To appear in *Proceedings of the 14th Annual Conference of the Cognitive Science Society*, Bloomington, IN, August, 1992.

Relearning after Damage in Connectionist Networks: Implications for Patient Rehabilitation*

David C. Plaut
Department of Psychology
Carnegie Mellon University
Pittsburgh, PA 15213–3890
plaut+@cmu.edu

Abstract

Connectionist modeling is applied to issues in cognitive rehabilitation, concerning the degree and speed of recovery through retraining, the extent of generalization to untreated items, and how treated items are selected to maximize this generalization. A network previously used to model impairments in mapping orthography to semantics is retrained after damage. The degree of relearning and generalization varies considerably for different lesion locations, and has interesting implications for understanding the nature and variability of recovery in patients. In a second simulation, retraining on words whose semantics are atypical of their category yields more generalization than retraining on more prototypical words, suggesting a surprising strategy for selecting items in patient therapy to maximize recovery.

Introduction

Cognitive neuropsychology aims to extend our understanding of normal cognitive mechanisms by studying their pattern of breakdown following brain damage in neurological patients. An underlying motivation for many researchers is that a more detailed analysis of the normal mechanism, and the way it is impaired in particular patients, should lead to the design of more effective therapy to remediate these impairments (Howard & Hatfield, 1987). Significant progress has been made in analyzing cognitive mechanisms and their impairments in terms of "box-and-arrow" information-processing diagrams, particularly in the domain of written language (Coltheart et al., 1980, 1987; Patterson et al., 1985). However, relatively few remediation studies have been based directly on these cognitive analyses, and while these few have been fairly successful, the specific contribution of the analysis is often unclear (for examples and general discussion, see Byng, 1988; Caramazza, 1989; Seron & Deloche, 1989; Wilson & Patterson, 1990). In large part the limited usefulness of box-and-arrow diagrams in this regard may stem from the general lack of attention paid to specifying the actual representations and computations that perform a task (Seidenberg, 1988).

Recently, a number of researchers employing connectionist models have attempted to go beyond the box-and-arrow approach by demonstrating that a fully-specified implementation of the normal process, when damaged, actually behaves like patients with analogous brain damage (e.g. Farah & McClelland, 1991; Hinton & Shallice, 1991; Mozer & Behrmann, 1990; Patterson et al., 1990; Plaut, 1991; Plaut & Shallice, 1991a, 1991b, 1992). This paper attempts to extend connectionist modeling in neuropsychology to address issues in cognitive rehabilitation. These issues concern degree and speed of recovery through retraining, the extent of generalization to untreated items, and how treated items can be selected to maximize this generalization.

The domain of investigation is impaired word reading, known as "acquired dyslexia." First, studies on remediation in acquired dyslexia based on cognitive models of normal reading are summarized, focusing on a study by Coltheart & Byng (1989) that attempted to reestablish the mapping between written words (orthography) and their meanings (semantics). A set of simulation experiments are presented in which a network, previously used to model impaired reading for meaning (Hinton & Shallice, 1991), is retrained after different lesions in which a proportion of the connections between groups of units are removed. The amount of recovery and generalization depends on the location of the lesion in the network and has interesting implications for understanding the effects seen in patients. The paper concludes with a second simulation demonstrating that retrain-

*I'd like to thank Marlene Behrmann and Geoff Hinton for their help with the research described in this paper. All of the simulations were run on a Silicon Graphics Iris-4D/240S using the Xerion simulator developed by Tony Plate. This research is supported by grant 87-2-36 from the Alfred P. Sloan Foundation, grant T89-01245-016 from the Pew Charitable Trusts, and grant ASC-9109215 from the National Science Foundation. Plaut (1992) presents an abstract of this work.

ing on words whose semantics are atypical of their category yields more generalization than retraining on more prototypical words, suggesting a surprising strategy for selecting items in patient therapy to maximize recovery.

Remediation of reading for meaning

Coltheart & Byng (1989) undertook a remediation study with an acquired dyslexic, EE, a 40-year-old left-handed postal worker who suffered left temporal-parietal damage from a fall. On the basis of a number of preliminary tests administered about 6 months later, they determined that EE had a specific deficit in deriving semantics from orthography. To improve the patient's word reading ability, Coltheart & Byng designed a study involving words containing the spelling pattern -OUGH (e.g. THROUGH, COUGH, BOUGH), which have highly irregular pronunciations and, thus, are difficult to read without semantics. EE was retrained on 12 of 24 such words, in which he studied the written words augmented with mnemonic pictures for their meaning (e.g. a picture of a tree drawn on the word BOUGH). Prior to therapy, four of the treated words were read correctly; after therapy, all 12 were read correctly. In addition, the *untreated* words also improved, from one correct prior to therapy, to seven correct after therapy. Thus, the improvement in the untreated set (6 words) was 75% as large as the improvement in the set that was actually treated (8 words). This generalization to untreated words is surprising because a word and its meaning are arbitrarily related—there is no intuitive reason why relearning the meanings of some words should help reestablish performance on other words with unrelated meanings.

In a second study, EE was given the 485 highest frequency words for oral reading. The 54 words he misread were divided in half randomly into treated and untreated sets. EE again learned to read the treated words by studying cards of the written words augmented with mnemonics for their meanings. As a result, his reading performance on the treated words improved from 44% to 100% correct. Once again, the untreated words also improved, from 44% to 85% correct (73% generalization). This improvement was not due to "spontaneous recovery" nor to other non-specific effects because performance on the words was stable both before therapy and after therapy.

Thus, in at least one patient, retraining the mapping from orthography to semantics for some words can generalize to other words. However, it should be noted that such improvement and generalization does not always occur. Some patients learn the treated items but show no generalization to untreated items. Others show generalization within but not between modalities. Still others may have difficulty learning the treated items themselves. For instance, Behrmann (1987) found no generalization from treated to untreated homophonic word pairs (e.g. RIGHT and WRITE) in the writing of acquired dysgraphic CCM, although the writing of 75 irregular words did improve significantly. Scott & Byng (1988) found that retraining the reading of homophonic word pairs of an acquired dyslexic, JB, generalized to reading untreated pairs but not to his *writing* of either treated or untreated pairs.

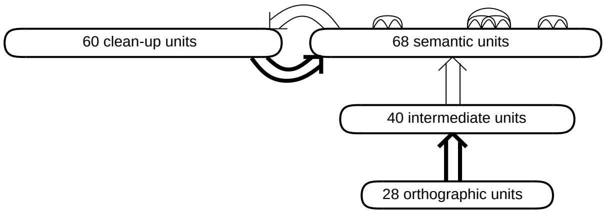

Figure 1: The network used by Hinton & Shallice (1991). Arrows in bold represent sets of connections that are lesioned in the present study.

Why some patients improve while others do not is not at all clear. An explanation of the effects seen in patient therapy in this domain should account not only for the occurrence of generalization in some patients and conditions, but also for its absence in others. Connectionist networks are proving useful in understanding the nature of impaired word reading—can they provide insight into the nature and variability of its recovery?

Modeling impaired reading for meaning

Hinton & Shallice (1991) have put forward a connectionist account of the process of accessing semantics from orthography, and the pattern of errors this process exhibits under damage. Based on previous work by Hinton & Sejnowski (1986), they trained a recurrent back-propagation network to map from the orthography of 40 three- or four-letter words to a simplified representation of their semantics, described in terms of 68 predetermined semantic features. The architecture of the network they used, shown in Figure 1, has two main pathways: (1) a "direct" pathway, from 28 orthographic units to 68 semantic units via 40 intermediate units, and (2) a "clean-up" pathway, from the semantic units to 40 clean-up units and back to the semantic units. The direct pathway generates initial semantic activity from visual (orthographic) input, while the clean-up pathway iteratively refines this initial activity into the exact correct semantics of the word.

After training the network, Hinton & Shallice systematically lesioned it by removing proportions of units or connections, or by adding noise to the weights. They found that the damaged network occasionally settled into a pattern of semantic activity that satisfied the response criteria for a word other than the one presented. These errors were more often semantically and/or visually similar to presented stimuli than would be expected by chance. While the network showed a greater tendency to produce visual errors (e.g. CAT ⇒ "cot") with lesions near the input layer and semantic errors (e.g. CAT ⇒ "dog") with lesions near the output layer, both types of error occurred for almost all sites of damage. This pattern of errors is similar to that of patients with deep dyslexia (Coltheart et al., 1980).

More recently, Plaut & Shallice (1991a, 1991b) have extended these initial findings in two ways. First, they established the generality of the co-occurrence of semantic, visual, and mixed visual-and-semantic errors by showing that it does not depend on peculiar characteristics of the network architecture, the learning procedure, or the way responses are generated from semantic activity. Second, they extended the approach to account for many of the remaining characteristics of deep dyslexia, including the effects of concreteness/imageability and their interaction with visual errors, the occurrence of visual-then-semantic errors, greater confidence in visual as compared with semantic errors, relatively preserved lexical decision with impaired naming, and the existence of different subvarieties of deep dyslexia.

The replication of the diverse set of symptoms of deep dyslexia through unitary lesions of the network strongly suggests that the underlying computational principles of the network capture important aspects of the process of mapping orthography to semantics in humans. Extending this claim further, we would expect relearning in the lesioned network to show similar effects to those observed in rehabilitation studies with analogous neurological patients. The following experiments test this claim.

Experiments in relearning after damage

A version of the Hinton & Shallice network was trained without momentum until it could read all 40 words perfectly (see Plaut, 1991, for details). The effects of lesions near orthography (orthography ⇒ intermediate connections) were compared with those of lesions within semantics (clean-up ⇒ semantics connections). For each of these two sets of connections, a severity of lesion was selected which lowered correct performance to near 20% (30% of orthography ⇒ intermediate connections, and 50% of clean-up ⇒ semantics connections).

For a given instance of a lesion, the responses to the 40 words were categorized as correct or incorrect. A response was considered correct if the proximity (i.e. normalized dot-product) of the semantics generated by the network was within 0.8 of the correct semantics of the presented word, and the proximity of the next best word was at least 0.05 further. Half of the correct words and half of the incorrect words were randomly selected and placed in the "treated" set; the remaining words were placed in the "untreated" set. Thus, both the treated and untreated sets always contained 20 words and were balanced for correct performance.

The lesioned network was then retrained for 50 sweeps on the treated words only. Performance was measured at each sweep during relearning separately for the treated and untreated word sets, in terms of the average percentage of words read correctly using the response criteria. The two sets were then exchanged and the retraining was repeated, starting from the same initial (lesioned) set of weights. Finally, the weights were again reinitialized and the lesioned network was retrained on all 40 words.

Figure 2 presents the retraining results for both locations of lesion, averaged over all 20 lesion instances and over exchanges of the treated and untreated word sets. First consider lesions within semantics (left of the figure). The treated words are quickly relearned by the network, with performance improving from near 20% to over 90% correct in under 20 sweeps through the word set. In addition, there is considerable generalization from the treated to untreated word sets (mean generalization 0.61, $t(39) = 28.1$, $p < .001$). Correct performance on the untreated words improves from 20% to 68% even though these words are never presented to the lesioned network. In fact, relearning on all of the words is quite dramatic, with performance recovering completely after 50 sweeps. These results replicate earlier findings on relearning and generalization in connectionist networks after corrupting weights with noise (Hinton & Plaut, 1987; Hinton & Sejnowski, 1986).

In contrast, retraining after lesions near orthography results in a quite different pattern of performance (see the right of Figure 2) . Relearning the treated words proceeds more slowly, with over 40 sweeps required to raise performance above 90%. Relearning all 40 words is even slower and more erratic. More importantly, there is no evidence of generalization to the untreated words—if anything, average correct performance on these words shows a trend towards getting slightly worse (mean generalization: -0.024,

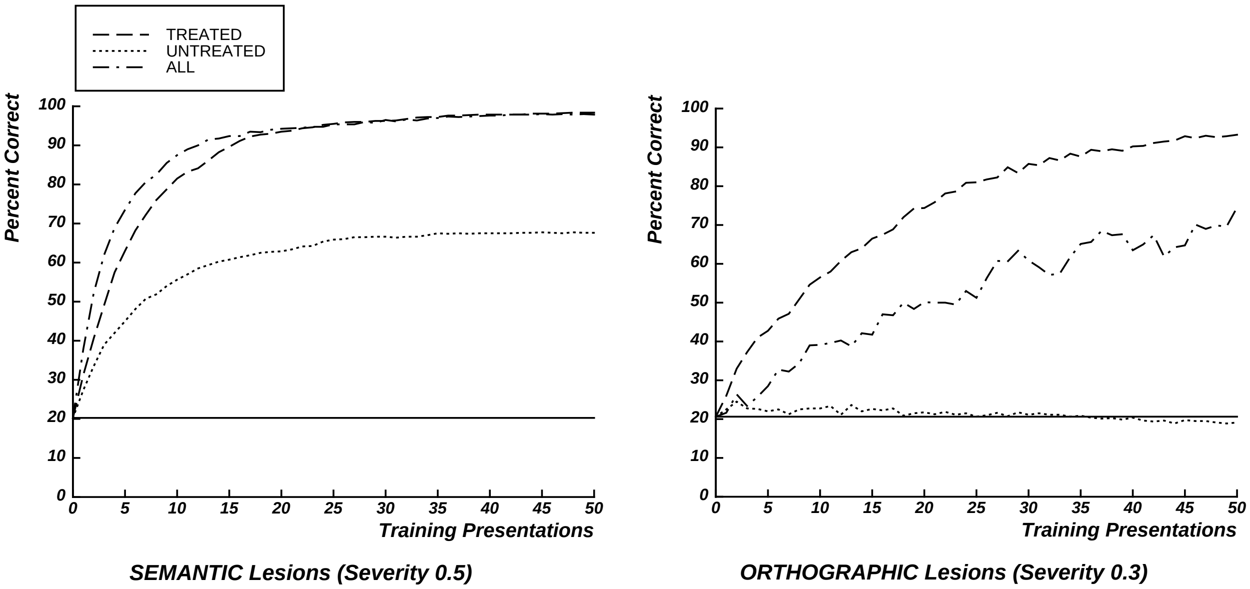

Figure 2: Retraining performance after clean-up ⇒ semantics lesions (left) and orthography ⇒ intermediate lesions (right). The solid horizontal lines represent the levels of performance at the onset of retraining.

$t(39) = 1.17$, $p = .25$).

Why does retraining after lesions within semantics yield rapid relearning and considerable generalization while retraining after lesions near orthography produces much worse relearning and no generalization? The degree of relearning and generalization depends on the consistency of the weight changes (i.e. directions of movement in "weight space") that would be optimal for individual words. While this is typically described in terms of the degree of overlap in the distributed representations of words, it depends more precisely on the consistency or structure in the mapping from input to output. Viewed as an abstract task, there is no systematic structure in mapping orthographic strings onto their semantics—input similarity is unrelated to output similarity. However, when instantiated in a network, the task is broken down by the learning procedure into a number of separate transformations involving intermediate representations carried out by different parts of the network. These transformations constitute "subtasks" that may differ considerably in their degree of structure. For example, the subtask of the clean-up ⇒ semantics connections is to refine the initial semantic activity generated by the direct pathway into the exact correct semantics of the presented word. Since semantically similar words require similar clean-up, this subtask is highly structured. In contrast, the subtask of the orthographic ⇒ intermediate connections is to generate intermediate layer representations that are as semantically organized as possible from visually organized inputs. Since semantic similarity is unrelated to visual similarity, there is no structure in this subtask. However, to the extent that the orthographic ⇒ intermediate connections succeed in generating semantically organized representations, the subtask of the intermediate ⇒ semantics connections becomes (semantically) structured. Consistent with this interpretation, relearning after lesions to these connections yields moderate but significant generalization (24%; see Plaut, 1991). Thus, the effectiveness of relearning after a lesion to a set of connections reflects the degree to which the mapping those connections carry out is structured.

As described above, studies of cognitive rehabilitation of acquired dyslexics in the domain of reading for meaning have demonstrated considerable relearning of treated items and (often) improvement on untreated but related items. Relearning after lesions to a network that operates in the same domain results in similar qualitative effects for lesions within semantics but not for lesions near orthography. Thus, at a general level, the cause of rapid relearning and generalization in the network—distributed representations and structure in subtasks—may provide an explanation for the nature of recovery, and lack of recovery, in these patients.

A specific hypothesis that comes out of the relearning simulations relates to the systematic differences observed in the degree of relearning and generalization as a function of lesion location. The simulations predict that a patient with a functional impairment close to or within semantics should show considerable generalization, while one with an impairment close to orthography should show little or none. Conversely, the degree of generalization observed in a patient can be used to predict the fine-grained location of their functional impairment *within* the mapping from orthography to semantics.

Designing therapy to maximize generalization

Ideally, we would like to use our understanding of the impairment in a particular patient to lead to the design of a rehabilitation strategy that maximizes recovery. The previous simulation clarifies the conditions under which retraining yields generalization. Under these conditions, how can items be selected for retraining so as to maximize this generalization? A critical variable in semantic representation is prototypicality—how close a concept is to the central tendency of its category (Rosch, 1975). The question is, is it better to retrain on prototypical or non-prototypical words?

Unfortunately, the limited size and complexity of the original training set precludes a reasonable comparison. Accordingly, a second simulation study was carried out, analogous to the first except that it involved 100 "words" whose orthographic and semantic representations were artificially generated. First, a single semantic "prototype" was created by randomly setting each of 50 semantic features to be present with probability $p = 0.2$. Two sets of 50 word meanings were generated from this prototype using different levels of random distortion (Chauvin, 1988). A "prototypical" set consisted of small distortions of the prototype (each feature of a word had a probability $d = 0.1$ of being randomly regenerated with $p = 0.2$). A "nonprototypical" set consisted of large distortions ($d = 0.5$). Orthography was represented as random patterns of activity ($p = 0.2$) over 20 input features. Using the same architecture and learning procedure as in the first study, a network was trained to generate the appropriate semantic features from each orthographic pattern. We investigated relearning after lesions to the intermediate $\Rightarrow$ semantics connections because they yielded only moderate generalization. Seventy instances of lesions of severity 0.25 reduced overall correct performance to 35.6% on average.

After each lesion, words were divided into prototypical and non-prototypical groups as described above, and then one group was further divided in half (balanced for correct performance). One of these halves formed the treated set, while the other formed one untreated set, and the words of the opposite type formed a second untreated set. The network was then retrained for 50 presentations of the treated set. Figure 3 presents the average generalization (i.e. ratio of untreated to treated improvement in correct performance, using a simple best-match criterion) from prototypical and non-prototypical treated sets to prototypical and non-prototypical untreated sets. Overall, retraining on non-prototypical words produces more generalization than retraining on prototypical words ($F(1, 69) = 337.4$, $p < .001$). The figure shows that this effect is due primarily to the fact that retraining on prototypical words significantly *reduces* performance on untreated non-prototypical words.

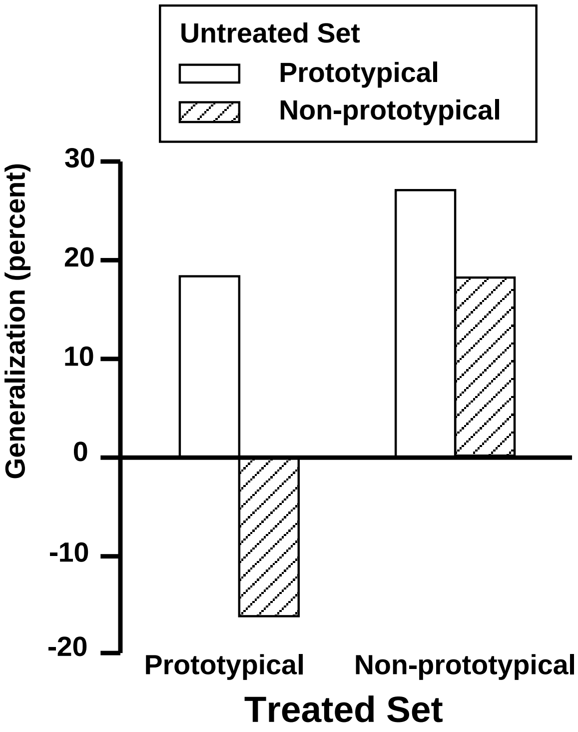

Figure 3: Generalization from prototypical or non-prototypical treated sets, to prototypical or non-prototypical untreated sets.

We can understand this effect by analogy with a set of randomly distributed points, where each point represents the effects of training on a particular word. The average of the outliers (non-prototypical words) may well-approximate the central points (prototypical words), but the average of the central points is still quite far from the outliers.

Summary

Theoretical analyses of cognitive impairments following brain damage should lead to the design of more effective strategies for rehabilitation. Simulations in this paper extend the relevance of connectionist modeling in neuropsychology to address issues in rehabilitation.

Attempts at cognitive rehabilitation of the mapping between orthography and semantics in patients have resulted in considerable improvement in performance on treated words, as well as significant generalization to untreated but related words, although the degree of recovery can vary considerably. The degree of relearning and generalization after damage in a network that performs the analogous task depends considerably on the location of lesion. These differences can be understood in terms of the amount of structure in the subtasks performed by parts of the network. The differences also provide a possible explanation for the variability in recovery observed in patients, and generate hypotheses about the specific location of their underlying functional impairment.

A potential benefit of connectionist modeling in neuropsychological rehabilitation is that it provides a framework for investigating the relative effectiveness of alternative rehabilitation strategies. A second simulation found that retraining on less prototypical words produced more generalization that retraining on more prototypical words, suggesting a surprising strategy for selecting items in patient therapy to maximize recovery.

Overall, the results demonstrate that investigations of relearning after damage in connectionist networks can provide an account of the general nature of relearning and generalization in patients and can generate interesting hypotheses about the design of effective patient therapy.

References

Behrmann, M. (1987). The rites of righting writing: Homophone remediation in acquired dysgraphia. *Cognitive Neuropsychology*, 4(3):365–384.

Byng, S. (1988). Sentence processing deficits: Theory and therapy. *Cognitive Neuropsychology*, 5(6):629–676.

Caramazza, A. (1989). Cognitive neuropsychology and rehabilitation: An unfulfilled promise? In Seron, X. & Deloche, G., editors, *Cognitive Approaches in Neuropsychological Rehabilitation*, chapter 12, pages 383–398. Lawrence Erlbaum Associates, Hillsdale, NJ.

Chauvin, Y. (1988). *Symbol Acquisition in Humans and Neural (PDP) Networks*. PhD thesis, University of California, San Diego.

Coltheart, M. & Byng, S. (1989). A treatment for surface dyslexia. In Seron, X. & Deloche, G., editors, *Cognitive Approaches in Neuropsychological Rehabilitation*, chapter 5, pages 159–174. Lawrence Erlbaum Associates, Hillsdale, NJ.

Coltheart, M., Patterson, K. E., & Marshall, J. C. (1980). *Deep Dyslexia*. Routledge, London.

Coltheart, M., Sartori, G., & Job, R. (1987). *The Cognitive Neuropsychology of Language*. Lawrence Erlbaum Associates, Hillsdale, NJ.

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.

Hinton, G. E. & Plaut, D. C. (1987). Using fast weights to deblur old memories. In *Proceedings of the 9th Annual Conference of the Cognitive Science Society*, pages 177–186, Hillsdale, NJ. Lawrence Erlbaum Associates.

Hinton, G. E. & Sejnowski, T. J. (1986). Learning and relearning in Boltzmann Machines. In Rumelhart, D. E., McClelland, J. L., & the PDP research group, editors, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Volume 1: Foundations*, chapter 7, pages 282–317. MIT Press, Cambridge, MA.

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

Howard, D. & Hatfield, F. M. (1987). *Aphasia therapy*. Lawrence Erlbaum Associates, Hillsdale, NJ.

Mozer, M. C. & Behrmann, M. (1990). On the interaction of selective attention and lexical knowledge: A connectionist account of neglect dyslexia. *Journal of Cognitive Neuroscience*, 2(2):96–123.

Patterson, K. E., Coltheart, M., & Marshall, J. C. (1985). *Surface Dyslexia*. Lawrence Erlbaum Associates, Hillsdale, NJ.

Patterson, K. E., Seidenberg, M. S., & McClelland, J. L. (1990). Connections and disconnections: Acquired dyslexia in a computational model of reading processes. In Morris, R. G. M., editor, *Parallel Distributed Processing: Implications for Psychology and Neuroscience*. Oxford University Press, London.

Plaut, D. C. (1991). *Connectionist Neuropsychology: The Breakdown and Recovery of Behavior in Lesioned Attractor Networks*. PhD thesis, School of Computer Science, Carnegie Mellon University. Available as Technical Report CMU-CS-91-185.

Plaut, D. C. (1992). Rehabilitating reading for meaning: Experiments in relearning after damage in connectionist networks. *Journal of Clinical and Experimental Neuropsychology*, 14(1):49.

Plaut, D. C. & Shallice, T. (1991a). Deep dyslexia: A case study of connectionist neuropsychology. Technical Report CRG-TR-91-3, Connectionist Research Group, Department of Computer Science, University of Toronto, Toronto, Ontario, Canada. Submitted to *Cognitive Neuropsychology*.

Plaut, D. C. & Shallice, T. (1991b). Effects of abstractness in a connectionist model of deep dyslexia. In *Proceedings of the 13th Annual Conference of the Cognitive Science Society*, pages 73–78, Hillsdale, NJ. Lawrence Erlbaum Associates.

Plaut, D. C. & Shallice, T. (1992). Perseverative and semantic influences on visual object naming errors in optic aphasia: A connectionist account. Technical Report PDP.CNS.92.1, Series on Parallel Distributed Processing and Cognitive Neuroscience, Department of Psychology, Carnegie Mellon University, Pittsburgh, PA. Submitted to *Journal of Cognitive Neuroscience*.

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

Scott, C. & Byng, S. (1988). Computer assisted remediation of a homophone comprehension disorder in surface dyslexia. *Aphasiology*.

Seidenberg, M. (1988). Cognitive neuropsychology and language: The state of the art. *Cognitive Neuropsychology*, 5(4):403–426.

Seron, X. & Deloche, G. (1989). *Cognitive Approaches in Neuropsychological Rehabilitation*. Lawrence Erlbaum Associates, Hillsdale, NJ.

Wilson, B. & Patterson, K. E. (1990). Rehabilitation for cognitive impairment: Does cognitive psychology apply? *Applied Cognitive Psychology*, 4:247–260.