored this hypothesis by constructing a connectionist model of semantic memory and lesioning it. In this model, both animate and inanimate words were associated with functional as well as perceptual features. However, the proportion of functional and perceptual features differed for animate and inanimate words respectively. They found that by lesioning either the perceptual *or* functional components of semantic memory, animate or inanimate words were impaired respectively. This was used to corroborate Warrington and Shallice's account of category-specific semantic deficits. Category-specific semantic deficits arose even though words were not stored with respect to semantic category.

While this is a plausible account of the source of category-specific semantic deficits that does not appeal to the prior taxonomic organization of semantic memory, it still implies that there exists an intrinsic dissociation in the way that functional and perceptual features are stored in semantic memory. The Farah and McClelland model only works because there are identifiable regions that encode one or the other type of semantic information, and that can be lesioned selectively. This account still relies on an *a priori* structuring of semantic memory to explain the observed semantic dissociation. The only difference is that rather than positing an explicit taxonomic order, the taxonomic ordering is mediated by a high correlation between perceptual features with animate objects, and functional features with inanimate objects.

The greatest shortcoming of the model is that it fails to explain why damage should occur either (a) selectively to the perceptual features (thereby preserving knowledge of inanimate words) or (b) selectively to the functional features (thereby preserving knowledge of animate objects).

[1] Now at the Department of Psychology, Birkbeck College, University of London, Malet St., London, WC1E 7HX, UK.

Some evidence for separate perceptual or functional memory damage comes from the neuropathologies associated with category-specific semantic deficits. Localized damage (e.g., from herpes encephalitis) to the temporolimbic system (resulting in a loss of perceptual features), or to the frontoparietal regions (resulting in a loss of functional features have been associated with the loss of one semantic category or the other (Saffran & Schwartz, 1988). However, category specific impairments have also been found in patients with Alzheimer’s disease, a widespread pathology, causing damage to both the temporolimbic system and the frontoparietal regions. (Gonnerman, Andersen, Devlin, Kempler, & Seidenberg, in press; McKrae, De Sa, & Seidenberg, 1997; Silveri & Gainotti, 1988). This sort pathology cannot be modelled by selectively leasioning neurons in separate memories. One would expect that the diffuse neural damage found in Alzheimer’s patients with category-specific semantic deficits would result in equal damage to perceptual and functional features. Hence, even if inanimate words have more functional features and animate words have more perceptual features both categories would be equally impaired by the random damage.

In this paper we present a connectionist model of category-specific semantic deficits that does not assume an initial partitioning of semantic memory along either a taxonomic or a perceptual/functional divide. The model posits a unified semantic memory in which all features are treated equally (e.g., Caramazza, Hills, Rapp, & Romani, 1990). Category-specific semantic deficits arises from random lesioning of the network. The model we propose suggests that category-specific semantic deficits reflect differences in the *variability* of features encoding both animate and inanimate objects.

The rest of this paper proceeds as follows. First, we briefly present the pseudo-recurrent network architecture developed by French (1997a) and used for modeling semantic memory (French, 1997b). An explanation for category-specific semantic deficits is then presented. The next section illustrates this process by lesioning networks that have been trained with two real world categories. Finally, the likelihood of recovery from damage is discussed.

Pseudo-Recurrent Connectionist Networks

The architecture discussed in this paper was first developed by French (1997a) to overcome catastrophic interference in backpropagation (BP) networks. It suggests that catastrophic interference in memory can be overcome by mixing in approximations of previously learned patterns (“pseudopatterns” of Robins, 1995) with new information during learning. Learning proceeds in two stages. The first stage (which involves mixing new and old information) takes place in an early-processing area of the network. The second stage (which involves laying down the new knowledge) takes place in the final-storage area.

This method is analogous to that used by McClelland, McNaughton, and O’Reilly (1995) to model the exchange of information between the hippocampus and the neocortex involved in the laying down of memories. The shunting of information between two memory systems is believed to have evolved as a natural way of overcoming the problem of catastrophic interference in a distributed system such as the brain.

The network consists of a feedforward BP network that is divided into two parts, one used to help train the other (Figure 1). We will call the left-hand side of the network the “early-processing memory” and the right-hand side the “final-storage memory.” It is perhaps easiest to explain how the network works in terms of a specific example.

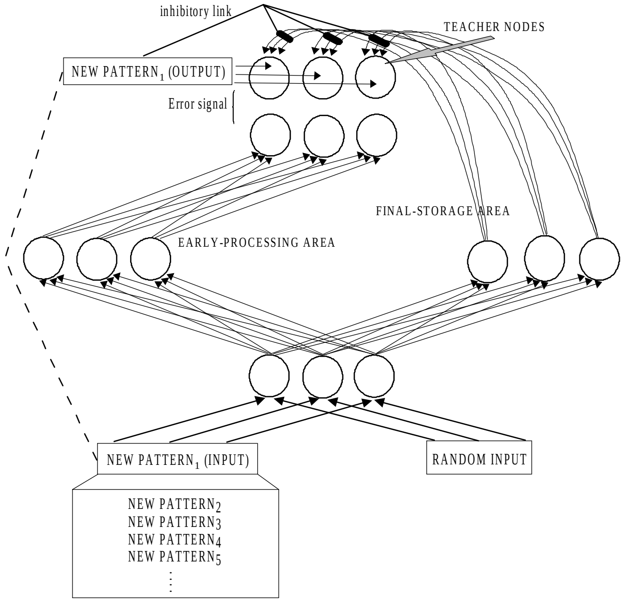


Figure 1. The pseudo-recurrent network architecture

Suppose that the “final-storage” area contains what the network has learned up to the present time. The network is then asked to sequentially learn 20 new patterns, P_1, P_2, ... P_{20}. Each of these patterns, P_i, consists of an input and an output (“teacher”) association: (I_i, T_i). By sequentially learning these patterns we mean that each individual pattern must be learned to criterion before the system can begin to learn the subsequent pattern. To learn pattern P_1, its input I_1 is presented to the network. Activation flows through both parts of the network, but the output from the final-storage part is prevented from reaching the teacher nodes by the “real” teacher T_1. In other words, the teacher pattern T_1 fills the teacher nodes. The early-processing network then adjusts its weights with the standard backpropagation algorithm using as the error signal the difference between T_1 and the output O_1 of the early-processing network. Crucially, however, the early-processing network does not only learn the pattern P_1. Internally created *pseudopatterns*, reflecting the contents of final-storage, are also generated by the final-storage memory and will be learned by the early-processing memory along with P_1.

Pseudopatterns are generated by final-storage and learned by the early-processing memory as follows. A random input pattern, i_1, is presented to the input nodes of the system. This input produces an output, o_1, at the output layer of the early-processing memory and also produces an output, t_1, on the teacher nodes of the final-storage memory. This input-output pair (i_1, t_1) defines a pseudopattern, ψ_1, that reflects the contents of the final-storage memory. The difference between t_1 and o_1 determines the error signal for changing the weights in the early-processing memory. Similarly, the other random inputs, i_2, i_3, . . . i_n, produce pseudopatterns, ψ_2, ψ_3, . . . ψ_n that are also be learned by the early-processing memory. Once the weight changes have been made for the first epoch for the set of patterns {P_1, ψ_1, ψ_2, . . . ψ_n}, the early-processing memory cycles through this set of patterns again and again until it has learned them all to criterion. By learning the pattern P_1 the early-processing memory is learning the new information presented to it; by learning the pseudopatterns $\psi_{1, \ldots,}$ ψ_n, the early-processing memory is, in addition, learning an approximation of the information previously stored in final storage. Obviously, the more pseudopatterns that are generated, the more accurately they will reflect the contents of final storage. Once learning in the early-processing network has converged for P_1, ψ_1, ψ_2, . . . ψ_n, the early-processing weights then replace the final-storage weights. In other words, the early-processing memory *becomes* the final storage memory and the network is ready to learn the next pattern, P_2. (Note that this weight-copying strategy is certainly not biologically plausible. However, it has been shown (French, 1997a) that information transfer can also be effectively done from early-processing to final-storage by means of the above type of pseudo-pattern transfer.

The essence of this technique is to interleave new information to be learned with pseudopatterns that reflect the contents of final-storage. Thus, rather than interleaving the real, originally learned patterns with the new input coming to the early-processing memory, we do the next best thing — namely, we interleave pseudopatterns that are *approximations* of the previously stored patterns. Once the new pattern and the pseudopatterns are learned in the early-processing area, the weights from the early-processing network are copied to the corresponding weights in the final-storage network (or, more plausibly, the early-processing area trains the final-storage area using its own set of pseudopatterns).

The model is called "pseudo-recurrent" not only because of the recurrent nature of the training of the early-processing memory by the final-storage memory — approximations of previously learned information is continually fed back into the early-processing area from final-storage —, but also as a means of acknowledging the all-important mechanism of information transfer from final-storage to early-processing storage — namely, pseudopatterns. During sequential learning, information is continually passed back and forth between the two memory areas by means of pseudopatterns.

One unanticipated result of this use of pseudo patterns is the compression of the representations that develop in final storage. This is illustrated for a particular example in Figures 2 and 3, and discussed in more detail below. Compression has numerous advantages. In particular, there is a decrease in the number of resources required to activate any given word, and a decrease in the amount of overlap in final storage. Compact representations may, presumably, allow for more efficient processing of incoming stimuli because of their reduced demand on system resources (i.e., less activation is required to fully activate a compact representation. However, highly compact representations are more vulnerable to selective damage than highly distributed representations.

It is worth repeating that the pseudorecurrent architecture is meant to capture the natural process by which the brain may be overcoming catastrophic interference (McClelland, McNaughton, and O'Reilly; 1995). The compression of categorical representations is a processing by-product that falls naturally out of the pseudo-recurrent mechanism.

A Mechanism for Category-Specific Loss

In contrast to explanations of category-specific semantic deficits that rely on the perception/function distinction (Warrington & Shallice, 1984; Durrant-Paetfield, Tyler, Moss, & Levy, 1997; Farah & McClelland, 1991), we suggest that this selective memory loss is due, at least in part, to the considerable difference in the average variability within most biological and artificial kinds. This difference, is combined with the phenomenon of gradual compression of representations as they are consolidated in final-storage — making them increasingly susceptible to damage.

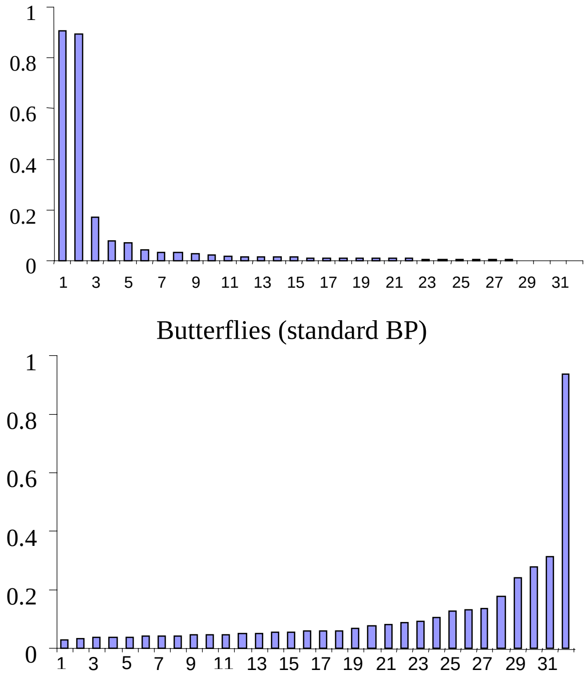
When two real-world categories that have very different variance are stored in a network — connectionist or human — the difference in variance will be reflected in a difference in the variance of the internal representations of the two categories. The greater the variance in the real-world category, the greater the variance in the internal representation of that category, where the variance of an internal representation is determined by the "spread" of the distribution of the hidden-unit activation pattern corresponding to a representation when it is activated.

The more compact the distribution (i.e., the lower the variance) the more vulnerable the category is to catastrophic damage. This is because the loss of one or two nodes in a narrowly defined category corresponds to a greater proportional loss of information. This is true in any distributed connectionist network. The pseudo-recurrent network enhances this effect by effectively reducing the number of nodes participating in the representations. We explore this account further by training a pseudorecurrent network with two categories: one artificial (CHAIR) and the other natural (BUTTERFLY).

Animate vs. Inanimate Semantic Dissociations

Twenty standard Backpropagation (BP) and 20 pseudo-recurrent networks with 13 input units, 13 output units, and 32 hidden units each were trained to autoassociate 20 examples of both butterflies and tables (for a total of 40 tokens). The parameter training values were as follows: initial weight range: [–2, 2], learning rate: 0.1, momentum: 0.9, and Fahlman offset: 0.1. The pseudo-recurrent networks used 15 pseudo-patterns in learning.

The categories of CHAIR and BUTTERFLY were chosen because they were familiar categories with extremely high naming reliability (100% for both) and very similar image agreements (chair: 3.22 and butterfly: 3.92) according to the Snodgrass and Vanderwart (1980) picture naming data. Subjects recognize these categories easily and have similarly well-defined mental images for both categories. The 40 exemplars were coded along the following 13 dimensions: head-length, head width, eye separation, antenna length, dominant colour, leg length, number of legs, vertical extent, horizontal extent, number of angles, material, surface incline, deformability. Measurements were taken from randomly selected actual examples of butterflies and chairs as detailed in Howarth (1973) and Humphreys (1970) respectively.



Butterflies (standard BP)

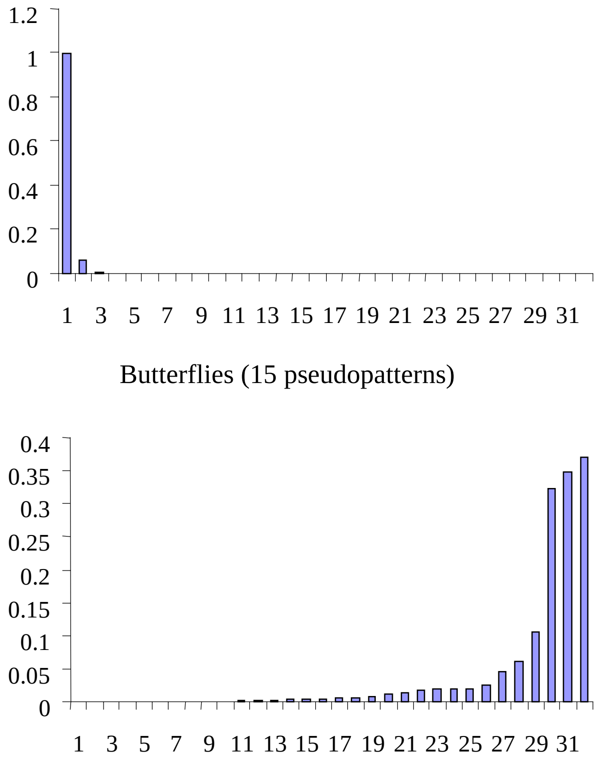
Chairs (standard BP)

Figure 2. Hidden unit activation profile for Butterflies and Chairs in a standard BP network.

Exemplars were selected randomly and were learned sequentially: each exemplar was learned to criterion before the next was presented. Training (autoassociation) was stopped when all outputs were within 0.2 of their target or after 1000 epochs of training. Figure 2 shows an example of the internal representation developed across the hidden units for the BUTTERFLY and CHAIR categories for the standard BP network.

Both categories are encoded over the whole band of hidden units. Presentation of a butterfly exemplar produces activation on 28 out of 32 hidden units while presentation of a chair produces activation on all 32 units.

Figure 3 shows the same internal representations for the pseudo-recurrent networks trained with 15 pseudo-patterns. These hidden unit representations are much more compact. The BUTTERFLY category is only coded across 3 hidden units while the CHAIR category is coded across 22 units. As compared with the standard BP networks, there is a net decrease in the variance of the internal representations of both categories as measured by the number of units required to encode them. It is worth noting that a loss in information also accompanies this representational compression. Although the pseudo-recurrent networks can autoassociate as well as the standard BP networks, the finer details of the exemplars are lost during the compression process.



Butterflies (15 pseudopatterns)

Chairs (15 pseudopatterns)

Figure 3. Hidden unit activation profile for Butterflies and Chairs in a network trained with 15 pseudopatterns.

To explore the robustness of these representations, the 40 networks were systematically lesioned by removing each of the hidden units one at a time. There were 32 possible lesions for each network for a total of 640 lesioning instances. The systematic lesioning approach guarantees that the whole space of possible damage is explored. Table 1 shows the proportion of networks having completely lost the BUTTERFLY or CHAIR categories (but having preserved the other category) for both standard BP and pseudo-recurrent networks.

Table 1. Percentage of lesions resulting in total category loss (n=640)

	Standard BP	Pseudorecurrent
CHAIR	0%	0.3%
BUTTERFLY	0%	3%

With standard BP none of the lesions resulted in total category loss. The distributed representations are immune to this type of lesioning. However category loss did appear in a small but significant number of pseudo-recurrent networks. Three percent of lesions resulted in the complete loss of BUTTERFLY (while preserving CHAIR) and 0.3 % resulted in the loss of CHAIR (while preserving BUTTERFLY). The natural kind category was more likely to be selectively lost than the artificial kind. However, it is important to note that about 1/10th as many lesions resulted in the opposite effect: the selective loss of the artificial category. This is compatible with the finding that for a small number of patients, the naming of animate objects is better preserved (Warrington & McCarthy, 1987; Hillis & Caramazza, 1991, Sacchett & Humphreys, 1992). Both phenomena can be explained by appealing to the same random damage and the different distribution characteristics of the categories.

One implication of this approach is that relative category loss can only be meaningfully evaluated between categories that are at a similar taxonomic level. Basic individual categories such as BUTTERFLY and CHAIR would both be lost before superordinate categories such as ANIMAL and FURNITURE because both the latter categories have much more variation than either of the basic level categories.

Recovery From Damage

Early in learning few pseudopatterns have been mixed in with the categorical information. Each category remains relatively broadly defined across the hidden units. As learning progresses (as the network gets older and more pseudo-patterns are mixed in) the categories become more compact and more tightly defined. One implication is that random damage early in learning (at a young age) will produce general damage to all categories but is unlikely to catastrophically damage any one category. Because no category is eradicated, there is a much better chance that a small amount of subsequent exposure to examples of that category will produce a complete recovery of the category.

In contrast, older networks have narrowly defined, relatively sparse category representations. As a result, random damage is less likely to effect any of them. However, if a category is damaged, it is more likely to be catastrophically damaged and unable to recover with subsequent exposure to examples of that category.

In short, young networks are more susceptible to minor damage but can recover from the damage whereas older networks are more resilient to damage but more brittle and less able to recover from damage.

Discussion

In this paper we have presented a simple model of category-specific semantic deficits. The model uses the pseudorecurrent architecture devised by French (1997a). Learning occurs by mixing in information already present within a network with the new information before laying it down in a network by using backpropagation. One result of this process is that categorical representations become more compact, less distributed, and more susceptible to catastrophic damage. We suggest that such a mechanism could account for a range of category-specific semantic deficits.

The pseudorecurrent architecture was used to model the category-specific semantic deficits observed between animate and inanimate objects. This was illustrated by training the network on one animate category (BUTTERFLY) and one inanimate category (CHAIR). After consolidation with pseudopatterns the networks were exposed to random diffuse lesioning. There was a predominant loss of the animate category (Butterfly) with a small minority of networks showing the reverse effect of losing the inanimate category. This closely matches what is found with human patients. The present model does not have to appeal to a structured semantic memory (for example by positing taxonomic or percept/function structures in memory) and is therefore more parsimonious than previous models of semantic dissociation.

Minimal systematic lesioning was used to explore the robustness of the category representations. Clearly, increasing the amount of lesioning would increase the amount of loss in both categories. An important implication of the model is that category-specific semantic deficits can occur even with minimal lesioning. However, the large majority of individuals experiencing this type of damage would not report any loss. This suggests that the number of people having suffered damage may be far greater than the number who are actually diagnosed with an semantic deficits.

We do not wish to claim that there are no differences between perceptual or functional object information. There are many reasons to believe differences exist (at the very least in terms of encoding) and that these difference may impact on the ability to retrieve animate or inanimate words. The model we present is very simple and explores a single, simple mechanism that can produce category-specific deficits as a results of random damage. The basic point it makes is that one does not need to appeal to a structured, or separate semantic memories to account for category-specific dissociations. McRae, De Sa and Seidenberg (1997) make a related point. They suggest that knowledge of animate objects is more susceptible to diffuse damage in a unified memory because that knowledge is encoded across a smaller set of more highly correlated features units than knowledge of inanimate objects.

Finally, the model suggests that more attention should be paid to the input statistics of the categories used for testing semantic deficits. A strong prediction of the model is that — within a given subject — the variability of categories should be a strong predictor of whether they are preserved or not. Thus, within a subject showing semantic deficits, the categories with broader definitions should be preserved independently of whether they are animate, inanimate, concrete, or abstract.

Acknowledgements

This work was supported in part by the Belgian science foundation grant PAI/4-19 and by a grant from the CGRI-FRNS/British Council collaborative research fund. The authors would also like to thank Jean-Pierre Thibault for his significant contribution to the ideas in this paper and Janice Kay for her critical reading of a previous draft of this paper.

References

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

Durrant-Peatfield, M., Tyler, L., Moss, H. & Levy, J. (1997). The distinctiveness of form and function in category structure: A connectionist model. *Proceedings of the Nineteenth Annual Conference of the Cognitive Science Society*, Mahwah, NJ:LEA.

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

Farah, M., Meyer, & McMullen, (1996). The living/non-living distinction is not an artefact: Giving an a priori implausible hypothesis a strong test. *Cognitive Neuropsychology, 13, 137-154.*

Franklin, S., Howard, D., & Patterson, K. (1995). Abstract word anomia. *Cognitive Neuropsychology, 12, 549-566.*

French, R. M. (1997a). Pseudo-recurrent connectionist networks: An approach to the “sensitivity–stability” dilemma. *Connection Science, 9*(4), 353-379.

French, R. M. (1997b). Selective memory loss in aphasics: An insight from pseudo-recurrent connectionist networks. In J. Bullanaria, G. Houghton, D. Glasspool (Eds.). *Connectionist representations: Proceedings of the Fourth Annual Computation and Psychology Workshop* (pp. 183-195). Spinger-Verlag.

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

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

Gonnerman, L. M., Andersen, E. S., Devlin, J. T., Kempler, S., & Seidenberg, M. S. (in press). Double dissociation of semantic categories in Alzheimer’s disease. *Brain and Language*.

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

Hollis, R. (1970). *Modern chairs*. London, UK:Lund Humphries.

Howarth, T. G. (1973). *Butterflies of the British isles*. London, UK: Viking.

McClelland, J., McNaughton, B., & O'Reilly, R. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*. *102*, 419–457.

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

Robins, A. (1995). Catastrophic forgetting, rehearsal and pseudorehearsal. *Connection Science, 7*(2), 123–146.

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

Safran, E. M., Bogoyo, L. C. Schwartz, M. F. & Marin, O. S. M. (1980). Does deep dyslexia reflect right hemisphere reading? In M. Coltheart, K. E. Patterson & J. C. Marshal (Eds.). *Deep dyslexia*. London: ß & Kegan Paul.

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

Sartori, G., Miozzo, M., & Job, R. (1993). Category-specific form-knowledge deficit in a patient with Herpes Simplex Virus Encephalitis. *The Quarterly Journal of Experimental Psychology, 46A*(3) 489-504.

Silvari, M. & Gainotti, G (1988). Interaction between vision and language in category-specific semantic impairment. *Cognitive Neuropsychology, 5*, 677-710.

Snodgrass, J. G. & Vanderwart, M. (1980). A standardized set of 260 pictures: Norms for name agreement, Image agreement, Familiarity, and visual complexity. *Journal of Experimental Psychology: Human Learning and Memory, 6*, 174-215.

Warrington, E. K. (1981). Concrete word dyslexia. *British Journal of Psychology, 72*, 175-196.

Warrington, E. K. & McCarthy, R. (1987). Categories of knowledge: Further fractionisations and an attempted integration. *Brain, 11*, 1273-1296.

Warrington, E, K. & Shallice, T. (1984). Category-specific semantic impairments. *Brain, 107*, 829-859.