Implications of Distributed Representations for Semantic Processing: Evidence from Alzheimer's Disease

Justin M. Aronoff[1], Laura M. Gonnerman[2], Elaine S. Andersen[1, 3,4], Daniel Kempler[5] and Amit Almor[3]

Departments of Linguistics[1], Psychology[3], and Program in Neuroscience[4],
University of Southern California, Los Angeles, CA 90089
[2]Department of Psychology, Lehigh University, Bethlehem, PA 18015
[5]Comunication Sciences and Disorders, Emerson College, Boston, MA 02116-4624

Abstract

Prior work by Gonnerman and colleagues presented a theory of semantic processing in normal and impaired populations. Their account incorporates distributed representations and predicts a complex relationship between semantic knowledge and naming ability. According to this account, during the course of progressive brain damage, one should observe different relationships between damage to semantic knowledge and naming ability for natural kinds versus artifacts. For artifacts, the theory predicts that naming ability will not be strongly correlated with damage to semantic category structure, whereas for natural kinds the nature of the relationship will change as damage to the system progresses. To test this theory, young and elderly participants and patients with Alzheimer's disease named a series of pictures and completed a board sorting task, in which they placed words from a semantic category on a two dimensional grid in a way that represented their inter-similarities, thus reflecting the nature of their semantic knowledge. Results confirmed the prediction that a strong relationship between picture naming and disrupted category structure is evident only for natural kinds categories at later stages of damage. For natural kinds in earlier stages and artifacts throughout the progression of the disease, disrupted category structure is not directly reflected in naming performance. These data point to a complex relationship between the underlying category structure and its realization in naming ability.

Introduction

The investigation of the consequences of brain damage on specific semantic categories has revealed much about the normal organization of concepts and categories. Although there have been a wide array of category-specific deficits reported in the literature (e.g. Caramazza & Shelton, 1998 and Sheridan & Humphreys, 1993), the type that have generated the most interest are those that are domain specific, most typically affecting natural kinds (*e.g., animals*, *plants*) or artifacts (*e.g., vehicles*, *tools*). Each type of domain-specific deficit has become associated with damage to specific brain areas. Natural kinds deficits have been associated with damage to the bilateral antero-medial and inferior temporal lobe (Sartori, Job, Miozzo, Zago, & Marchiori, 1993), while artifacts are associated with damage to fronto-parietal regions (Sacchett & Humphreys, 1992; Warrington & McCarthy, 1983).

In an early account of category specific impairments, Warrington and McCarthy (1987) proposed that these deficits could be explained by damage to particular types of features that underlie the representations of natural kind or artifact categories. They argued that concepts that represent natural kinds are predominantly recognized based on their perceptual features (*e.g., zebras* and *horses* are differentiated by the feature *has-stripes*), while concepts that represent artifacts are predominantly recognized based on their functional features (*e.g., hammers* are differentiated from *saws* because they are used to pound things, whereas *saws* are used to cut things). By this account, individuals have category-specific deficits for natural kinds because of damage to areas responsible for storing and processing the visual semantic information (perceptual features) so crucial to natural kinds. In contrast, damage to areas representing functional features (presumably related to action and movement), will cause selective deficits for artifacts concepts. Additional support for feature-based theories comes from evidence for the anatomical localization of features (e.g., Martin, Haxby, Lalonde, Wiggs, & Ungerleider, 1995).

Such a feature-based account can explain why focal damage causes category specific deficits, as can accounts that posit anatomical localization of specific categories (e.g., Caramazza & Shelton, 1998). However, category specific deficits have also been reported in patients with Alzheimer's disease (AD), which is characterized by more patchy, widespread damage. Below we review the evidence for category specific impairments in AD and a theory proposed to account for these deficits.

Expanding the Featural Approach to Account for Category-Specific Deficits in AD

A number of studies have found category specific naming deficits in AD patients (e.g., Chan, Salmon, & De La Pena, 2001; Gonnerman, Andersen, Devlin, Kempler, & Seidenberg, 1997; Silveri, Daniele, Giustolisi, & Gainotti,

1991). These findings pose a problem not only for category localizing accounts such as Caramazza and Shelton's (1998), but also for feature based accounts that rely *solely* on the localization of features to explain category specific impairments. The question that arises is how the widespread damage characteristic of AD could affect only isolated brain regions (where either categories or features are localized, depending on the theory) and produce selective damage to natural kinds or artifacts. Why doesn't AD affect both domains equally?

To explain how patchy and widespread damage can cause domain-specific deficits, Gonnerman *et al.* (1997) proposed that, in addition to the difference between the principle type of features that define natural kinds and artifacts (Warrington and McCarthy, 1987), there is a critical difference between the two domains in terms of both the degree of feature intercorrelations and the nature of distinguishing features.[1] Features become intercorrelated when they occur together frequently such that one feature is predictive of another (*e.g., has wings* and *has a beak*). Unlike artifact concepts, natural kinds tend to share a large number of intercorrelated features with other category members (*e.g.,* many animals *have four legs* and *have tails*). These intercorrelations allow natural kinds to be more resilient than artifacts to limited damage, for two reasons: 1) because the proportion of intercorrelated to distinguishing features is so high, a small amount of damage is less likely to affect a distinctive feature; and 2) because intercorrelated features receive collateral support from all the concepts with which they are associated, damaged features within a particular semantic representation may be activated indirectly through collateral support from other features (see Devlin, Gonnerman, Andersen, & Seidenberg, 1998).

Progressive Semantic Damage in Natural Kinds versus Artifacts

These differences in the structure of semantic representations lead to different predictions about the effect of progressive semantic damage on the ability to name natural kinds versus artifacts. For a particular concept, damage that affects the availability of any single non-distinguishing feature will not necessarily cause a naming deficit for that concept. For example, losing the feature *grows-on-trees* for *apple* will not prevent an individual from naming an apple if other features (*is round, has a core, etc.)* are maintained, just as the loss of *has a windshield* should not prevent naming a car, if other features are available. In contrast, inability to activate a distinguishing feature can cause immediate naming problems (without *has stripes*, it is hard to differentiate a zebra from a horse). Because artifacts have a higher ratio of distinguishing to non-distinguishing features (Devlin *et al*. 1998) and fewer intercorrelated features than natural kinds, minimal damage is more likely to affect a distinguishing feature and thus interfere with naming ability. As damage progresses, there will be a linear increase in the number of artifact concepts affected, as more distinguishing features are lost. The pattern of decline for natural kinds looks very different; the intercorrelated features that characterize this category initially help support the maintenance of features and concepts, but increasing damage has a profound effect, as the network is no longer able to sustain these features and whole clusters of concepts that depend on them become unavailable.

Relationship Between Naming and Category Structure

Given this account of differences in the structure of natural kinds versus artifact categories, one would expect a different relationship between the degree of disruption to semantic knowledge and the ability to name individual concepts for each domain. Small amounts of damage to natural kinds may disrupt the category structure, but this will not be reflected in naming scores because of the collateral support provided by intercorrrelations. In effect, for natural kinds categories at low levels of damage, disordered category structure is masked by a preserved ability to name items. However, with increasing damage to the semantic system, sets of intercorrelated features will drop out *en masse*, resulting in a corresponding inability to name the entire set of items that relied on those intercorrelated features. These items may be a subset of a particular natural kind category, for example *four legged animals*, or they may affect a larger set of items such as *animals*. Thus, for natural kinds the theory predicts a weak correlation between naming ability and disruption of category structure at low levels of damage, and a strong correlation at higher levels of damage.

The picture is different for artifacts. Small amounts of damage within the artifacts domain are more likely to affect a distinguishing feature and thus cause a naming problem. Therefore, one would expect a correlation between the amount of damage and naming errors. However, since the errors are not related to sets of intercorrelated features as they are in natural kinds, the damage and the naming errors may be distributed across several artifact categories. For example, a patient may name a vehicle, a piece of furniture and a couple of tools incorrectly. If one then looks at the structure of the *vehicles* category, it would be relatively intact, even though the patient is making several naming errors. While damage and naming errors are more closely related throughout the progression of damage for artifacts than for natural kinds globally, examining any *single* artifact category may not reveal a strong correlation with naming ability.

Thus, our approach makes the following specific predictions. First, to the extent that AD patients have damaged semantic representations, their impairment should be reflected in reduced differentiation of items within a category. Second, since we believe that greater disruption of internal category structure will be reflected in more severe picture naming deficits, this reduced differentiation

[1] Distinguishing features are those features that serve to differentiate concepts within a semantic category. For example, the feature *has-stripes* differentiates *tigers* from *lions*.

should be even more apparent for categories in which they show naming difficulties. Third, the relationship between picture naming ability and impaired category structure will be the strongest for more severely impaired natural kinds categories. Thus, we expect a strong correlation between picture naming scores and loss of item differentiation for natural kinds categories, but only for those natural kinds categories for which AD participants show poor naming performance. In what follows, we describe a study designed to test these predictions.

Participants

The participants were 64 individuals who were paid for their participation. The young normal (YN) group consisted of 25 undergraduate students from the University of Southern California. The old normal (ON) group consisted of 24 elderly individuals. The group of patients with Alzheimer's disease (AD) included 15 individuals who were diagnosed with probable Alzheimer's disease using the NINCDS-ADRDA criteria (McKhann, Drachman, Folstein, Katzman, Price, & Stadlan, 1984). Results of neurological, laboratory (including computed tomography or magnetic resonance scan), and neuropsychological assessment failed to suggest other causes of dementia. All participants were native speakers of standard American English. See Table 1 for mean age and MMSE scores.

Table 1: Participants information

Group	Mean Age (SD)	Mean MMSE (SD)
YN	20.2 (2.4)	29.3 (0.6)
ON	78.1 (5.3)	29.1 (0.9)
AD	83.5 (3.8)	20 (3.0)

Methods

Picture Naming

Participants named items from twelve categories, six natural kinds and six artifacts, with twelve items in each category. Stimuli were controlled for familiarity, imageability (Snodgrass & Vanderwart, 1980; Wilson, 1988) and frequency (Francis & Kucera, 1982), as well as being matched for typicality across the two domains (Battig & Montague, 1969). 144 color pictures were then selected from various sources (e.g., graphics libraries). The pictures were displayed using the PsyScope experiment software (Cohen, MacWhinney, Flatt, & Provost, 1993) on a Macintosh Powerbook computer. Participants were asked to name each picture and responses were audio recorded and transcribed. Participants were allowed as much time as needed to respond to each stimulus.

Board Sort

The board sort task was adapted from one used by Ober and Shenaut (1999) as well as by Bonilla and Johnson (1995). Participants were presented with a set of printed words on 1" X 1" foam-board chips and instructed to study them and then place them on the board such that they represented the relationship between the words, with more similar items being placed closer together. The task was completed on a laminated board with a 10 x 10 square grid. The 12 word sets were the same as those used for the picture naming task. Each participant completed four boards, usually two selected from the natural kinds categories and two from the artifacts. In addition, participants completed a board with a set of 12 colored chips that did not have any words written on them. This color board functioned as a control to verify that the participants understood and were able to complete the task.

Results

Below we report first the results of the picture naming task, and of the board sort task. We then examine the relationship between picture naming and board sorting for both natural kinds and artifacts at different levels of performance.

Picture Naming

Naming responses were coded as correct, incorrect, or machine error. Synonyms were considered correct for the purpose of this study, since we are interested in participants' semantic knowledge rather than their word knowledge.

Results revealed that YN controls correctly named the pictures 86% of the time, ON controls 85% of the time, and AD patients 62% of the time (see Figure 1). T-tests were conducted comparing all three groups. To control for family-wise error, alpha was adjusted to .02 (alpha/c). The results demonstrate a significant difference in picture naming scores between YN and AD (t (15) = -4.15, p < .0009), and ON and AD (t (16) = 3.95, p < .002), but no significant difference between YN and ON (t (44)= -.48, p = .63).

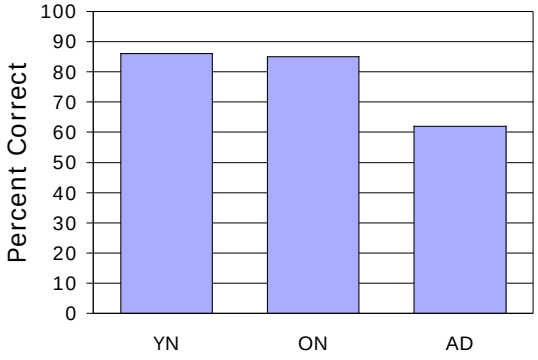

Figure 1: Percent correct on picture naming task by group: Young Normal (YN); Old Normal (ON); and Alzheimer's disease patients (AD).

Board Sort

To analyze the sorting data, each board was first converted into a set of 66 data points representing the Euclidian distances between any two chips on the board. The unit of

measurement used was board spaces, thus the closest together any two chips could be placed on the board was 1 space apart, and the farthest two chips could be from each other on the board was 12.73 spaces[2].

For each board there are several possible ways to group the items. For instance, a participant can group insects by manner of motion (which would have *flying* and *crawling* insects in separate groups), or they could group the items by number of legs (e.g., *spiders*, *ants*, and *flies* versus *caterpillars* and *centipedes*). In this respect, the task resembles traditional similarity judgments, where similarity can be judged based on a variety of features, not all of which are equally important (Medin & Goldstone, 1993). In both types of tasks what is important is the average "distance" between any two items, averaged across the different groupings.

Moreover, placement of items on the board may be affected by the extent to which individual participants choose to divide categories into multiple subgroups. For instance, the category *insects* can be divided into *flying insects* and *crawling insects*. *Crawling insects*, in turn, can be divided into crawling *insects with four or fewer legs*, and *crawling insects with more than four legs*. *Crawling insects with more than four legs* can be divided into *those with a countable number of legs*, and *those with too many legs to count*. Natural kinds categories tend to have a richer set of subgroups than artifact categories (see Garrard, Ralph, & Hodges, 2001). Each additional subgroup isolates its members (to some degree) from the other members of the category, thus increasing the mean distance between all the chips on the board (see Figure 2). Thus, we expect the natural kinds boards of normal participants to show greater overall mean chip distances than their artifacts boards. In addition, the boards of AD patients should differ from those of normal participants in that disrupted representations will lead to chips being placed close together, reflecting less fine differentiation within a category.

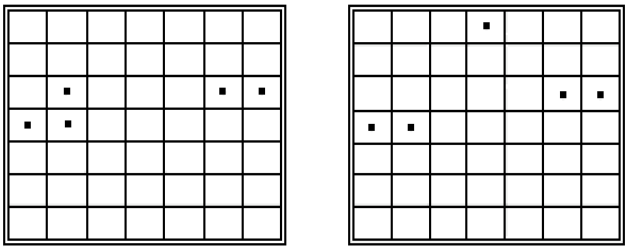

Figure 2: Two example boards. The one on the left contains two subgroups and has a mean distance of 3.38. The one on the right has the same number of chips but arranged in three subgroups and has a mean distance of 3.67.

[2] Scores are derived using the Pythagorean theorem. The board is square, and the distance of each side is 9, thus the diagonal distance from a chip in the bottom corner to one in the opposite top corner is the square root of 9^2 plus 9^2, or 12.73.

To compare the performance of AD patients to YN and ON controls, we ran a one-way, repeated measures ANOVA using the distances between word pairs as the dependent measure. The results revealed a main effect of Group, $F(2, 63) = 6.32$, $p < .003$ (see Figure 3). Planned comparisons revealed significant differences between all three groups, with YN having a higher mean chip distance than ON ($t = 11.9$, $p < .0001$), and ON having a higher mean chip distance than AD ($t = 6.42$, $p < .0001$).

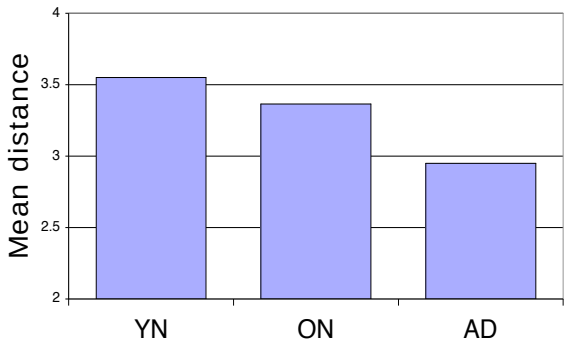

Figure 3: Board sort task: mean distances across 12 categories for each group: Young Normal (YN); Old Normal (ON); and Alzheimer's disease patients (AD).

We also analyzed the color board responses for comparison (see Figure 4). The results indicated that AD patients did not have a tendency to place chips closer together than YN/ON for color boards (YN mean: 2.7, ON mean: 3.11, AD mean: 3.1). We can therefore be confident that the AD patients are capable of performing the task and any differences between AD patients and normal controls are due to underlying differences in category structure and not simply task demands.

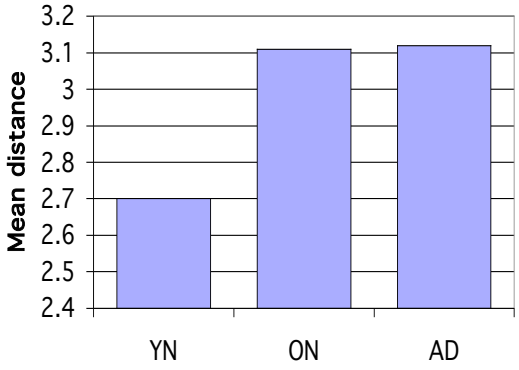

Figure 4: Mean distances for color board by group: Young Normal (YN); Old Normal (ON); and Alzheimer's disease patients (AD).

Our second prediction was that the mean chip distance should be greater for boards with items AD patients were able to name (Good boards) than for those boards with items AD patients were unable to name (Bad boards). The Good boards included the natural kinds and artifacts boards for which the patient made the fewest naming errors, and the

Bad boards were those for which they made the most naming errors. Thus, every subject had two Good boards and two Bad boards. As expected, the results demonstrate a significant difference between the two types of boards ($F(1, 14) = 8.47$, $p < .011$) with Good boards having a higher mean distance than Bad boards (see Figure 5).

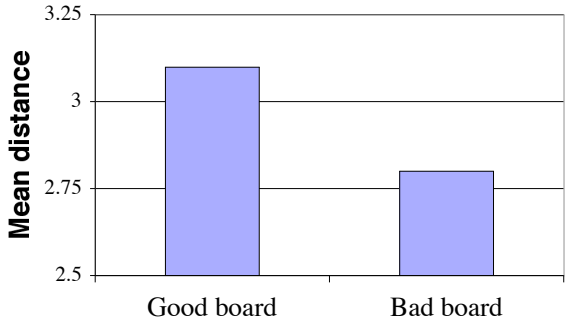

Figure 5: Alzheimer's patients' performance on board sort task: Mean chip distance for boards with corresponding high naming performance (Good boards) compared to boards for which patients made many errors naming the items (Bad boards).

Our final prediction was that semantic damage as reflected in the board sort task should be most highly correlated with naming at later stages of damage for natural kinds categories. To verify this, we calculated the correlation between mean distance on the board sort task and percentage of errors on the naming task. As predicted, only one group, the Natural Kind High Error group (*i.e.*, those that had a high percentage of errors for naming items in a natural kinds category) demonstrated a significant (inverse) correlation ($r = -.64$, $p < .007$). This means that as picture naming errors increased, the mean distance (*i.e.*, internal category structure) decreased (see Figure 6). Neither the Natural Kind Low Error, the Artifact High Error, nor the Artifact Low Error board sort distances showed any correlation with naming error ($r = -.2$ and $p = .56$, $r = -.36$ and $p = .25$, $r = -.3$ and $p = .36$, respectively).

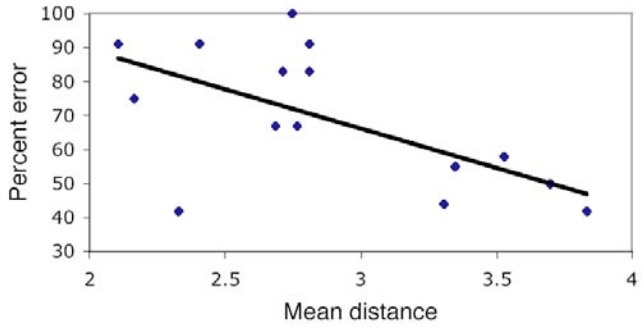

Figure 6: Correlation between mean distance measure from the board sort task and percent error for natural kind categories with high error rates.

Discussion

Our results confirm that AD patients have semantic deficits, which can be seen in picture naming and board sorting tasks. Contrary to Ober and Shenaut (1999), our data indicates that AD patients do have impaired semantic representations as demonstrated by their lower mean score compared to YN and ON on the board sort task. Although both studies shared a similar method, there was a crucial difference between how we analyzed our data and how they analyzed theirs. Ober and Shenaut compared the results from patients with Alzheimer's disease with those from normal controls using a pathfinder algorithm. This analysis provides a rank-based output. Because the pathfinder algorithm only demonstrates that, for example, *horse* is more related to *camel* than to *duck*, we decided to use a different analysis that would reveal more precise differences within a category than rank orderings allow.

In addition, our results show that good picture naming performance does not necessarily reflect preserved underlying category structure. The comparison of naming scores and category structure, measured by mean chip distances, confirmed our predictions. First, AD patients' performance differed significantly from that of both YN and ON controls; although they were not different from normal controls on the color boards, AD patients consistently placed their chips closer than YN and ON controls on the word boards. Thus, the disrupted internal category structure in the semantic systems of AD patients resulted in less fine differentiation of items within a category. In addition, this disrupted semantic structure was reflected in differences in naming performance, with more naming errors on items that were placed closer together on the board sort task.

Finally, as predicted, there was a weak, nonsignificant correlation between naming scores and category structure (measured by the board sort task) for artifact categories at all levels of naming performance. This is because increases in naming errors reflect damage from several categories across the artifact domain, but our board sort tasks looks at artifact categories one at a time. As for the natural kinds categories, when AD patients had only minor or no trouble naming, there was no significant correlation between the mean distance of the chips (*i.e.*, the amount of internal structure in the category) and the picture naming score. As proposed, this indicates that small amounts of semantic damage can affect the integrity of natural kinds categories as seen in the board sort task, while having only slight effects on the naming of the same words. However, since natural kinds concepts are largely dependent on intercorrelated features, once this core information is lost for the category, the high level of category damage should be reflected in poor naming performance for many items in the category. This is exactly what was seen, as evidenced by the strong correlation between picture naming scores and the category structures revealed in the board sort task for increasing levels of damage.

These results provide evidence for the importance of intercorrelated and distinguishing features, not only in

accounting for the semantic deficits in AD, but also in normal processing. They support a view where category structure can differ across categories because each category is dynamically shaped by its members. Our data suggest that natural kinds and artifacts categories differ, not only in the features that make up these categories, but also in the organization of the internal category structure. These findings also highlight the importance of task differences in revealing underlying damage to semantics and other aspects of language processing in AD (see also Kempler, Almor, Tyler, Andersen, & MacDonald, 1998).

Acknowledgments

This research was supported by NIA grant R01 AG-11774-04 and by NIH training grant 5T32MH20003-05. We thank the participants and their families for their generous contribution of time and effort. We would also like to thank Unja Hayes and Houri Hintiryan for their help with data collection and Alison Flipse, Neil Rampal, and Neva Ayn Rovner for their help in coding and verifying data.

References

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 Monograph*, 80 (3), 1-46.

Bonilla, J. L. & Johnson, M. K. (1995). Semantic space in Alzheimer's disease patients. *Neuropsychology*, 9 (3), 345-353.

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

Chan, A. S., Salmon, D. P., & De La Pena, J. (2001). Abnormal semantic network for "animals" but not "tools" in patients with Alzheimer's disease. *Cortex*, 37(2), 197-217.

Cohen, J. D., MacWhinney, B., Flatt, M. & Provost, J. (1993). PsyScope: An interactive graphic system for designing and controlling experiments in the psychology laboratory using Macintosh computers. *Behavior Research Methods, Instruments, and Computers*, 25, 257-271.

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.

Francis, W. N., & Kucera, H. (1982). *Frequency Analysis of English Usage: Lexicon and Grammar*. Boston: Houghton-Mifflin.

Garrard, P., Ralph, M. A. L., & Hodges, J. R. (2001). Prototypicality, distinctiveness, and intercorrelation: analyses of the semantic attributes of living and nonliving concepts. *Cognitive Neuropsychology*, 18(2), 125-174.

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

Kempler, D., Almor, A., Tyler, L. K., Andersen, E. S., & MacDonald, M. C. (1998). Sentence comprehension deficits in Alzheimer's disease: A comparison of off-line vs. on-line sentence processing. *Brain and Language*, 64, 297-316.

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

McKhann, G., Drachman, D., Folstein, M., Katzman, R., Price, D., & Stadlan, E. M. (1984). Clinical diagnosis of Alzheimer's disease: Report of the NINCDS-ADRDA work group under the auspices of Health and Human Services Task Force on Alzheimer's Disease. *Neurology*, 34, 939-944.

Medin, D. L., Goldstone, R. L. & Gentner, D. (1993). Respects for similarity. *Psychological Review*, 100(2), 254-278.

Ober, B. A., & Shenaut, G. K. (1999). Well-organized conceptual domains in Alzheimer's disease. *Journal of the International Neuropsychological Society*, 5(7), 676-684.

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

Sartori, G., Job, R., Miozzo, M., Zago, S., Marchiori, G. (1993). Category-specific form-knowledge deficit in a patient with herpes-simplex virus encephalitis. *Journal of Clincal and Experimental Neuropsychology*, 15(2), 280-299.

Sheridan, J., & Humphreys, G. (1993). A verbal-semantic category-specific recognition impairment. *Cognitive Neuropsychology*, 10, 143-184.

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

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 & Memory*, 6(2), 174-215.

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

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

Wilson, M.D. (1988) The MRC Psycholinguistic Database: Machine Readable Dictionary, Version 2. *Behavioural Research Methods, Instruments and Computers*, 20, 6-11.