Chapter 9

General discussion

Connectionist networks would appear *a priori* to be an appropriate formalism within which to develop computational models of neuropsychological disorders. Although the specific relationship between these networks and neurobiology is far from clear (Sejnowski et al., 1989; Smolensky, 1988), the belief that representation and computation in these networks directly resembles neural computation at some level remains one of their strongest attractions. In fact, the degree to which the behavior of connectionist networks after damage resembles that of neurological patients supports the claim that the apparent similarity is, in fact, substantial. Furthermore, studying the breakdown and recovery of behavior in damaged networks can shed light on their normal computational characteristics.

Connectionist modeling is most interesting when the formalism significantly contributes to a natural explanation for empirical phenomena that are counterintuitive when viewed within other formalisms. In the thesis, we focus on deep dyslexia, a neurobehavioral disorder in which patients exhibit a wide variety of symptoms in oral reading and related tasks, the most notable being the production of semantic errors. While the syndrome can certainly be described in terms of impairments within traditional "box-and-arrow" information-processing models of reading, such accounts offer little in the way of underlying principles that explain why such a diverse set of symptoms should co-occur in virtually all known patients who make semantic errors. Hinton & Shallice (1991) offer a connectionist account in which the central aspects of deep dyslexia—the existence of semantic errors and their co-occurrence with visual and mixed visual-and-semantic errors—arise naturally as a result of damage to a network that builds attractors in mapping orthography to semantics. While the approach has the advantage over traditional models of being far more computationally explicit, it has the limitation that there is little understanding of the underlying principles of the model which give rise to its behavior under damage. The current research involves a set of connectionist simulation experiments aimed both at developing our understanding of these principles, and at extending the empirical adequacy of the approach on the basis of this understanding. The results demonstrate the usefulness of a connectionist approach to understanding deep dyslexia in particular, and the viability of connectionist neuropsychology in general.

In this final chapter, we begin by discussing computational issues, focusing on the relationship between our work and other modeling efforts, and the nature of the principles that underly the ability of networks to reproduce the characteristics of deep dyslexia. We then turn to empirical considerations, evaluating the degree to which these computational principles account for the full range of patient behavior. The relationship between the current approach and other theoretical accounts of deep dyslexia is considered next. We then evaluate the adequacy and promise of extensions of the approach, relating to issues in rehabilitation and in modeling object naming deficits. We conclude by considering more general issues regarding the impact of connectionist modeling in neuropsychology, the importance of what can be learned about connectionist networks from their behavior under damage, future directions of research, and final conclusions.

9.1 Computational generality

Most connectionist efforts in modeling acquired dyslexia (e.g. Mozer & Behrmann, 1990; Patterson et al., 1990) have followed the standard approach in cognitive neuropsychology of using a particular model of normal reading to account for disorders of reading as a result of damage. In contrast, H&S never intended their model to be anything but the coarsest approximation to the mechanism by which normal subjects derive the meanings of words. Rather, their network was intended to embody particular computational principles, involving distributed representations and attractors, that were claimed to underly the effects seen in patients. In this way, the H&S model was put forth as representative of a wide class of models, all of which share the same basic principles but differ in other respects, and all of which, it was implicitly claimed, would show the characteristics of deep dyslexia under damage. However, H&S did not demonstrate that models which lacked the properties they claimed were central would *not* show the characteristics of deep dyslexia, nor did they investigate the actual nature and scope of the class of models that would. The present research is aimed, in part, at clarifying exactly what aspects of the original model are responsible for its similarity under damage to deep dyslexic patients, and what aspects are less central. To this end, simulations were carried out that explored the implications of each of the major design decisions that went into the H&S model: the definition of the task including the representation of the orthographic input and semantic output, the specification of network architecture, the use of a particular training procedure, and the means by which the performance of the network is evaluated.

9.1.1 Response generation

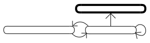
From a purely computational point of view the current simulations represent an advance over related work in some respects. The most important of these is the development of networks that generate explicit phonological responses without the use of a best-match procedure. Connectionist networks typically produce as output patterns of activity—that is, vectors of real numbers—in response to input. When using a network to model the reading behavior of normal or impaired subjects, what is compared with subject behavior is not the behavior of the network *per se*, but the behavior of the network *together with a procedure for interpreting vectors of real numbers as overt responses*. When the two together behave similarly to subjects, it is typically the network alone that is put forward as the explanation. However, there is always the issue of the extent to which the results depend on characteristics of the interpretation procedure. For this reason, if we wish to ascribe the modeling success to properties of the network, it is important that the interpretation procedure be neutral with respect to the observed effects and be as simple as possible.

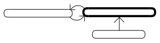
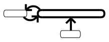
Most connectionist modeling work, including H&S, uses a best-match interpretation procedure, in which the output of the network is compared with all of the outputs it has been trained to produce, with the nearest one being selected as the overt response. These comparisons require a significant amount of knowledge about the task and can be rather involved—in fact, the ability of connectionist networks themselves to perform a best-match (categorization) operation is often put forward as a significant strength of the approach. The use of a simple error score (Seidenberg & McClelland, 1989) has the same failing as it requires knowledge of the correct response. The problem is particularly acute when a distributed output representation is used. A best-match procedure hides much of the difficulty of deciding on one of the 2^n possible binary responses over n output units given limited training data. In this way, the production of legal but unfamiliar and inappropriate responses, such as “blends,” goes unnoticed—but avoiding the problem by sidestepping the difficulty of generating a coherent response in a distributed representation is far from satisfactory.

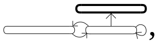
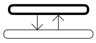
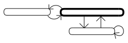
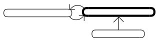
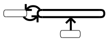
Our procedure for interpreting phonological output does not require any knowledge about the particular words on which the network has been trained. However, it does embody phonological knowledge about what constitutes a legal pronunciation. Since the set of legal pronunciations is far greater than the set of *familiar* ones, our interpretation procedure involves many fewer constraints, and hence much less knowledge, than one based on the training set. In fact, the DBM results showing the lack of importance of a probability criterion for individual phonemes suggests that very simple phonological knowledge—one phoneme active in each position—suffices.

9.1.2 The importance of attractors

The main empirical result of the simulation experiments is clear: 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, perhaps disturbingly so (see Section 9.5 below). 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. These results were shown not to be due to idiosyncratic effects of particular words, or of our procedure of averaging results over different instances of lesion. The generality of the effects argues against the possible criticism (e.g. Massaro, 1988) that the original results were due to the sophisticated manipulation of parameters that could have produced *any* observed phenomenon. Clearly the results do not depend on the detailed aspects of the model that were under the direct control of the experimenters.

However, if the co-occurrence of error types held under *all* conditions, we still could not infer what principles are responsible for them. In fact, among the simulations that were run, there were some conditions under which the mixture of error types did not occur. The most basic of these is where there are no attractors downstream from a lesion to provide clean-up. This was observed for I⇒S lesions in the  network, and for lesions to the phonological clean-up pathway in both of the output networks (with and without intra-phoneme connections). Under these conditions, the networks produced virtually no explicit error responses, even though correct performance may still be reasonable. Furthermore, the strong correlation between correct rate and explicit error rate across all of the simulation conditions demonstrates that the processes that underly correct performance in the normal network—attractors—are also responsible for the error responses in the damaged network. This provides strong evidence for H&S's claim that attractors are essential to produce the effects observed in their network.

While the *existence* of the various error types held across a wide variety of conditions, their quantitative distribution varied considerably over lesions in different locations in different networks. There were general trends of higher proportions of visual errors for lesions near orthography, and higher proportions of semantic errors for lesions near or within semantics. In fact, some lesions within semantics produced virtually no purely visual errors, although semantic and mixed visual-and-semantic errors still occurred (e.g. C⇒S lesions in the replication of the H&S network and the  and  networks, when using the response criteria). In this way, the systematic variation of proportions of error types in the model offers the possibility of accounting for similar systematic differences observed in patients (e.g. "input," "central," and "output" deep dyslexics, Friedman & Perlman, 1982; Shallice & Warrington, 1980) while still demonstrating the basic commonalities of all of these patients (see Section 9.2.1 below).

One effect observed by H&S that appears to be less general is that of higher rates of mixed visual-and-semantic errors than predicted by the independent rates of visual errors and semantic errors. When the pressure to build strong attractors was increased by training with noisy input, this effect was observed only in networks in which the intermediate units between orthography and semantics were involved in developing attractors (i.e. the , , and  networks). The mixed rate was not higher than predicted in networks in which the attractors operated separately from, and subsequently to, the direct access of semantics from orthography (i.e. the  and  networks). To the degree that patients exhibit a sufficiently high rate of mixed visual-and-semantic errors, the results place constraints on the nature of network architectures that can account for these effects. The non-generality of this effect also emphasizes the necessity of exploring a range of models that vary systematically from a particular model that shows some effect. It is difficult to determine which empirical results are robust and which are not on the basis of intuitions alone.

A potential limitation of the original H&S work that has not been addressed in subsequent simulations is the possible effects of using such a small training set. Although we demonstrated that the basic effects hold for two separate word sets—the original set and the abstract/concrete set—both sets contain only 40 words. The question arises as to whether the results are strongly biased by this limitation. In fact, Seidenberg & McClelland (1990) have argued that many of the limitations of their model are due to the fact that it was only trained on about 2900 words. However, there are significant differences between the tasks that the two models perform that provide reasonable justification for the reliability of effects produced in the current networks with only 40 words. Mapping directly from orthography to phonology involves learning statistical relationships among mappings that can then be applied to novel inputs in reasonable ways. Thus, a large number of training cases are required to estimate these statistics reliably, and performance would be expected to improve with a larger training set. In contrast, mapping from orthography to semantics involves *overcoming* statistical regularities, since visual similarity is not predictive of semantic similarity. It is true that a small training set limits the range of similarity that can be expressed *within* orthography or semantics, but it is unlikely to fundamentally alter the nature of the mapping between them. Thus the small size of the word sets prevented us from investigating the effects of variables such as frequency and syntactic class that are known to significantly influence deep dyslexic reading, and these issues remain open for future research. However, the basic findings of mixtures of error types would still hold if a much larger set of words were used.

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

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).

9.2 Empirical adequacy

9.2.1 Extensions of the Hinton & Shallice results

The H&S simulation was concerned with only some of the properties of deep dyslexia. A major strand of the current investigation was to explore whether other characteristics of the disorder would also be observed when a connectionist network that mapped orthographic to semantic representations was lesioned.

Three issues were specifically addressed: the effects of abstractness/concreteness, how confidence relates to error type, and lexical decision. Information relevant to a fourth issue—visual-then-semantic errors—came to light in the course of the study. A fifth issue—the different subvarieties of deep dyslexia—was indirectly confronted when the problem of generating a lexical phonological output was tackled. It should be noted, though, that our investigations of these five issues were not carried out with the same wide range of simulations as was done with regard to the more basic effects.

Effects of abstractness

In the simulation described in Chapter 6, an additional assumption was made, following Gentner (1981) and Jones (1985), that concrete nouns have a “richer” semantic representation than do other words. Specifically, the number of dimensions on which the semantic representation of a word has a specific value independent of the values it has on other dimensions is assumed to be greater for concrete nouns than for other words. This corresponds in our model to concrete nouns having more semantic features than do abstract nouns.

When this assumption is made, lesions to the direct pathway of the input network lead to an advantage in correct performance for concrete over abstract words.[1] It appears that the greater number of active semantic features gives the clean-up circuit more raw material on which to work,

[1] In further experiments not reported in this thesis, lesions to the output network also resulted in better correct performance on concrete vs. abstract words, although the difference was not as large as for input lesions.

allowing stronger attractors to be built. The magnitude of the effect in the network is not quite as large as that shown in some deep dyslexic patients, where patients such as D.E. (Patterson & Marcel, 1977) and K.F. (Shallice & Warrington, 1975) can show a $(C - A)/(C + A)$ ratio of 0.75 or 0.68 (where C and A are the correct rates on concrete and abstract words, respectively). Values approaching 0.5 were the largest obtained in the simulation, but a quantitative difference of this sort is not unexpected given the great difference in scale between the model and the human cognitive system.

More surprising than the mere existence of an abstract/concrete effect is the fact that it interacts with the occurrence of visual errors in a similar way to that found in most deep dyslexic patients in whom it has been investigated. After lesions to the direct route in the network, visual errors on average occur on more abstract words than do semantic errors, and the responses of visual errors tend to be more concrete than the stimuli. The one patient who differed in this respect was G.R. (Barry & Richardson, 1988). Like the simulation, G.R. produced visual errors much more frequently on abstract words, but the stimuli producing visual errors and semantic errors were roughly equally concrete. However, G.R. made semantic errors in matching spoken as well as written words to pictures (Newcombe & Marshall, 1980a). His impairment would therefore seem to involve the semantic system itself, which, when lesioned, might be expected to give rise to a higher number of semantic errors, even for concrete words.

Better performance in reading concrete than abstract words is not always found in acquired dyslexic patients. Warrington (1981) reported a patient, C.A.V., who read abstract words significantly better than concrete words, although the difference (55% vs. 36%) was not as dramatic as the complementary contrast found in certain deep dyslexic patients. The apparent double dissociation of concrete vs. abstract word reading between C.A.V. and deep dyslexics is difficult to account for without resorting to the rather unpalatable position that the semantics for concrete and abstract words are *neuroanatomically* separate (Warrington, 1981). The simulation provides an alternative explanation. Severe lesions to the clean-up pathway lead to an abstract word superiority which is, though, smaller than the concrete word advantage obtained from lesions to the direct pathway.

The difference between our explanation and Warrington's is subtle but important. Since in our simulations we allow damage to impair the direct and clean-up pathways independently, we are implicitly assuming that these pathways are anatomically separate (in Warrington's sense). However, it is *not* the case that the direct pathway processes abstract semantics while the clean-up pathway processes concrete semantics. The entire network is involved in generating the semantics of both concrete and abstract words. Rather, the direct and clean-up pathways serve different roles in this process, and these roles are differentially important for reading these two classes of words. As in Warrington's account, the dissociations arises from the selective impairment of a specialized process, but the specialization is not in terms of the surface distinction (i.e. concrete vs. abstract words) but rather in terms of underlying representational and computational principles (e.g. the influence of differing number of semantic features on the development of attractors).[2]

The fact that the model is consistent both with patients showing a concrete word advantage and with patients showing an abstract word advantage may suggest to some readers that the model is underconstrained by the data. There are three possible replies. First, overall, both patients and the model show a concrete word superiority. Second, for both types of superiority, the model predicts that visual error responses will tend to come from the class of words that are read more accurately. As predicted, C.A.V.'s visual error response were more *abstract* than the stimuli (Warrington, 1981). Finally, the model predicts that the complementary patterns would differ on other characteristics, corresponding to the different effects of direct vs. clean-up pathway lesions. C.A.V. also showed an advantage in matching auditorily-presented words with pictures, suggesting modality-independent damage at the level of the semantic system. Thus, there are additional aspects of our simulation that counter the challenge that it is underconstrained. However, given the uniqueness of concrete word dyslexia in C.A.V., its occurrence in the model should be considered suggestive rather than conclusive.

Confidence judgments

Chapter 5 examined the relative confidence with which visual and semantic errors are produced. Two analogues for confidence were developed in the DBM and GRAIN networks: the speed of settling, measured in terms of the number of iterations, and the "goodness" of the resulting representation, measured in terms of the energy in different parts of the network. Using both measures, visual errors were produced with more confidence than semantic errors, as has been observed in three deep dyslexic patients by Patterson (1978) and Kapur & Perl (1978), although the differences were small.[3]

Lexical decision

Coltheart (1980a) in his review rates lexical decision as being "surprisingly good" in nine patients, but most of the evidence is based on personal communication. The published results that are cited pertain only to two of the more recently described patients (D.E., P.W.; Patterson, 1979). Lexical decision was not rated "surprisingly good" in three patients; J.R. (Saffran, personal communication), P.S. (Shallice & Coughlan, 1980), and A.R. (Warrington & Shallice, 1979).[4] Moreover, our

[2] One could always introduce a "direct mapping" box and a "semantic clean-up" box into a conventional model, and explain the double dissociation in terms of separate impairments to these two boxes. However, to do so would violate the principle that modules are supposed to be individuated on the basis of the type of information they process (which in this case is the same—deriving the semantics of both concrete and abstract words from orthography). Such a proposal would also be pointless as it would contribute nothing to our understanding of the functions of these "modules."

[3] A somewhat different pattern of findings on G.R. (Newcombe & Marshall, 1980a) is not based on an adequate amount of data.

[4] A.R. differs from prototypical deep dyslexia patients in a number of ways (see Coltheart, 1980a). Also, his lexical decision was assessed in an unusual fashion.

attempts to demonstrate preserved lexical decision performance in a lesioned network have also been somewhat indeterminate. In an early investigation, Hinton & Shallice (1989) defined a "yes" response in lexical decision in the network by using a lower value of the proximity criterion than required for explicit naming (0.7, down from 0.8) and no gap criterion. This procedure did not result in relatively preserved lexical decision for words that could not be read. However, this effect was obtained in the present investigation (see Section 5.4) when a procedure similar to that employed by Seidenberg & McClelland (1989) was used with the DBM network. According to this procedure, letter strings are given a "yes" response in lexical decision when they can be "re-created" on the basis of orthographic and semantic knowledge. For words that could not be read, this yielded a d' value (1.94) of the same sort of range as that found in D.E. (1.74; Patterson, 1979). While these more recent results are promising, it should be kept in mind that aspects of the simulations—in particular, the definition of the task of lexical decision—are too unconstrained for the simulations to constitute a completely adequate characterization of preserved lexical decision in deep dyslexic patients.

Visual-then-semantic errors

A phenomenon that was not specifically investigated is the occurrence of visual-then-semantic errors in deep dyslexia (e.g. SYMPATHY ⇒ "orchestra", presumably mediated by *symphony*; Marshall & Newcombe, 1966) These are generally thought of as a visual error followed by a semantic error (Coltheart, 1980a), which presumably implies that two different impairments are involved. The present simulations provide a more parsimonious explanation, as the errors can arise when only a single set of connections is lesioned. They were observed unexpectedly using both the original H&S word set (Section 4.8) and the abstract/concrete word set (Section 6.5). The mechanism by which they arise is most clearly seen in the case where the network includes an output system. A lesion to the input system can produce a semantic representation very close to that of a word visually related to the stimulus. However, the attractors in the output system may map this slightly inaccurate semantic activity onto the phonology of a semantic neighbor of this visually related word rather than the phonology of the word itself. It is the *normal* operation of the output system that produces the semantic part of the visual-then-semantic error.

Subvarieties of deep dyslexia

The final empirical issue addressed by the present investigation of deep dyslexia is that it can arise in a number of forms. In some patients, such as V.S. (Saffran & Marin, 1977) and G.R. (Patterson, personal communication), comprehension performance is very similar for auditory word presentation as for visual. If a unitary impairment is assumed, then it must lie at or beyond the level of the semantic system. On the other hand, patients like P.S. (Shallice & Coughlan, 1980) and K.F. (Shallice & Warrington, 1980) were much better at comprehending spoken than written words, suggesting an earlier locus of impairment, between orthography and semantics. This contrast has led to the assumption that deep dyslexia can exist in two or more forms, with the impairment primarily involving input pathways in one case, and output pathways in the other (Friedman & Perlman, 1982; Shallice & Warrington, 1980). However, it remained totally unexplained why the two loci of impairment should give rise to a qualitatively similar pattern of errors.

The current simulations provide a simple explanation. When an output system was added to the model, and a lesion was made to either the first or second set of connections within this system, the resulting error pattern was qualitatively similar to the one obtained after input lesions (see Section 3.4). Indeed, qualitatively equivalent error patterns arise in the simulations from lesions to any stage along the semantic route, from the first set of connections after the graphemic units to the last set before the phonemic units.

9.2.2 Remaining empirical issues

No evidence was obtained relating to certain aspects of the deep dyslexia symptom-complex. Some of these—derivational errors, and part-of-speech effects—can be accounted for by natural extrapolations from the current results. The situation is less clear for others: associative semantic errors, patients who make no visual errors, and the relation with impairments in writing (deep agraphia). We consider each of these in turn.

Derivational errors

Deep dyslexic patients often make derivational errors, giving a response that is a different inflectional form of the stimulus (e.g. HITTING ⇒ "hit"). Since the word sets and orthographic representations we have used do not involve inflections, we could not have directly reproduced this type of error in our simulations. However, derivational errors can be considered to be one of a variety of mixed visual-and-semantic error, as they almost always have both a visual and a semantic relation to the stimulus. Therefore, above-chance rates of such errors are to be expected given the rates of mixed errors produced in the simulations. This is not to deny that the representations of inflectional forms of a word are related in a special way, unlike other visually or semantically related sets of words (Patterson, 1978; 1980)—only to point out that the occurrence of derivational errors in deep dyslexia can be explained without such an assumption.

Part-of-speech effects

In general, deep dyslexics read nouns better than adjectives, adjectives better than verbs, and verbs better than function words. Both the H&S word set and the abstract/concrete word set contain only nouns. However, Jones (1985) showed that ordering words in term of ease-of-predication results in the same overall rank ordering of syntactic classes. In addition, Barry & Richardson (1988) found that part-of-speech had no effect on the reading performance of G.R. when concreteness, frequency, and "associative difficulty" (closely related to ease-of-predication) were statistically controlled. In the abstract/concrete simulations, we reflected the ease-of-predication of a word in terms of the number of active features in its semantic representation, and found that concrete words, with greater ease-of-predication, are read better than abstract words. It would seem appropriate to give different parts-of-speech semantic representations in which the average number of features varied in a similar fashion. By analogy with the effects found with the abstract/concrete word set, one would expect that damage to the main part of the network would result in the same rank order of correct performance, with nouns > adjectives > verbs > function words. Thus the approach taken in the simulations seems likely to produce the part-of-speech effects found in deep dyslexia (also see Marin et al., 1976).

Associative semantic errors

Coltheart (1980c) argued that two types of semantic errors occur in deep dyslexia: a *shared-feature* type, and an *associative* type. In the present simulations, only the shared-feature type was formally investigated. Comparing Tables 6.1 and 6.2 of Coltheart (pp. 147-148, also see the error corpora in Appendix 2 of Coltheart et al., 1980), this type appears to be the larger group, and over half of those held to be associative by Coltheart appear to have visual (V) or shared-feature (SF) characteristics as well.[5] In some errors, however, the associative aspect completely dominates (e.g. FREE ⇒ "enterprise", STAGE ⇒ "coach"). Could a network produce such errors?

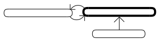
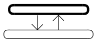
Notice that words with an associative relationship often follow one another in spoken and written language. In the course of normal fluent reading, the system must quickly move from the representation of one word to the next. Suppose that the system must start from the attractor of the current word, or at least is biased towards it, when beginning to process the next word. For word pairs that frequently follow each other (e.g. WRIST WATCH), the network will learn to lower the energy boundary between the attractor basins for the two words so that the transition can be accomplished more easily (Elman, 1990).[6] This lower boundary would be more easily corrupted or lost under damage than the boundaries between basins for other word pairs. As a result, presentation of the first word would become more likely to settle into the attractor for the second word, resulting in an associative semantic error. This explanation also predicts that the reverse ordering should also become more likely as an error, which is found in patients (e.g. DIAL

[5] WRIST ⇒ "watch" (SF), ANTIQUE ⇒ "vase" (SF), NEXT ⇒ "exit" (V), PALE ⇒ "ale" (V), COMFORT ⇒ "blanket" (SF), IDEAL ⇒ "milk" (SF), THERMOS ⇒ "flask" (SF), INCOME ⇒ "tax" (SF), MOTOR ⇒ "car" (SF), BRING ⇒ "towards" (SF), POSTAGE ⇒ "stamps" (SF), WEAR ⇒ "clothes" (SF), STY ⇒ "pig" (SF), BLOWING ⇒ "wind" (SF), SHINING ⇒ "sun" (SF), CONE ⇒ "ice-cream" (SF).

[6] This explanation does not imply that sequences of interpretations are *caused* by temporarily adjusting the energy boundaries between them, but only that an *effect* of learning sequences would be to lower the boundaries between frequent transitions.

⇒ "sun" and CONE ⇒ "ice-cream"; Coltheart, 1980c).[7] Of course, these errors would become even more likely if the two words shared any visual or semantic features.

Patients who make no visual errors

A major contribution of the current connectionist approach to deep dyslexia is the ubiquitous co-occurrence of visual, semantic, mixed visual-and-semantic errors when an attractor network that maps orthography to semantics is lesioned. Thus, possibly the strongest empirical challenge to the current account is the existence of three patients who make semantic and derivational errors in reading, but no purely visual errors (K.E., Hillis et al., 1990; R.G.B. and H.W., Caramazza & Hillis, 1990). K.E. made semantic errors in all other lexical processing tasks as well (e.g. writing to dictation, spoken and written picture-word matching), suggesting damage within the semantic system. In contrast, R.G.B. and H.W. made semantic errors only in tasks requiring a spoken response, suggesting damage in the output system after semantics. While a number of the network architectures we examined in Chapter 4 produced no visual errors with some types of clean-up damage when the response criteria were used (e.g.  C⇒S lesions;  S⇒I lesions), all of the networks produced visual/phonological errors for every lesion location when an output system was used. The primary motivation for developing an output system was to obtain an unbiased procedure for generating explicit responses from semantic activity, rather than to model the human speech production system *per se*. In fact, there are many ways in which it is clearly inadequate for the latter purpose (cf. Dell, 1986; 1988; Levelt, 1989). However, we have considered the pattern of errors produced by lesioning the output network as helping to explain the existence of an output form of deep dyslexia. Therefore, we can hardly argue that the deficits of R.G.B. and H.W., much less K.E., are outside the scope of the model.

As far as patient K.E. is concerned, the initial report on word reading refers to most errors being semantic, but remaining errors include phonologically and/or visually related ones. These only amounted to 1.4% of all non-correct responses in the main experiments reported. However, these experiments involved the presentation of a considerable number of items (e.g. 14) from each of a number of categories (4 or 10), with each item presented in a number of different tasks (e.g. 5). Thus, items in a small set of categories were repeatedly presented. It seems likely that K.E. would learn the categories and use this to limit the number of visual responses, as these would tend not to fall in one of the categories. In any case, the experimental context was clearly different from the standard one where the deep dyslexic reading pattern is reported. In addition, a considerable number of mixed errors seem to occur, but this is not analyzed in the paper.

There appear to be two very different ways in which the absence of visual/phonological errors in

[7] The explanation does not imply that both directions of an associative error need be *equally* likely after damage, because there can be differences in the paths that the network follows in state space, settling from the initial pattern for one word to the final pattern for the other.

R.G.B. and H.W. can be explained. The first is on the basis of a difference in the status of semantic errors as "speech acts" (Searle, 1969) in these patients as compared with other deep dyslexia patients. Deep dyslexic patients at times produce a circumlocutory response—they describe the meaning of the word rather than attempting to read it aloud. However, in general, such responses form only a small part of the deep dyslexic's output (e.g. G.R., K.F.). In contrast, both R.G.B. and H.W. produce many responses which are described as "definitions" of the words they are trying to read (21% and 28% of all non-correct responses, respectively). Caramazza & Hillis (1990) report that, in repetition tasks, R.G.B. produced many circumlocutions, while H.W. often followed her errors with the comment, "I can't say what you said but that is the idea." As the patients were clearly frequently trying to communicate that they understood the word, it seems quite plausible that any potential visual/phonological error (that would not be sense-preserving) would be edited out prior to articulation. After all, it is convincingly demonstrated that semantic access from the written word was unimpaired in both patients. Semantic errors, on the other hand, would be more difficult to detect as errors at the semantic level and could, in fact, serve as an approximation to the meaning for communication purposes.

Alternatively, the lack of visual/phonological errors in a few patients may be explained by individual differences in the effects of qualitatively equivalent lesions in connectionist networks. The reported simulation results are the sum of a number (typically 20) of random samples of a given lesion type. In a network, qualitatively and quantitatively equivalent lesions, such as instances of O⇒I(0.3), have quantitatively different effects depending on the particular connections removed (also see Patterson et al., 1990). The reported results are means of distributions—the patients who make no visual/phonological errors may correspond to the tail of one of the distributions.

Neither of these solutions to the problem posed to our modeling work by the patients of Hillis et al. (1990) and Caramazza & Hillis (1990) is completely satisfactory. In our account of deep dyslexia, we have accepted that the response produced by the patient can be modeled directly by the output of our network(s), and that the means of the effects of 20 qualitatively and quantitatively equivalent lesions can model the responses produced by a patient with only one lesion. Our two possible responses to the patients who make no visual errors imply that at least one of these assumptions can at best hold only for the large majority of patients. The theory cannot apply in its strongest form to the results produced by *all* patients who read by the semantic route as a result of neurological damage.

Acquired dysgraphia

The final characteristic of deep dyslexia that Coltheart, Patterson and Marshall (1987a) describe is that "if a patient makes semantic errors in reading isolated words aloud he or she will also....have impaired writing and spelling" (p. 415) which, they argue, will involve either a global or a deep dysgraphia. However, the converse relation does not hold; there are deep dysgraphic patients who are not deep dyslexic (e.g. Bub & Kertesz, 1982; Newcombe & Marshall, 1984; Howard & Franklin, 1988). These results challenge the simple presumption that the orthographic processing systems involved in writing are the same as those involved in reading.

According to the present account, deep dyslexia depends on the co-occurrence of at least two major types of damage: the first to the phonological route, and the second (less severe) to the semantic route. One possible explanation of deep or global dysgraphia without deep dyslexia is that, in most people, writing is a less well-learned skill than reading, and so would be more vulnerable to the effects of brain damage. Given this, and the fact that both reading and writing make use of common semantic and phonological systems, damage that is sufficient to produce deep dyslexia would seem likely to impair writing and spelling as well. On this account, though, deep dyslexia without deep or global dysgraphia should eventually be observed. Indeed, relatively recovered pure alexic patients (Coslett & Saffran, 1989a) would seem to fit this pattern (also see the patients of Beringer & Stein, 1930, and Faust, 1955, discussed by Marshall & Newcombe, 1980).

Visual vs. phonological errors

It has frequently been suggested that some deep dyslexic patients have an impairment in accessing phonological lexical representations from semantics (e.g. Friedman & Perlman, 1982; Patterson, 1978; Shallice & Warrington, 1980). There are three main lines of evidence that lead to this conclusion. First, certain patients (e.g. P.W. and D.E.; Patterson, 1978) frequently select the presented word when offered a choice between it and their semantic error, implying that they know the presented word. Second, in auditory-visual matching these patients again usually select the presented word rather than their visual error. Third, certain patients perform as well on visual word-picture matching as for auditory word-picture matching, and perform both at close to normal levels (e.g. V.S., Saffran & Marin, 1977; P.W., Patterson, 1979), although others are much worse with visual than with auditory presentation of words (e.g. P.S., Shallice & Coughlan, 1980; K.F., Shallice & Warrington, 1980).

Our simulations present a potential problem for this argument. The output network develops strong phonological attractors in the same way that the input network develops strong semantic attractors. Thus, for the same reason that damage to the input network produces visual and semantic errors, damage to the output network would be expected to produce semantic and *phonological* rather than visual errors. This prediction stands in contrast with the inclusion of visual errors *per se* as a symptom of deep dyslexia.

The word sets used in the current simulations were not designed to differentiate phonological from visual errors. Yet pure phonological errors (e.g. HAWK ⇒ "tor" with British pronunciations) certainly occur when the output pathways are lesioned. Whether phonological errors occur in deep dyslexia has never to our knowledge been empirically investigated, although Goldblum (1985) suggests that the so-called visual errors are actually phonological. However, inspection of the error corpora for a number of patients (Coltheart et al., 1980, Appendix 2) do not support this interpretation. If one takes P.W., for example, many errors are more easily explained as a visual error (e.g. ORATE ⇒ "over", CAMPAIGN ⇒ "camping") but only one is easier to explain as a phonological error (GRIEF ⇒ "greed"). Attempts to simulate the three empirical phenomena that suggest an output lesion might reveal that they are compatible with an input lesion, or more particularly a lesion to the semantic system itself. In any case, the area requires further empirical study and simulations.

9.3 Theoretical issues

The connectionist account of deep dyslexia that we have developed from the position advocated by Hinton & Shallice (1991) is based upon four assumptions, listed in Section 9.1.2 above, about the process of mapping orthography to semantics. The first two of these are standard assumptions within connectionist modeling. Another, on the difference between representations of abstract and concrete words, is derived from earlier theorizing. Only the third, concerning attractors, is at all original to the present approach. In addition to these four assumptions, two more are necessary to account for additional characteristics of deep dyslexia. The first—that the mapping from orthography to semantics is isolated from phonological influences—is standard in accounts of deep dyslexia (see Coltheart et al., 1980). The second—that the pathway from orthography to semantics is also affected by a lesion—is widely but not universally held (see Shallice, 1988, for discussion).

If one takes the nine characteristics held to apply to deep dyslexia by Coltheart, Patterson and Marshall (1987a), three are directly explained in a principled fashion on the present account (semantic errors, visual errors, concrete word superiority). Three more (derivational/morphological errors, the part-of-speech effects, and function word substitutions) follow in a straightforward fashion from the simulations, even though they have yet to be implemented. An additional two are an immediate consequence of one standard assumption, that of the absence of phonological processing. Only one—the relation between reading and writing—is at all problematic. In addition, the simulations offer principled accounts of five other phenomena which have been widely investigated empirically: mixed errors, the interaction of semantic factors in the genesis of visual errors, confidence in error types, lexical decision, and most surprisingly of all, the visual-then-semantic errors. However, as discussed in the preceding section, there are a number of other less central aspects of the disorder which are not yet well accommodated within the approach.

This 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 what we have called a "principled account." By this, we mean that (a) many aspects of the syndrome are explained from a common set of basic assumptions, rather than requiring specific extra assumptions for each aspect; and (b) the explanations are derived from the assumptions computationally rather than intuitively.

Consider, as an example, the shared-feature semantic error itself. Various theoretical accounts have been given as to why such errors should occur. Coltheart (1980c), in his review of the phenomenon, considers two theories, but rejects one—the imagery explanation—as being empirically much inferior to the other. The second one—the Marshall & Newcombe (1966) account—takes a position derived from Katz & Fodor (1963) in arguing that the patient lacks the ability to descend a hierarchically organized semantic tree to the appropriate terminal leaf when deriving a phonological form from a semantic representation. Yet, as Coltheart points out, this account would not explain the standard non-synonymous co-ordinate errors (e.g. NIECE ⇒ "aunt"). He suggests "one needs to suppose that when a determiner is lost, sometimes it leaves some trace: the patient knows that a determiner is lost, so supplies one, without having any way of selecting the correct determiner" (p. 153). While Coltheart provides some limited empirical arguments in favor of this amended Marshall & Newcombe position, his amendment is not derived from any deeper assumptions and is not used in the explanation of any other phenomenon. It remains, therefore, theoretically *ad hoc*. The account given by Shallice & Warrington (1980) suffers from similar problems to that of Marshall & Newcombe (1966), and that of Morton & Patterson (1980) introduces specific *ad hoc* assumptions. By contrast, on the present account the existence of semantic errors essentially derives from the assumption of attractors, which is also used in explaining many other aspects of the syndrome.

9.3.1 The right hemisphere theory

Two main classes of theory have been put forward to account for deep dyslexia: the multiple functional impairments position (e.g. Morton & Patterson, 1980; Shallice & Warrington, 1980) and the right hemisphere theory (e.g. Coltheart, 1980b; 1983; Saffran et al., 1980; Zaidel & Peters, 1981). The current account account adopts the "subtraction" assumptions taken by the multiple functional impairment theories, whereby impaired behavior is explained by the damaged operation of the same mechanism that subserves normal behavior. In a sense our account is a specific version of this class of theory. However, as discussed in Section 2.4, multiple functional impairment theories have problems in limiting the number of postulated impairments, and the locus of damage that explains one symptom often differs from that assumed for another. The present version has two advantages in addition to the principled nature of its predictions: it can explain a wide range of symptoms assuming that the isolated semantic route is subject to only one locus of lesion, and can also explain why a number of different loci of lesions give rise to qualitatively similar patterns of symptoms.

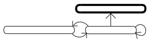
The right hemisphere theory differs from the multiple functional impairment theories in that many aspects of the syndrome are derived from a common cause. Here, though, the extrapolation from the basic assumption is an empirical one—the reading behavior of deep dyslexic shares aspects with that of other patients known to be reading with the right hemisphere (and normal subjects under brief lateralized presentation). The adequacy of these correspondences is a matter of ongoing debate (see Barry & Richardson, 1988; Baynes, 1990; Coltheart et al., 1987a; Jones & Martin, 1985; Marshall & Patterson, 1983; 1985; Patterson & Besner, 1984a; 1984b; Patterson et al., 1989; Rabinowicz & Moscovitch, 1984; Shallice, 1988; Zaidel & Schweiger, 1984). The important point is that the present connectionist account is orthogonal to one based on right hemisphere reading. If the right hemisphere reads by the same principles as the normal mechanism for reading via meaning (although perhaps less effectively), then the connectionist account would still apply. In addition, one would not have to postulate that the right hemisphere reading process has a particular set of properties—they could be inferred from the connectionist account. Moreover, the connectionist account could also explain reading patterns similar to deep dyslexia which *are* based on left-hemisphere reading (and so can be abolished by a second, left hemisphere stroke; Roeltgen, 1987). In such an account, the total reading system would contain both left hemisphere and right hemisphere units and connections (as well as inter-hemispheric corrections) with the left hemisphere ones being more numerous. However, the compatibility of the connectionist and right hemisphere accounts of deep dyslexia depends on the assumption that right hemisphere reading differs from normal reading only quantitatively and not qualitatively. In their review which is broadly favorable to the right hemisphere theory, Coltheart, Patterson and Marshall (1987a) leave this issue open.

9.3.2 Attractors vs. logogens

At a more detailed level, the operation of attractors plays a central role in our account of deep dyslexia. How do attractors relate to other theoretical concepts that have been used in explaining deep dyslexic reading behavior? The most commonly used concept with some relation to an attractor is that of a "logogen" (Morton, 1969; Morton & Patterson, 1980; also see Section 2.1.2). We take the defining characteristic of a logogen to be that it is a representation of a word, with an associated activity level, in which all of the information (of a particular type) relating to the word is packaged together. Words are related to other words via information that is *external* to the logogens themselves. In this way, logogens operate much like "localist" representations in connectionist networks (Feldman & Ballard, 1982; McClelland & Rumelhart, 1981), and the relationship between attractors and logogens is much the same as that between distributed and localist connectionist representations.

A full consideration of the impact of the distinction between localist and distributed representation of concepts is far beyond the scope of this thesis. We only raise two issues. The first relates to the degree to which concepts (words) can operate *independently*. In a localist representation, words can influence other parts of the system in a manner unrelated to the way other words have influence (e.g. in generating a pronunciation from semantics). This is a strong advantage because the meanings of words are arbitrarily related to their spelling and pronunciation. For this reason, reading for meaning is the paradigmatic domain in which localist representations appear most appropriate (Hinton et al., 1986). However, a localist representation of words prior to semantics would eliminate all effects of *visual* similarity in later processing, which seems inconsistent with the ubiquitous occurrence of these errors in all subvarieties of deep dyslexia. In contrast, words in a distributed representation can have effects *only by virtue of their features*, and so other words tend to have similar effects to the degree that they share those features. The use of attractors is a way of compensating for this bias of distributed representations in domains where it is problematic, but the underlying effects of similarity are revealed under damage.

The second issue has to do with the status of concepts that do not explicitly correspond to words.[8] Localist representations have difficulty representing concepts for which there is no existing unit. One approach is to use a distributed representation over units for related concepts—this is how McClelland & Rumelhart's (1981) Interactive Activation model produces a pseudo-word superiority effect in letter identification tasks. But ultimately a new unit must be allocated for the new concept, and connected with appropriate weights to other units. How this might be accomplished in a biologically and psychologically reasonable way is still an open issue (but see Feldman, 1982). Distributed representations can represent new concepts quite naturally as novel combinations of features. The pattern of features can be made into an attractor by appropriate modification of connection weights within semantics. However, learning the associations of the new concept with its spelling and pronunciation is more difficult. The weights from individual semantic features to phonological features must be altered so that the novel semantic pattern activates an initial phonological pattern within the basin of attraction of the new pronunciation, without disrupting the mappings for other words. This is also an open research issue.

The attractor network which would appear to be closest to the updated logogen model of Morton & Patterson (1980), as far as the process of reading via meaning is concerned, is the  one, in which attractors are built at the level of the units intermediate between letter representations and semantic ones. However, a major difference between the logogen approach and this attractor one should be noted. The similarity metric of the relation between logogens is purely visual/orthographic. If the activation level of a second logogen is near to that of one that reaches threshold then this implies only that the two represent stimuli that are visually similar. In contrast, the similarity metric for attractors is both visual and semantic. Thus damage to attractors can produce both visual and semantic influence in errors, while damage to logogens can result only in visual confusions.

[8] Geoff Hinton has called these "dark concepts" by analogy with regions of space which cannot be seen because any matter they contain does not emit light (rather appropriately known as "dark attractors").

9.4 Extensions of the approach

9.4.1 Other "deep" syndromes

The connectionist account we have provided for deep dyslexia would seem to be directly generalizable in two ways: first, to other syndromes in which an input/output mapping can be accomplished only via semantics; and second, more generally to syndromes in which a single-route mechanism maps between arbitrarily-related input and output domains. In the first case, 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).

Deep dysphasia involves the co-occurrence of semantic and phonological errors in repetition, and a concrete word superiority (see e.g. Morton, 1980; Michel & Andreewsky, 1983; Howard & Franklin, 1988; Katz & Goodglass, 1990; Martin & Saffran, personal communication). In some patients (e.g. N.C. of Martin & Saffran), the parallel with deep dyslexia is very close, as the phonological errors are phonologically related words. In other patients (e.g. R. of Michel & Andreewsky), responses which are phonologically related to the target are often paraphasic (i.e. non-words). In general, though, this syndrome would fit with an explanation in which repetition must rely on partially impaired semantic mediation, because damage has eliminated the standard, direct route from input phonology to output phonology (see Morton, 1980; Howard & Franklin, 1988; Katz & Goodglass, 1990).

If semantic mediation in writing operates by principles analogous to those for reading, then the corresponding pattern of symptoms would be expected to result from lesions. In fact, essentially the same arguments that apply for deep dyslexia also apply for deep dysgraphia (see e.g. Bub & Kertesz, 1982; Newcombe & Marshall, 1984; Howard & Franklin, 1988). Specifically, phonological mediation in writing is inoperative, and semantic mediation suffers from damage complimentary to that in the reading processes simulated in current work.

More generally, any domain that involves mapping between arbitrarily-related domains, analogous to orthography and semantics, would be expected to give rise to error patterns that are analogous to those found in deep dyslexia (except for aspects that are specific to orthography or semantics, such as the effects of abstractness). To what extent such relations hold is beyond the scope of this thesis.

Beyond accounting for the specific pattern of reading behavior of deep dyslexic and closely related patients, and analyzing the underlying computational principles, a major goal of the thesis was to extend the approach to other domains of interest in cognitive neuropsychology and connectionist modeling. Specifically, Chapter 7 investigated the effects of relearning after damage and their implications for cognitive remediation in patients, and Chapter 8 presented a simulation of the pattern of visual object naming errors in optic aphasia. While both of these studies were more exploratory than the simulations of deep dyslexia, it seems warranted to consider how well they address the relevant empirical issues.

9.4.2 Cognitive remediation

The goal of work in cognitive remediation is to maximize the general recovery of the cognitive functions that are impaired in a neurological patient. What makes the approach *cognitive* is that the therapy is directed at reestablishing the operation of particular computational subsystems based on a functional analysis of the cognitive processes involved in the task domain, and their specific impairment in the patient. Connectionist modeling offers specific hypotheses about the nature of the representations and computations that underly cognitive processes, and how these can be reestablished through relearning after damage. Of course, the ultimate test of the empirical adequacy of relearning simulations is the extent to which the hypotheses they generate lead to improved therapy for patients. While such a demonstration is beyond the scope of this thesis, we can tentatively evaluate the promise of a connectionist approach to cognitive remediation based on the general correspondence of the effects of relearning and generalization in patients and networks.

Many patients with an impairment in mapping between orthography and semantics show benefits from treatment of specific words, as well as generalized improvement on untreated but related words (e.g. Behrmann, 1987; Coltheart & Byng, 1989; Scott & Byng, 1988). Why should this occur? In general, there is little understanding of the underlying mechanism by which cognitive functions recover, either spontaneously or as a direct result of therapy. In part this is due to the lack of specific proposals about the neural implementation of cognitive processes, and how this implementation is learned initially and relearned after brain damage.

From a purely methodological point of view, connectionist networks appear particularly appropriate for investigating issues in remediation because they are computationally powerful enough to carry out complex tasks, allow a natural analogue of neurological damage in terms of removing units and/or connections, and can improve their performance with local learning rules that apply equally well in the damaged and undamaged state. More importantly, their *behavior* in relearning after damage is qualitatively similar to that shown by patients. The networks relearn a task quickly after damage, and exhibit spontaneous recovery of related knowledge that is not explicitly retrained. These effects occur because the knowledge that accomplishes all of the associations in the task is distributed across all of the weights in the network—to the degree that the task is structured, relearning some of the associations produces weight changes that improve performance on all of them. In this way, the principles that underly the effects of relearning in connectionist networks provides a specific hypothesis about the basis of recovery and generalization in patients.

While the simulations presented in Chapter 7 in the domain of reading via meaning are too limited to warrant a detailed comparison with particular patients, they have implications for the *degree* of relearning and generalization exhibited by patients. Thus the simulations predict greater generalization in patients who have damage to a part of the system that carries out a more structured mapping. For example, patients with damage within semantics should show greater generalization than patients with damage closer to orthography (see Behrmann & Lieberthal, 1989). Conversely, the degree to which a patient shows generalization during remediation can generate a hypothesis about the detailed location of their functional impairment (i.e. whether it more involves semantics or orthography).

A particularly interesting, but unfortunately inconclusive, aspect of the current approach is the possibility that connectionist modeling might provide a framework within which to generate and evaluate hypotheses about how to design the most effective therapy. In a preliminary step in this direction, the current simulations found that retraining on words whose semantics are atypical of their category yields more generalization to more prototypical words than *vice versa*. However, an additional experiment demonstrated that the effect was due to the nature of the unretrained (test) set rather than the retrained set, and in general, the word set is too limited to support definitive implications for patient therapy.

9.4.3 Optic aphasia

Simulations presented in Chapter 8 attempted to extend the principles that explain deep dyslexic reading behavior to account for the pattern of errors made by optic aphasics in naming visually presented objects. Three aspects of the behavior of these patients were problematic: (a) responses are often influenced by previously presented objects; (b) errors are primarily semantically rather than visually related to the stimulus; and (c) performance in gesturing and semantic categorization tasks can be much better than naming performance.

In order to account for the influence of previous objects, we introduced short-term correlation weights that bias the network towards recent interpretations. As a result, like optic aphasics, the damaged network often misnames objects as the preceding object or as one semantically related to it. Although there is independent computational and empirical motivation for short-term weights, they were included in the current simulations specifically to reproduce the perseverative effects of optic aphasics, and so must be viewed as rather *ad hoc*. Thus, the mere occurrence of perseverations is less interesting than the interactions of these perseverative effects with semantic influences, which are not inherent in the operation of the short-term weights. Nonetheless, it may be safest to interpret the use of short-term weights in object recognition as a hypothesis that may warrant further empirical investigation.

Both the network and optic aphasics produce predominantly semantic rather than visual errors in visual object naming. This effect is not simply due to the difference “chance” rates of semantic vs. visual similarity because it is found *relative* to these chance rates. Rather, it is argued that the structure in the mapping from visual to semantic representations for objects reduces the influence of visual similarity in shaping semantic attractors. Further research is required to test whether differences in task structure can account for analogous differences in error patterns in other domains.

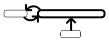
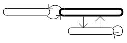
The current simulations did not directly address what is perhaps the most problematic aspect of optic aphasia—apparently intact recognition with impaired naming. For this reason, the simulations do not reproduce optic aphasia *per se* but only the corresponding error pattern in visual object naming. A two-sided argument was presented to reconcile the current work with the preserved categorization and gesturing abilities of optic aphasics. The first side contests the claim that recognition is intact in these patients. The second side suggests how the remaining capabilities might be subserved by the residual operation of a semantic system with some degree of specialization by modality (Allport, 1985; Farah & McClelland, 1991; Warrington & McCarthy, 1987). While the argument suggests the possibility of a complete simulation of optic aphasia within the current framework, a working simulation is required before it should be taken as anything more than speculation.

9.5 The impact of connectionist modeling in neuropsychology

Deep dyslexia was first described in a single patient, G.R. (Marshall & Newcombe, 1966), but it soon began to be conceived as a "symptom-complex" (Marshall & Newcombe, 1973), and then as a "syndrome"—that is, as a collection of behaviors arising from a specific functional impairment (Coltheart, 1980a; Marshall & Newcombe, 1980). Almost immediately this position was criticized. Morton & Patterson (1980) rejected the concept of a syndrome. Shallice & Warrington (1980) argued that the pattern of symptoms could have a number of different origins (also see Coltheart & Funnell, 1987). Caramazza (1984) and Schwartz (1984) argued against the general methodology of assuming that frequently observed combinations of symptoms represented the effects of a single underlying impairment. Shallice (1988), while willing to accept syndromes based on dissociations, rejected errors in particular as a fruitful basis on which to generalize across patients. Even Coltheart, Patterson and Marshall (1987a), in their later review, seem rather pessimistic about characterizing deep dyslexia as a syndrome, unless the right hemisphere theory were correct.

The present investigation has both positive and negative theoretical implications for the validity of the concept of a "syndrome," in deep dyslexia and more generally. On the positive side, the work was motivated by the possibility that deep dyslexia is indeed a coherent functional entity. However, there is a critical difference in the nature of the functional entity as envisaged in the current research, and the formulation that has been accepted, either implicitly or explicitly, both by critics (e.g. Caramazza, 1984; 1986) and by defenders (e.g. Coltheart, 1980a; Shallice, 1988) of the syndrome concept. According to this standard formulation, if a symptom-complex is to be of theoretical interest, it must arise from the same functional lesion site for all patients who exhibit it. If it can be demonstrated that some aspects of the symptom-complex do not always co-occur across patients, then this is considered evidence that the symptom-complex can arise from more than one locus of damage. The symptom-complex becomes a "psychologically weak syndrome" and hence of little or no theoretical interest (see Caramazza, 1984; Coltheart, 1980a, for relevant discussion).

While this logic seems appropriate for theoretical analyses in terms of conventional "box-and-arrow" systems, the present research shows that it is not appropriate for at least some connectionist systems. Part of the overall symptom pattern may occur as a result of lesions in many parts of a complex system, for reasons that derive directly from the nature of the computation that the whole system is carrying out. An example is given in the present simulations by the qualitative similarity of error patterns whenever lesions are made between orthographic input and semantic output. At the same time, other aspects of the symptom-complex may differ between lesion sites. Thus lesions to the clean-up network do not show the concrete word superiority effects shown by lesions to the direct pathway, even though they produce the same patterns of visual and semantic similarity in errors (see Section 6.5). This means that, even when patients differ in some respects, the aspects of their behavior that are similar may still arise from a common functional origin. Thus considering these patients together may be a valuable guide to understanding the impaired system. In this way, even the existence of so-called "weak syndromes" can be theoretically productive.

There is also a negative side to the general methodological implications of the current simulations. Hinton & Shallice (1991) showed that a "strong dissociation" (Shallice, 1988) between the processing of different semantic categories can occur when particular lesions are made to the clean-up pathway. The category "foods" was selectively preserved in a striking manner. However, when lesions were made to a second network which was essentially the same except for the use of a different random starting point for the learning procedure, the dissociation did not occur. The present simulations show similarly dramatic effects when the same set of connections are lesioned, but again, minor changes in architecture lead to different category effects: "animals" were performed over 20 times better than "body parts" for the  network, and over three times better than "outdoor objects" in the  network (see Figure 4.18, p. 105). It would appear that the strong dissociations obtained may reflect idiosyncrasies in the learning experience of particular networks. If this is so, and if the cognitive system operates according to connectionist principles, then these results casts doubt on the enterprise, advocated by Shallice (1988) and others, of identifying the major divisions in the cognitive system based on the existence of strong dissociations.

Fifteen years ago, Marin, Saffran and Schwartz (1976) responded to criticisms of the relevance of neuropsychological findings for understanding normal cognition by pointing to high-energy physics, where studying the effects of random damage has produced substantial theoretical results. The results obtained in this thesis, together with analyses of equivalent depth that are beginning to be made of other syndromes as well, suggest that the analogy may be closer than Marin and colleagues intended. If our simulations are valid, in principle even if not in detail, then neuropsychological evidence, such as the deep dyslexia syndrome, will provide strong support for a particular organization of the cognitive system which would probably prove difficult to obtain by the use of experiments on normal subjects. On the other hand, without detailed simulations, appropriate interpretations of many aspects of the syndrome would be virtually impossible. In this case, cognitive neuropsychology will benefit most extensively from an interplay between empirical and computational approaches in future work.

9.6 The impact of neuropsychology on connectionist modeling

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 results point to the importance of the interaction of task structure (both within and between domains) with other aspects of network design. The results involve (a) the natural occurrence of "blend" responses under damage, (b) the above-independent rates of mixed visual-and-semantic errors, (c) the degree of generalization in relearning, and (d) the relative differences in visual error rates across tasks.

Blends

Chapter 3 illustrated the tendency of attractor networks to develop spurious attractors which, under damage, result in responses that are inappropriate "blends" of familiar responses. These additional attractors are not problematic in most applications because the network is only tested on input that is similar to trained input. Damage provides a stronger challenge to the network because portions of the network "downstream" from the damage may receive input that is drastically different from any received during normal operation. In fact, the effects of internal damage cannot be well-approximated by simply corrupting the input to the network, because the normal operation of earlier portions of the network will mitigate the effects on later portions. This was shown in the current simulations by the fact that lesions further from semantics (e.g. O⇒I) tended to be less debilitating than lesions closer to semantics (e.g. I⇒S). In this way, investigating the effects of damage can reveal properties of the network that would be difficult to identify from its normal operation.

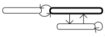
The fact that the spurious attractors tend to correspond to *blends* of familiar responses, rather than random responses, reflects the bias of the network to produce similar outputs to similar inputs. When the phonology of a word is trained as a response, other phonological patterns are also reinforced to the extent that they overlap with the trained pattern. As the simulations demonstrate, the occurrence of blends is significantly reduced by training the network in a way that encourages the formation of strong attractors for familiar patterns. These attractors overcome the similarity bias within phonology by enforcing *coherence* among each familiar pattern that may run counter to its similarity with other patterns. Thus, the frequency of blend responses under damage provides a measure of the degree to which the attractors can overcome the bias of similarity in the output domain. The fact that the DBM produced few blends without recourse to special training techniques (see Section 5.1.7) suggests that contrastive Hebbian learning may naturally produce stronger attractors than back-propagation.

Rates of mixed errors

The second issue concerning the effects of task structure involves the conditions under which damage produces rates of mixed visual-and-semantic errors above those expected from the independent rates of visual errors and semantic errors. As described above, among the deep dyslexic networks trained with noise, mixed rates were higher than expected only when the intermediate units between orthography and semantics were involved in implementing the attractors. In other words, if the network could separate the influences of visual and semantic similarity (by having a separate semantic clean-up pathway), it did so. This makes sense given that visual and semantic similarity are mutually uninformative. The need to overcome misleading similarity between some inputs and outputs induces a degree of specialization for different types of information in different portions of the network. In contrast, when visual and semantic similarity *are* related (in the object recognition/naming task), the rates of mixed errors are far above the predicted rates, suggesting less specialization.[9] Thus, the structure in the mapping from input to output domains strongly influences the way that the task is distributed throughout the network.

Generalization in relearning

Relearning experiments in Chapter 7 found significant generalization to untrained words after lesions to the semantic clean-up pathway (`C`⇒`S`), but little if any generalization after lesions near the orthographic input (`O`⇒`I`). Generalization occurs because the weight changes induced by the trained words approximate the changes that are appropriate for the untrained words as well. This is true to the extent that the structure in the mappings for trained words is shared by the mappings for untrained words. Thus, the relative difference in the degree of generalization in relearning after `C`⇒`S` vs. `O`⇒`I` lesions is interpreted as reflecting the amount of structure in the "subtasks" performed by these sets of connections. While there is a sense in which the subtask performed by the `O`⇒`I` connections is structured, since visually similar inputs tend to have similar intermediate representations, this similarity actually *hinders* the network in generating the appropriate semantic

[9] Unfortunately, the comparison of tasks is confounded by the fact that the optic aphasia network has feedback connections from semantics to the intermediate units, and so these units are involved in implementing the attractors. However, the ratios of mixed to other error rates in the optic aphasia network are much larger than in the equivalent deep dyslexia network (), indicating an influence of task structure on mixed rates in addition to the architectural influences.

representations. Hence, reestablishing the mappings for trained words after damage does not help generate the semantics for untrained words. In contrast, words with similar semantics require similar clean-up, and this similarity directly contributes to generating the correct semantics. In this case, the resulting weight changes from the trained words also help untrained words with similar semantics, producing substantial generalization. Furthermore, since the combined influence of non-prototypical words better approximates the changes required by prototypical words than *vice versa*, retraining on the former leads to more generalization.

Rates of visual errors

The final result of the current research relating to task structure involves its effect on the influence of input similarity in shaping attractors, as reflected in the rates of purely visual errors.

In mapping orthography to semantics, there are strong constraints on the shapes of the boundaries between attractor basins. The network must shape the basins so as to divide between visually similar words with very different meanings (see Figure 2.10), without becoming so large as to capture nearby semantically related words. Since this division can be quite far from the final semantics of each word, the basins must become rather elongated (or equivalently distorted) in order to compensate for the lack of structure in the task. This elongation gives rise higher rates of visual errors under lesions to the direct pathway, and to a lesser extend under semantic damage as well.

In contrast, there is less pressure to distort the attractor basins when solving a more structured task, such as mapping visual object representations onto semantics. Visually similar objects, that generate similar initial semantic activity, tend to remain in nearby regions of semantic space. Hence, most of the boundary of the attractor basin for each object adjoins basins for semantically related objects. As a result, semantic errors predominate after damage. While visual errors still occur at above-chance rates, the increased structure of the task reduces the influence of visual similarity in shaping attractor basins.

Taken together, the results of the thesis clearly demonstrate the importance of task structure in understanding the layout of attractor basins in state space, and the consequent behavior of the network under damage. There are certainly ways of studying the effects of task structure in networks without resorting to damaging them. However, the current investigation has lead to a number of interesting insights into operation of attractors that may not have arisen in the course of more conventional connectionist research. Perhaps further analysis of the normal operation of attractor networks can serve to clarify these insights, in the same way that psychological experimentation with normal subjects can often elaborate the details of processes that are first isolated in neuropsychological investigations of brain-injured patients.

9.7 Future work

Only relatively recently have developments in both cognitive neuropsychology and connectionist modeling progressed to the point where attempts to develop detailed models of neuropsychological phenomena are feasible. While preliminary attempts in the area, including the present research, are quite promising, considerable work remains before a comprehensive account of normal and impaired cognition can be developed in any domain, including reading. With regard to the current investigation, each limitation that we have noted about aspects of the simulations could serve to motivate future research. Here we only mention avenues of research on a larger scale that we intend to pursue.

9.7.1 Implementing a dual-route model of reading

The current simulations reproduce many characteristics of patients who read via the semantic route, and simulations of Patterson et al. (1990), while less than satisfactory thus far, hold promise for reproducing the reading behavior of at least some patients who read via the phonological route. A natural next step is to develop an integrated model of reading containing both a semantic and a phonological route, along the lines of Seidenberg & McClelland's (1989) more general framework for lexical processing (see Figure 2.5). Developing such a "dual-route" model would allow a number of interesting issues to be investigated that cannot be addressed in the context of models of either route in isolation.

- What is the most effective means of developing the ability to read for meaning? Assuming that the associations between phonology and semantics have been previously established through speaking and listening, is the mapping from orthography to semantics best established by direct training or by explicitly learning spelling-to-sound correspondences? If both routes are trained simultaneously, do they initially specialize for particular types of words (e.g. regular vs. exception)?
- In normal reading, how is evidence from the two routes best combined in pronouncing a word? Is an output representation based on single phonemes sufficient or should additional phonological structure be encoded? What are the relative contributions of each route in processing particular types of words? Seidenberg and McClelland's model demonstrated that a phonological route *can* generate pronunciations for both regular and exception words, but in normal reading the semantic route may play a more important role in pronouncing the latter.
- What kinds of effects emerge from various types of partial damage in a model with two functional routes? The previous simulations of each isolated route could only consider neurological patients in which the unimplemented route is *entirely* non-functional, and yet

patients with pure neuropsychological syndromes are rare. The behavior of a dual-route model under partial damage would provide insight into a wider range of acquired dyslexics. For example, phonological alexics (Beauvois & Derouesné, 1979; Shallice & Warrington, 1980), who can read words but not nonwords (like deep dyslexics) but do not make semantic errors, may have an intact semantic route and only partial damage to the phonological route. Also, the behavior of non-semantic lexical readers (Sartori et al., 1987), who can read words (with poor comprehension) but not nonwords, may result from partial damage to both routes, thus obviating the introduction of a third, lexical non-semantic route (Morton & Patterson, 1980). In addition, a number of characteristics of surface dyslexics are not well-accounted for by the existing simulations (Behrmann & Bub, in press) and could be more fully investigated in a dual-route model.

- How can the intact operation of one route be used most effectively in mediating recovery from partial damage of the other route? What sort of retraining regime maximizes the effectiveness and generality of recovery?

The major challenge of developing such a dual-route model is in the design of an adequate semantic system that could represent the meanings of sufficiently many words as is required for developing the phonological route (see Seidenberg & McClelland, 1990).

9.7.2 Rehabilitation study

The relearning studies presented in the thesis relate to patient therapy only in the most general way. Much more could be learned from a detailed attempt to model the pattern of recovery of a particular patient. The current work makes predictions primarily about the influence of semantic variables, such as prototypicality, on relearning and generalization. Hence, the most appropriate type of patient would be one with semantic deficits, such as a global, Wernicke, or transcortical sensory aphasic. The intention is to develop a more elaborate version of the semantic system (in conjunction with the dual-route simulations described above), in which the effects of a wider range of semantic variables can be investigated than is possible in the current simulations. Predictions about the relative impact of these variables on relearning and generalization would then be tested by assessing the effectiveness of analogous patient therapy. A joint study involving patient therapy and modeling work, in collaboration with M. Behrmann, is currently being planned.

9.7.3 Modality-specific semantic impairments

The current simulations of visual naming in optic aphasia must be extended to account for the relative preservation of semantic categorization and gesturing before they can constitute a full simulation of optic aphasia. As suggested at the end of Chapter 8, a promising approach to developing such a simulation would involve a semantic system that interacted with multiple modalities for input (e.g. visual, auditory, tactile) and output (e.g. written, spoken, gestural). Under the pressure of differing amounts of structure in the possible input/output mappings, portions of semantics are likely to become particularly important for particular tasks. Under some types of damage, such a system would show dissociations of the form seen in optic aphasia or other types of modality-specific aphasias (e.g. bilateral tactile aphasia, Beauvois et al., 1978; auditory aphasia, Denes & Semenza, 1975; optic aphasia for colors, Beauvois & Saillant, 1985). In addition, such a simulation would extend the work of Farah & McClelland (1991) in accounting for modality-specific semantic impairments (e.g. Warrington & Shallice, 1984) and semantically-bounded anomias (e.g. Hart et al., 1985).

9.7.4 Pure alexia

The final major direction of research we intend to pursue involves modeling the reading behavior of patients with “pure alexia” (Dejerine, 1892). These patients are essentially unable to read via the normal means, and so resort to a “letter-by-letter” strategy. Once all or most of the individual letters have been identified, the word is successfully identified. On one account of this syndrome (Warrington & Shallice, 1980) the representations of words have been eliminated by damage, while on another (Patterson & Kay, 1982), word representations are intact but can no longer be accessed by letters in parallel. On both accounts, the letter-by-letter reading is a compensatory strategy, but the accounts differ as to whether this strategy requires mechanisms not normally involved in reading.

Following Patterson & Kay, we intend to model letter-by-letter reading in pure alexia by implementing and damaging a version of the model of visual-spatial processing put forward by Hinton (1981a; 1981b; Hinton & Parsons, 1988) in the domain of letter and word reading (also see Plaut, 1989). The model can recognize individual letters or words by mapping retina-based features onto object-based features via units that implement the appropriate retina-to-object coordinate transformations. The letter or word is then stored in a scene-based visual short-term memory via a second set of object-to-scene “mapping” units. If the retina-to-object mapping for a whole word is impaired, the network can still read the word by successively recognizing and storing each individual letter, and then accessing the object-based representation of the word top-down from the stored letters in the scene-based memory.

9.8 Conclusion

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 of both patient behavior and the nature of computation in connectionist networks.