Thesis Summary

Connectionist Neuropsychology: The Breakdown and Recovery of Behavior in Lesioned Attractor Networks

David C. Plaut
School of Computer Science
Carnegie Mellon University
dcp@cs.cmu.edu

September 1991
Thesis available as TR CMU-CS-91-185

Chapters 1 and 2: Introduction and background

Connectionist networks, also known as neural networks or parallel distributed processing (PDP) networks, are becoming increasingly influential in artificial intelligence and psychology as a methodology for developing computational models of cognitive processes. In contrast to conventional programs that compute via the sequential application of stored commands, these networks compute via the parallel cooperative and competitive interactions of a large number of simple neuron-like processing units. Learning involves modifying the weights on connections between these units to alter how their states influence each other. Networks of this form have been applied to problems in a wide range of cognitive domains, such as high-level vision and attention, learning and memory, language, speech recognition and production, and sequential reasoning (see McClelland et al., 1986, and recent Cognitive Science Society conference proceedings).

Beyond their computational properties, a major attraction of using connectionist networks to model cognitive processes is that their general similarity to neurobiology suggests that the nature of computation in these networks may provide insight into how cognitive processes are implemented in the brain (Sejnowski et al., 1989). Evidence that is often put forward in support of this claim is that, like brains, connectionist networks degrade gracefully under damage. That is, with partial damage, the network's performance on a task is only partially impaired rather than being completely lost. However, most demonstrations of this property have only considered the effects of damage on very general measures of performance, such as total error on a task. The relevance of connectionist modeling to the neural implementation of cognitive processes would be far better established if the specific effects of damage in networks were shown to be qualitatively similar to the specific effects of brain damage.

Some particularly encouraging preliminary work in this direction involves modeling the cognitive deficits of certain types of brain-injured patients by "lesioning" a connectionist model of the normal process (Farah & McClelland, in press; Hinton & Shallice, 1991; Mozer & Behrmann, 1990; Patterson et al., 1990). Most of these initial investigations have focused on deficits in word reading, known as "acquired dyslexias." The work of Mozer & Behrmann (1990) and Patterson et al. (1990) followed the standard approach in cognitive neuropsychology of using models

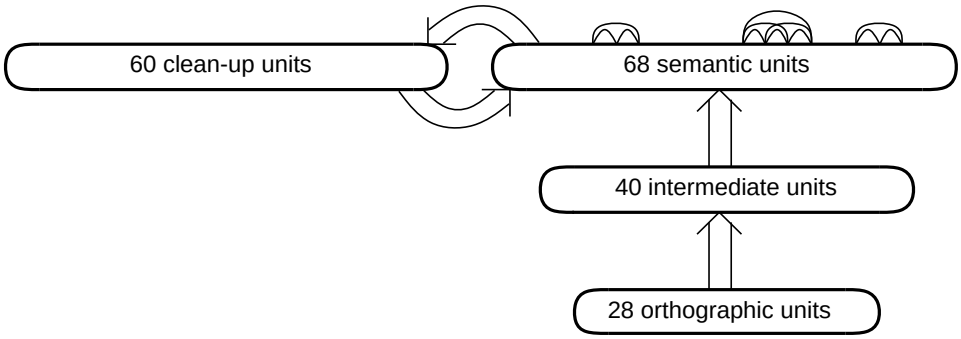

Figure 1: The network used by H&S. Arrows represent sets of connections that are lesioned.

of the normal reading process to account for disorders of reading as a result of damage. The work of Hinton & Shallice (1991, hereafter H&S) is somewhat different in nature. They were primarily concerned with "deep dyslexia," an acquired reading disorder in which patients can only pronounce written words by first accessing their meaning, and occasionally make errors in this process (e.g. misreading the word RIVER as "ocean"). In addition to these so-called "semantic" errors, a number of other error types occur: visual (SWORD $\Rightarrow$ "words"), derivational (GROWN $\Rightarrow$ "growing"), visual-and-semantic (TROUBLE $\Rightarrow$ "terrible") and visual-then-semantic (SYMPATHY $\Rightarrow$ "orchestra," presumably via *symphony*). Furthermore, these patients exhibit a variety of other symptoms, including an effect of part-of-speech in correct performance (nouns $>$ adjectives $>$ verbs $>$ function words), better performance on concrete, highly imageable words than on abstract, less imageable words, and an inability to read pronounceable non-words. Deep dyslexia poses a distinct challenge for cognitive neuropsychology because there is little understanding of why such a variety of symptoms should co-occur in virtually all known patients.

H&S put forward a connectionist account of the co-occurrence of semantic, visual, and mixed visual-and-semantic errors in deep dyslexia. They trained a network to map from the orthography (i.e. written form) of 40 words to a simplified representation of their semantics, described in terms of 68 predetermined semantic features. This task is unstructured in the sense that there is no intrinsic relationship between a word and its meaning—similarly spelled words need not have similar meanings. The architecture of the network, shown in Figure 1, has two main pathways: (1) a "direct" pathway, from orthographic units to semantic units via intermediate units, and (2) a "clean-up" pathway, from the semantic units to clean-up units and back to the semantic units. The direct pathway generates initial semantic activity from visual input. The clean-up pathway iteratively refines this initial activity into the exact correct semantics of the word. Only 25% of the possible connections in each pathway were included. However, additional direct connections among semantic units that represent closely related features were added. H&S then damaged the network in a variety of ways, and 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 error responses were related to the input semantically (i.e. in the same category) and/or visually (i.e. shared a letter) more often than would be expected by chance. In addition, mixed visual-and-semantic errors occurred more often than would be expected from the independent rates of visual errors and semantic errors. While the network showed a greater tendency to produce visual errors with damage near the orthographic layer and semantic errors with damage near the semantic layer, both types of error occurred for almost all sites of damage.

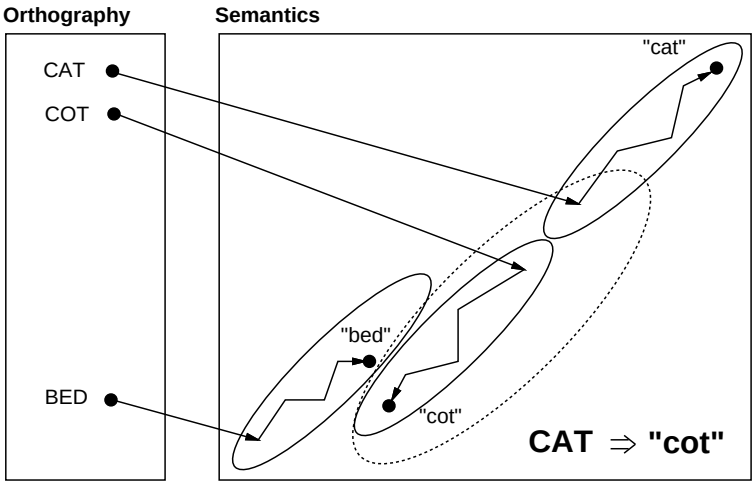

Figure 2: How damage to semantic attractors can cause visual errors. The solid ovals depict the normal basins of attraction; the dotted one depicts a basin after semantic damage.

An important concept in understanding these results is that of an "attractor." The semantic units in the network change their states over time in response to a particular orthographic input. The initial pattern of semantic activity generated by the direct pathway may be quite different from the exact semantics of the word. Interactions among semantic units, either via direct connections or via the clean-up units, serve to gradually modify and "clean-up" the initial pattern into the final, correct pattern. This process can be conceptualized as movement in a multi-dimensional space in which each dimension represents the state of a unit in the network. At any instant, the current pattern of activity of the network is represented as a particular point in this "state" space. When an input is presented to the network, the initial pattern of unit states constitutes the starting point. As units update their states, the point representing the current pattern of activity moves in state space, eventually arriving at the point corresponding to the correct interpretation of the input. In fact, there is a region in state space around this final point such that, if the network starts anywhere within this region, it will still settle into the same final pattern of activity. The point corresponding to this pattern is called an "attractor" in state space, and the surrounding region is called its "basin" of attraction.

H&S explain the co-occurrence of visual and semantic influences on errors in terms of the effects of damage in a network that builds attractors in mapping between two arbitrarily related domains. Connectionist networks have difficulty learning to produce quite different outputs from very similar inputs, yet very often visually similar words (e.g. CAT and COT) have quite different meanings. One effective way a network can accomplish this mapping is to construct large basins of attraction around each familiar meaning, such that any initial semantic pattern within the basin will move to that meaning (see Figure 2). Visually similar words are then free to generate fairly similar initial semantic patterns as long as they each manage to fall somewhere within the appropriate basin of attraction. In this way the network learns to shape and position the basins so as to "pull apart" visually similar words into their final distinct semantics. Damage distorts these basins, occasionally causing the normal initial semantic pattern of a word to be "captured" within the basin of a visually similar word. Essentially, the layout of attractor basins must be sensitive to both

visual and semantic similarity, and so these metrics are reflected in the types of errors that occur as a result of damage.

H&S were not concerned with supporting a particular model of normal word comprehension performance. Rather, they were investigating the effects of damage on the behavior of a more general type of network that maps between two arbitrarily related domains. By not being tied to a particular normal model, it becomes possible to systematically explore the space of models that qualitatively reproduce the neuropsychological phenomena. This allows the relevance and implications of various design decisions to be evaluated: the definition of the task, the specification of network architecture, the training procedure, and the evaluation of the behavior of the damaged network. To the extent that these design decisions arise out of general connectionist principles, an understanding of their impact in the particular domain of acquired dyslexia sheds light on the overall enterprise of connectionist neuropsychology.

This thesis investigates the breakdown and recovery of behavior in connectionist networks that develop attractors, in order to analyze their behavior more thoroughly and identify the computational principles that enable them to reproduce detailed neuropsychological phenomena. The term "connectionist neuropsychology" is intended to apply to the thesis research in two ways. The first is doing cognitive neuropsychology *with* connectionist networks—using them to understand patient behavior. The second is doing cognitive neuropsychology *on* connectionist networks—analyzing their impaired behavior under damage as a way of gaining insight into the nature of their own representations and processes. An important insight that emerges from the research is that the structure of a task has a profound influence on the layout of attractors in state space.

Each of the next four chapters investigates a particular design decision of the H&S model. The subsequent two chapters present extensions of the approach.

Chapter 3: Response generation: Mapping semantics to phonology

The first design issue we address is the way the responses of the network are generated and evaluated. Most data on deep dyslexic reading comes from tasks in which the patient produces a verbal response to a visually presented word. Since the output of the H&S model consists of semantic activity, an *external* procedure is required to generate explicit responses: the semantic activity produced by the damaged network is compared with the correct semantics of all known words, with the best-matching word being selected as long as the match is sufficiently good, and sufficiently better than any other match. The rationale for these criteria is that semantic activity that is too unfamiliar or ambiguous would be unable to drive an output system effectively, but this assumption was never verified.

Ideally, the response criteria would be replaced by extending the network to produce actual phonological responses. Unfortunately, initial simulations presented in this chapter demonstrate that, when such an output network is developed in a straightforward manner, damage to the input network results in a high rate of phonological responses that are novel "blends" of familiar responses (e.g. MUG (+CUP) $\Rightarrow$ "mup"). Deep dyslexic patients almost never produce this type of response. The network produces blends because, when the phonology of a word is trained as a response, other phonological patterns are also reinforced to the extent that they overlap with the trained pattern. As a result, the network develops "spurious" attractors for these phonological blends as it develops the appropriate attractors for word pronunciations.

One way to eliminate blends is to alter the training procedure for the output network so that the

attractor basins for words are much larger and stronger. Then, initial phonological patterns that are mixtures of the phonemes of two words are much more likely to fall into the attractor for one or the other of the words, rather than into a spurious attractor for a blend. We employ a training procedure that differs from the standard approach in two main ways: the use of noisy input, and incremental training. Training on noisy input amounts to enforcing a particular kind of generalization: inputs which are *near* known patterns must give identical responses. Thus the basin of attraction for each trained pattern must be at least large enough to include the patterns that can be generated from it with the amount of noise used during training. Incremental training (i.e. initially training parts of the network separately and then combining them) encourages each part to accomplish as much of the task as possible, without relying on the strengths of the other parts. Specifically, when training the complete network, if the direct pathway can generate reasonable phonology from even noisy semantics, there is less pressure on the phonological clean-up pathway to develop strong attractors for the correct patterns. Training them separately forces each of them to compensate for the noise *independently* so that their combination is more robust.

An output network developed with such a training procedure is combined with an input network to produce a complete network that produces explicit pronunciations of written words via their meanings (i.e. maps orthography to phonology via semantics). Lesions throughout the input network reproduce the co-occurrence of error types, with very low rates of blend responses, allowing explicit phonological responses to replace the best-match criteria H&S used. Further experiments demonstrate that lesions within the output network itself have qualitatively similar effects as input lesions. However, virtually no explicit error responses occur when the lesion involves the phonological clean-up pathway, supporting the claim that the operation of attractors is essential for reproducing the patient behavior.

Chapter 4: The relevance of network architecture

The second aspect of the design of the H&S model we evaluate is the network architecture, by which we mean the number of units and their connectivity. H&S provide only a general justification for the network architecture they chose. Hidden units are needed because the problem of mapping orthography to semantics is not linearly separable. Recurrent connections are required to allow the network to develop attractors, whose existence constitutes the major theoretical claim of the work. The choices of numbers of intermediate and clean-up units, restrictions on connections among semantic units, and connectivity density were an attempt to give the network sufficient flexibility to solve the task and build strong semantic attractors, while keeping the size of the network manageable.

In this chapter we investigate the effects of damage in a range of alternative network architectures, each intended to evaluate the relevance of a particular aspect of the original network. Specifically, we compare the impact of (1) the presence or absence of connections among semantic units, (2) the effect of 25% vs. full connectivity density, (3) whether the clean-up pathway that implements the attractors is located at, or prior to, semantics, and (4) whether the clean-up pathway is separated from the direct pathway, or combined by using a single group of hidden units for both functions. The lesioned behavior of each network is evaluated using both the H&S response criteria and an output network that generates explicit phonological responses. The results demonstrate that the existence of attractors is critical for reproducing the error pattern of deep dyslexics—it does *not* occur when there are no attractors "downstream" from a lesion to provide clean-up. However, the

architectural details that enable the attractors to develop are relatively unimportant. Quantitative differences in behavior among networks provide further evidence for the importance of attractors, with those that develop stronger attractors (e.g. due to full connectivity density) producing higher rates of correct and error responses compared with omissions (i.e. failures to respond).

One effect observed by H&S that appears to be less general than the simple co-occurrence of error types is the above-independent rates of mixed visual-and-semantic errors. This effect was observed only in networks in which the intermediate units between orthography and semantics were involved in developing the attractors. In these networks, the attractors lie at a stage where visual and semantic influences cannot be separated, and so damage reveals a bias towards responses that share both types of similarity with the input.

Additional simulations presented in this chapter address more detailed aspects of the pattern of correct and impaired performance shown to varying degrees by all of the networks. The specific issues that are investigated include (1) the distribution of correct performance and error types across semantic categories and individual words, (2) the impact of the definitions of visual and semantic similarity on the error pattern, (3) the occurrence of visual-then-semantic errors in the networks (e.g. BOG $\Rightarrow$ "cat" via *dog*), (4) the effects of lesion severity on the distribution of error types, and (5) the error patterns for individual lesions. The results serve to verify that the occurrence of a wide range of aspects of deep dyslexia in the networks are not due to idiosyncratic characteristics of the word set or interpretation procedure. They also demonstrate that the networks behave like deep dyslexic patients in terms of the pattern of responses after individual lesions.

Chapter 5: The relevance of learning procedure

The next design issue we address is the choice of learning procedure. H&S trained their model with a powerful but biologically implausible learning procedure known as "back-propagation through time." They justify its use in two ways. First, they emphasize that they were not directly concerned with simulating aspects of reading *acquisition*, but only its breakdown in mature, skilled readers. Thus the learning procedure can be viewed solely as a programming technique for determining a set of weights that is effective for performing the task. Second, they point out that back-propagation is only one of a number of ways of performing gradient descent learning in connectionist networks. Other more plausible gradient descent procedures are more computationally intensive than back-propagation but typically develop similar representations. However, this claim is rarely substantiated by a demonstration of the similarity between systems developed with alternative procedures.

In this chapter, we replicate and extend the H&S results using the more plausible gradient descent procedure of contrastive Hebbian learning in a deterministic Boltzmann Machine (DBM, Hinton, 1989; Peterson & Anderson, 1987). The architecture of the network is broadly similar to the back-propagation networks that have a phonological output system, except that there are no separate clean-up units. There is also full connectivity within each of the orthographic, semantic, and phonological layers, and all sets of connections between layers have full density and are bi-directional. An important property of the network not shared by the back-propagation networks is that it minimizes a well-defined measure, called "energy," when settling in response to a given input. Activity patterns with low energy satisfy many of the constraints encoded in the weights, and correspond to good interpretations of the input. Learning is also different in a DBM—rather than adjusting weights based on explicit error signals, contrastive Hebbian learning adjusts weights

based on differences in the co-occurrence of unit activities when the network is given both the input and correct output, compared with when it is given only the input.

Lesions throughout the DBM reproduce the co-occurrence of visual, semantic and mixed visual-and-semantic errors found in patients and in the back-propagation networks. The DBM produces low rates of omissions and blends, characteristic of networks with strong attractors. No special effort was made to encourage this—it arises naturally out of the nature of the formalism. Thus, the replication of the H&S results in a DBM lends credence to the notion that the pattern of results depends more on the *strength* of the attractors for words, than on the procedure by which those attractors are developed.

Training with noise introduces variability that encourages the development of strong attractors in the back-propagation networks. Additional simulations in this chapter investigate the effect of noise in the DBM framework by developing a network that is identical to the DBM except that it has stochastic units (McClelland, 1990; 1991). Surprisingly, damage to the stochastic network produces a virtually identical pattern of errors as damage to the deterministic one. However, rather than indicating that variability from noise conveys no advantage in developing strong attractors, these results suggest that the DBM already enjoys this advantage for other reasons: strong attractors develop in the DBM due to variability in settling for the same input *over successive weight changes* during learning.

Overall, these results validate the claim that the nature of the attractors developed using back-propagation have properties under damage that are similar to those developed using alternative, more biologically plausible formalisms. Beyond this, there are some aspects of the reading behavior of deep dyslexics that are much more effectively addressed using networks that have a well-defined energy measure. In particular, (1) patients have greater confidence in their incorrect responses when they make visual errors compared with semantic errors, and (2) in a "lexical decision" task, patients can reliably distinguish words they cannot read from visually similar non-words. In fact, the DBM settles more quickly, and into a lower energy between semantics and phonology, for visual errors as compared with semantic errors. In addition, if a letter string is presented to the network by providing the orthographic units with external input rather than fixing their states, the DBM derives more accurate orthographic representations for letter strings it has been trained on (words), as compared with other letter strings (non-words). Thus these differences in the network account for the effects seen in the patients.

Chapter 6: Extending the task domain: Effects of abstractness

The final design issue of the H&S model that we address is the definition of the task of reading via meaning. The original training set was too limited to address important issues known to influence the reading behavior of deep dyslexic patients. This chapter demonstrates that the approach taken by H&S can be extended to account for additional characteristics of deep dyslexic reading behavior. These involve the effects of the abstractness/concreteness of stimuli and responses, and the interaction of these effects with visual errors. Specifically, in patients: (1) concrete words are read better than abstract words, (2) visual errors are more likely on abstract stimuli then on concrete stimuli, (3) the responses in visual errors tend to be more concrete than the stimuli, and (4) the stimuli producing visual errors are more abstract than the stimuli producing semantic errors. Thus a semantic variable—concreteness—not only affects correct performance, but clearly influences the nature of *visual* errors.

There is a natural extension of the H&S semantic feature representation to express the distinction between concrete and abstract words. Jones (1985) showed that there is a very high correlation between concreteness and a measure, called "ease-of-predication," that reflects how easy it is to generate facts about a word. He further showed that the relative difficulty of different parts-of-speech in deep dyslexia maps perfectly onto their ordered mean ease-of-predication scores. Jones argued that the effects of both concreteness and part-of-speech in deep dyslexia can be accounted for by assuming that semantic representations of words reflect their ease-of-predication. Since semantic features loosely correspond to predicates, we realize the ease-of-predication distinction within the present framework by defining concrete words to have many more semantic features than abstract words.

To examine the effect of concreteness on visual errors, a set of 20 abstract and 20 concrete words were chosen such that each pair of words differed by a single letter. An input network, analogous to the H&S model, is trained with back-propagation to map from orthography to semantics. A similarly structured output network is trained separately to map from semantics to phonology. These two networks are then combined into a complete network that maps from orthography to phonology via semantics.

Overall, the behavior of the network under damage replicates, on a different word set, the mixture of error types found by H&S. In addition, lesions to the direct pathway of the network reproduce all of the effects of abstractness and their interaction with visual errors described above. However, severe lesions to the clean-up pathway produce opposite effects: *concrete* words are read more poorly and produce more visual errors. In fact, this type of lesion and pattern of performance are consistent with what is known about the single, enigmatic patient with "concrete word dyslexia" (C.A.V., Warrington, 1981). His reading disorder was quite severe initially, and he also showed an advantage for abstract words in picture-word matching tasks and with auditory presentation, suggesting modality-independent damage at the level of the semantic system.

These effects can be understood in the following way. As abstract words have fewer semantic features, they are less effective than concrete words at engaging the semantic clean-up mechanism and must rely more heavily on the direct pathway. Concrete words are read better under lesions to this pathway because of the stronger semantic clean-up they receive. In addition, abstract words are more likely to produce visual errors as the influence of visual similarity is strongest in the direct pathway. Slight or moderate damage to the clean-up pathway impairs what little support abstract words receive from this system, but also impairs concrete words, producing no relative difference. Under severe damage to this pathway, the processing of most concrete words is impaired but many abstract words can be read solely by the direct pathway, producing an advantage of abstract over concrete words in correct performance. In fact, an analysis of the network demonstrates that the clean-up representations of concrete words are stronger and more diverse than those of abstract words.

Chapter 7: Relearning after damage

The previous chapters demonstrate how computational principles central to connectionist modeling are useful in understanding and reproducing the detailed behavior of one class of neurological patients, deep dyslexics. A major motivation for many cognitive neuropsychologists is that a more detailed analysis of the breakdown of cognitive mechanisms due to brain damage may lead to the design of more effective therapy to remediate these impairments. In this chapter we extend the

relevance and usefulness of connectionist modeling in neuropsychology to address issues in the rehabilitation of cognitive deficits following brain damage. The main issues we address concern the degree and speed with which behavior can be reestablished as a result of therapy, the extent that recovery due to treatment of particular words generalizes to other words, and the possible bases on which to select words for therapy so as to maximize this generalization.

Attempts at cognitive rehabilitation of the mapping between orthography and semantics in patients have produced considerable improvement in performance on treated words, as well as significant generalization to untreated but related words (Behrmann, 1987; Coltheart & Byng, 1989; Scott & Byng, 1988). The degree of recovery and generalization due to therapy varied considerably across patients, although the comparison is difficult because different remediation techniques were employed.

In this chapter, two of the previously developed deep dyslexic networks are subjected to lesions that reduce correct performance to near 20%. For lesions near semantics, when the network is retrained on the entire set of words, correct performance improves much faster than when the network was originally learning the words and had reached the same overall level of performance. In addition, when the network is retrained on only half of the words (balanced for correct performance), its performance on the remaining words also improves considerably. 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, lesions near orthography produce much slower relearning and no generalization.

Relearning produces generalization to the extent that the weight changes induced by the retrained words approximate those that would be optimal for the unretrained words. In this way, generalization depends on the consistency or structure in the set of mappings from input to output. Viewed as an abstract task, there is no systematic structure in mapping orthographic strings onto their semantics. 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 subtasks can differ significantly in the extent to which they are structured. In particular, connections near the orthographic input implement a much less structured mapping than connections within the semantic clean-up system. Thus, the relative differences in the degree of relearning and generalization for different lesion locations can be understood in terms of the amount of structure in the subtasks performed by those parts of the network.

The differences in generalization due to lesion location also provide a possible explanation for the observed variability in the recovery of patients, and generate hypotheses about the specific location of their underlying functional impairment. In particular, 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 semantic route.

Ideally, we would like to use our understanding of the impairment in a particular patient to lead to the design a rehabilitation strategy that maximizes recovery. A potential benefit of connectionist modeling in neuropsychological rehabilitation is that it provides a framework for investigating the relative effectiveness of alternative rehabilitation strategies. One possibility was explored in another set of experiments, in which words were divided into retrained and unretrained sets based on how similar their semantic representation is to the prototype of their semantic category.

Retraining on non-prototypical words yielded more generalization to more prototypical words than *vice versa*. However, an additional experiment cast doubt on the interpretation of these results, and in general, the word set is too limited to support definitive implications for patient therapy.

Chapter 8: Visual object naming in optic aphasia

All of the simulations described thus far involve pronouncing written words via their meaning. This task can be thought of as a specific instance of the more general problem of recognizing and naming visual objects. This chapter demonstrates the generality of the principles that help explain the reading deficits in deep dyslexia, by extending them to account for the characteristics of a syndrome in the broader domain of visual object recognition and naming.

"Optic aphasics" are neurological patients with a selective deficit in visual object naming that does not appear to be caused either by impaired visual recognition or by impaired naming: visual recognition demonstrated by means other than naming (e.g. gesturing or categorization) and naming via modalities other than vision (e.g. by tactile input) are relatively intact. Somewhat reminiscent of deep dyslexics, the error responses of optic aphasics in visual naming are frequently semantically related to the stimulus (e.g. *violin* $\Rightarrow$ "trumpet") or both semantically and visually/morphologically related (e.g. *apple* $\Rightarrow$ "orange") but only rarely are they purely visually related (e.g. *rubber band* $\Rightarrow$ "bracelet"). In addition, a striking characteristic of optic aphasics is the frequent occurrence of "perseverations," in which the responses to previously presented objects interfere with naming the current object. For example, not only are the exact responses (correct or incorrect) to previously presented objects often given subsequently, but responses that are *semantically related* to previous objects are also given. Thus J.F. (Lhermitte & Beauvois, 1973) correctly named a picture of a *baby*, but then misnamed a *knife* as a "child."

Since, in previous simulations, the network is completely reset after the presentation of each input, accounting for the perseverative effects in optic aphasia requires an elaboration of the computational formalism itself. The approach we adopt involves introducing short-term weights that depend on the recent correlations between unit states. These weights augment the standard weights that are slowly modified over the course of learning. While there is independent computational and empirical motivation for introducing short-term interactions among units, their inclusion to explain perseveration in optic aphasia is rather *ad hoc*. For this reason, the simple occurrence of perseverations in the network is less interesting than the interactions of these perseverative effects with other aspects of the network's behavior, that are not inherent in the extension of the formalism.

A network similar to the H&S model, augmented with short-term correlational weights, is trained to generate semantic representations of 40 objects from high-level visual representations. The task is defined so that the relationship between visual and semantic representations is more structured for objects than it is for words. That is, objects that look similar (e.g. furniture) are more likely to have similar characteristics and functions than are words that happen to share letters. Under damage, the network produces high rates of semantic and mixed errors but relatively few purely visual errors, much like optic aphasics.

The predominance of semantic rather than visual errors can be understood in terms of the way the layout of semantic attractor basins is influenced by the greater structure in the mapping from visual to semantic representations for object vs. words. For words, there is a strong pressure to shape and position the basins so as to separate the initial semantic activity for visually similar words into their quite distinct, final semantics. As a result, there are large regions within state

space where the attractor basins for visually related words adjoin, providing ample opportunity for visual errors. In contrast, in object naming it is less common that visually similar objects must be separated into completely different semantics. Rather, similar inputs are free to have similar effects on the output, and so there is much less direct influence of (visual) input similarity as distinct from (semantic) output similarity in the layout of attractor basins developed by the network. Semantic attractors must still provide clean-up to compensate for variations in the initial semantic activity, due either to noise or to the short-term weights from the previous object. Since this clean-up is semantically structured, the naming errors produced after damage show predominantly semantic similarity.

The damaged network also reproduces the perseverative effects in the naming errors of optic aphasics. The observed rate of "pure" perseverations, in which the previous response (called the "prime") is repeated even though it bears no relation to the current stimulus (e.g. ("spoon") *table* $\Rightarrow$ "spoon"), is about 3 times the chance rate. The prime also biases the network towards responses that are semantically related to it (e.g. ("spoon") *table* $\Rightarrow$ "fork"). In fact, even within a category, the prime biases responses towards itself compared with other objects in the category (e.g. ("spoon") *fork* $\Rightarrow$ "spoon"). The strength of the perseverative influence is strongest when the prime is an error response and weakest when it is an omission. Also, perseveration increases as lesions move closer to semantics.

The effect of the short-term weights is to bias the network towards the most recently occurring pattern of activity, corresponding to the prime. This bias strengthens the attractor for the prime, enlarging its basin of attraction. These modifications of the layout of attractor basins are superimposed on those that occur as a result of damage. Perseverations occur because the enlarged basin for the prime is more likely to capture the initial semantics of other objects more frequently than are the basins for other objects. Furthermore, if the prime is an error response, its attractor basin is already likely to be dominant, making it that much more likely that the prime will be given as the response to other objects as well. Finally, a bias in the semantic clean-up towards the prime has a greater opportunity to produce a perseverative error when that clean-up is weakened by a lesion than when it is operating normally and damage affects only the direct pathway.

The simulations presented in the chapter focus on reproducing the pattern of semantic and perseverative influences on naming errors in optic aphasia, without directly addressing how such a system might also support preserved visual recognition and non-visual naming in the context of impaired visual naming. Thus the results do not constitute a complete simulation of optic aphasia, but merely a demonstration of principles that might lead to a more complete account. In fact, a number of different theoretical accounts have been given for the intact recognition with impaired naming of visual stimuli in optic aphasia, although none of them are completely satisfactory. The one most similar to our approach (Riddoch & Humphreys, 1987) locates the impairment between vision and semantics and challenges the claim that recognition is intact in these patients. We conclude the chapter with a related proposal for how a complete account of optic aphasia might be developed within the current framework.

Chapter 9: General discussion

The final chapter summarizes and evaluates the contributions of the thesis. We begin by discussing computational issues, focusing on the relationship between our work and other modeling efforts, and the nature of the principles that underlie the ability of networks to reproduce the characteristics

of deep dyslexia. From a computational perspective, the most important advance over related work is the development of networks that generate explicit phonological responses without the use of a best-match procedure. From an empirical perspective, the main result of the simulations is that the co-occurrence of semantic, visual, and mixed visual-and-semantic errors after unitary lesions is not due to any idiosyncratic characteristics of the original H&S network. Rather, it is remarkably general. In addition to holding for different lesion locations, as H&S found, it also holds for networks with different architectures, using different output systems, trained with different learning procedures, and performing different versions of the task. However, the existence of attractors is critical, and the high rates of mixed errors occur only when the attractors directly mediate between orthography and semantics.

On the basis of the current simulations, we therefore put forward a hypothesis on the properties of a system that give rise to the central characteristics of deep dyslexia. Those characteristics are the following.

1. Semantic, visual, mixed visual-and-semantic, visual-then-semantic, and other (unrelated) errors occur;
2. Concrete words are read better than abstract words;
3. Visual errors (i) tend to have responses that are more concrete than the stimuli, (ii) occur more frequently on abstract than concrete words, and (iii) have stimuli that are are more abstract than for semantic errors.

We claim that these characteristics generally occur if a system with the following properties is lesioned.

1. Orthographic and semantic representations are distributed over separate groups of units, such that similar patterns represent similar words in each domain, but similarity is unrelated between domains;
2. Connection weights are learned by a procedure for performing gradient descent in some measure of performance on the task of mapping orthography to semantics;
3. Mapping orthography to semantics is accomplished through the operation of attractors;
4. The semantic representations of concrete words are much "richer" than those of abstract words (i.e. contain considerably more features).

One proviso of this hypothesis is that the lesion does not directly affect any connections primarily concerned with implementing the attractors (e.g. the clean-up pathway).

The chapter continues by evaluating the degree to which these computational principles account for the full range of patient behavior. First, we summarize the aspects of deep dyslexia addressed by the current research that go beyond the H&S simulations: the effects of abstractness and their interaction with visual errors, the relative confidence in visual vs. semantic error responses, the relatively preserved lexical decision performance with impaired naming, the occurrence of visual-then-semantic errors, and the existence of different subvarieties of deep dyslexia. Then, we describe how the current framework might also account for a number of remaining empirical issues: the occurrence of derivational errors, the part-of-speech effects, the occurrence of associative semantic errors, the existence of patients who make semantic errors but not visual errors, the co-occurrence of impairments in writing, and the distinction between visual vs. phonological similarity in errors.

The current account differs from others provided for deep dyslexia—and with few exceptions (e.g. Miceli & Caramazza, 1990; Mozer & Behrmann, 1990), for cognitive neuropsychology as a whole—in providing a "principled account." By this, we mean (1) many aspects of the syndrome are explained from a common set of basic assumptions, rather than requiring specific extra assumptions for each aspect; and (2) the explanations are derived from the assumptions computationally rather than intuitively. We argue that the computational principles we use to account for deep dyslexia are orthogonal to claims (e.g. Coltheart, 1980; Saffran et al., 1980) that deep dyslexic reading reflects the operation of the undamaged right hemisphere. At a more detailed level of comparison, the most commonly used concept with some relation to an attractor is that of a "logogen" (Morton, 1969; Morton & Patterson, 1980). Logogens operate much like "localist" representations in connectionist networks (Feldman & Ballard, 1982; McClelland & Rumelhart, 1981), in which concepts are represented by single units. Thus, the relationship between attractors and logogens is much the same as that between distributed and localist connectionist representations.

We then evaluate the adequacy and promise of extensions of the approach. The two most obvious syndromes for which an analogous explanation could be given are the parallels to deep dyslexia in the auditory domain (deep dysphasia) and in writing (deep dysgraphia). We also summarize the strengths and limitations of the simulations on relearning after damage and modeling visual object naming in optic aphasia.

We conclude by considering more general issues, the first of which concerns the impact of connectionist modeling in neuropsychology. The usefulness of the notion of a "syndrome," as a collection of behaviors arising from a specific functional impairment, has been strongly criticized in cognitive neuropsychology (e.g. Caramazza, 1984). The current research demonstrates that principles other than locus of impairment may provide fundamental insights into the commonalities of patients who may differ in other respects.

In addition to using connectionist networks to explain patient behavior, a goal of the current research has been to extend our understanding of the representational and computational properties of the networks themselves by studying their behavior under damage. Our focus has been on investigating aspects of the development of the network that influence the layout of attractor basins in state space. In this regard, four sets of results demonstrate that the structure of a task, both within and between domains, interacts in important ways with other aspects of network design. These results involve: (1) the natural occurrence of "blend" responses under damage, (2) the above-independent rates of mixed visual-and-semantic errors, (3) the degree of generalization in relearning, and (4) the relative differences in visual error rates across tasks.

Finally, we propose a range of future work: implementing a "dual-route" reading model that can pronounce words based on spelling-to-sound correspondences as well as via semantics, a controlled rehabilitation study, modeling patients with modality-specific semantic impairments, and reproducing the letter-by-letter reading of pure alexics within a model of high-level visual-spatial processing.

In summary, this thesis investigates the breakdown and recovery of behavior in lesioned attractor networks, in order to analyze their behavior more thoroughly and identify the computational principles that enable them to reproduce specific neuropsychological phenomena. The research establishes the importance of attractors in reproducing the detailed pattern of behavior of deep dyslexic patients, and demonstrates that the structure of a task has a profound influence on the nature of the layout of these attractors in state space. The simulations also provide a principled account of the otherwise perplexing combination of symptoms exhibited by deep dyslexic patients, and extend

the relevance of connectionist modeling to issues in cognitive rehabilitation and visual object naming deficits. Taken together, the results of the thesis establish the viability of connectionist neuropsychology as a means of extending our understanding both of patient behavior and of the nature of computation in connectionist networks.

References

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

Caramazza, A. (1984). The logic of neuropsychological research and the problem of patient classification in aphasia. *Brain and Language*, 21:9–20.

Coltheart, M. (1980). Deep dyslexia: A right-hemisphere hypothesis. In Coltheart, M., Patterson, K. E., & Marshall, J. C., editors, *Deep Dyslexia*, chapter 16, pages 326–380. Routledge, London.

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.

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

Feldman, J. A. & Ballard, D. H. (1982). Connectionist models and their properties. *Cognitive Science*, 6:205–254.

Hinton, G. E. (1989). Deterministic Boltzmann learning performs steepest descent in weight-space. *Neural Computation*, 1(1):143–150.

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, Seattle, WA.

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.

Jones, G. V. (1985). Deep dyslexia, imageability, and ease of predication. *Brain and Language*, 24:1–19.

Lhermitte, F. & Beauvois, M. F. (1973). A visual-speech disconnexion syndrome: Report of a case with optic aphasia, agnosic alexia and colour agnosia. *Brain*, 96:695–714.

McClelland, J. L. (1990). The GRAIN model: A framework for modeling the dynamics of information processing. In Meyer, D. E. & Kornblum, S., editors, *Attention and Performance XIV*. Lawrence Erlbaum Associates, Hillsdale, NJ.

McClelland, J. L. (1991). Stochastic interactive processes and the effect of context on perception. *Cognitive Psychology*, 23:1–44.

McClelland, J. L. & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: Part 1. An account of basic findings. *Psychological Review*, 88(5):375–407.

McClelland, J. L., Rumelhart, D. E., & the PDP research group (1986). *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Volume 2: Psychological and Biological Models*. MIT Press, Cambridge, MA.

Miceli, G. & Caramazza, A. (1990). The structure of orthographic representations in spelling. *Cognition*, 37:243–297.

Morton, J. (1969). The interaction of information in word recognition. *Psychological Review*, 76:165–178.

Morton, J. & Patterson, K. (1980). A new attempt at an interpretation, or, and attempt at a new interpretation. In Coltheart, M., Patterson, K. E., & Marshall, J. C., editors, *Deep Dyslexia*, chapter 4, pages 91–118. Routledge, London.

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., 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.

Peterson, C. & Anderson, J. R. (1987). A mean field theory learning algorithm for neural nets. *Complex Systems*, 1:995–1019.

Riddoch, M. J. & Humphreys, G. W. (1987). Visual object processing in optic aphasia: A case of semantic access agnosia. *Cognitive Neuropsychology*, 4(2):131–185.

Saffran, E. M., Bogyo, L. C., Schwartz, M. F., & Marin, O. S. M. (1980). Does deep dyslexia reflect right-hemisphere reading? In Coltheart, M., Patterson, K. E., & Marshall, J. C., editors, *Deep Dyslexia*, chapter 17, pages 381–406. Routledge, London.

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

Sejnowski, T. J., Koch, C., & Churchland, P. S. (1989). Computational neuroscience. *Science*.

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