BRAIN AND LANGUAGE **52,** 25–82 (1996)
ARTICLE NO. 0004

Relearning after Damage in Connectionist Networks: Toward a Theory of Rehabilitation

DAVID C. PLAUT

Carnegie Mellon University

Connectionist modeling offers a useful computational framework for exploring the nature of normal and impaired cognitive processes. The current work extends the relevance of connectionist modeling in neuropsychology to address issues in cognitive rehabilitation: the degree and speed of recovery through retraining, the extent to which improvement on treated items generalizes to untreated items, and how treated items are selected to maximize this generalization. A network previously used to model impairments in mapping orthography to semantics is retrained after damage. The degree of relearning and generalization varies considerably for different lesion locations, and has interesting implications for understanding the nature and variability of recovery in patients. In a second simulation, retraining on words whose semantics are atypical of their category yields more generalization than retraining on more typical words, suggesting a counterintuitive strategy for selecting items in patient therapy to maximize recovery. In a final simulation, changes in the pattern of errors produced by the network over the course of recovery is used to constrain explanations of the nature of recovery of analogous brain-damaged patients. Taken together, the findings demonstrate that the nature of relearning in damaged connectionist networks can make important contributions to a theory of rehabilitation in patients.

INTRODUCTION

Cognitive neuropsychology aims to extend our understanding of normal cognitive mechanisms on the basis of their pattern of breakdown due to brain damage in neurological patients. A major motivation for many researchers is that a more detailed analysis of the normal mechanism, and the way it is impaired in particular patients, may lead to the design of more effective therapy to remediate these impairments (Howard & Hatfield, 1987). Signifi-

I thank Marlene Behrmann, Gary Dell, Jay McClelland, Tim Shallice, and an anonymous reviewer for providing useful comments on an earlier version of this paper. The research was supported financially by the McDonnell-Pew Program in Cognitive Neuroscience (Grant T89-01245-016), the National Science Foundation (Grant ASC-9109215), and the National Institute of Mental Health (Grant MH47566). Address reprint requests to David C. Plaut, Department of Psychology, Carnegie Mellon University, Pittsburgh, PA 15213-3890. E-mail: plaut@cmu.edu.

0093-934X/96 $12.00

cant progress has been made in analyzing cognitive mechanisms and their impairments in terms of ''box-and-arrow'' information-processing models, particularly in the domain of written and spoken language (e.g., Coltheart, Patterson, & Marshall, 1980; Coltheart, Sartori, & Job, 1987; Patterson, Coltheart, & Marshall, 1985). However, relatively few remediation studies have been based directly on cognitive analyses, and while these few have been relatively successful, the specific contribution of the cognitive model is often unclear (for examples and general discussion, see Byng, 1988; Caramazza, 1989; Hillis, 1993; Riddoch & Humphreys, 1994; Margolin, 1992; Mitchum & Berndt, 1990; Patterson, 1994; Seron & Deloche, 1989; Wilson & Patterson, 1990). The limited usefulness of box-and-arrow diagrams in this regard may stem from a general lack of attention paid to specifying the actual representations and computations that perform the task (Seidenberg, 1988, but see Coltheart, Curtis, Atkins, & Haller, 1993, for a recent exception). As Hillis (1993, p. 6) states,

> Models of the language task to be treated and cognitive analyses to identify the individual patient's level(s) of damage within that language system, while necessary, are not sufficient for developing theory-based treatment approaches. . . . There is nothing within the models of normal cognitive processes that would *alone* support the introduction of specific intervention strategies for improving some particular cognitive deficit. That is, *in addition to* an indispensable theory of the normal cognitive processes underlying the skill that is impaired, we need a theory of rehabilitation of cognitive function in order to predict what sort of change might occur in those processes, and what sort of intervention would bring about such a change.

Connectionist or parallel distributed processing models offer an alternative framework within which to cast explicit computational theories of normal and impaired cognitive processes (see McClelland, Rumelhart, & the PDP Research Group, 1986; Quinlan, 1991; Rumelhart, McClelland, & the PDP Research Group, 1986). Within such models, information is represented as patterns of activity over large groups of simple, neuron-like units. Processing takes the form of cooperative and competitive interactions among the units on the basis of weighted connections between them. These weights encode the long-term knowledge of the system and are learned gradually through experience in the domain. The use of automatic learning procedures for adjusting the weights to solve a task has enabled detailed, implemented models of this form to be developed within a wide range of cognitive domains, including high-level vision and attention, learning and memory, speech and language processing, and the coordination and control of action.

More recently, researchers have begun to explore how the effects of damage in connectionist models of normal cognitive processes can help us understand the specific patterns of cognitive impairment that can arise following brain damage. Much of this work has been focused on deficits in word reading—the acquired dyslexias (Mozer & Behrmann, 1990; Patterson, Seidenberg, & McClelland, 1990; Plaut & Shallice, 1993a)—but there has also

been considerable work in other domains, including spelling (Shallice, Glasspool, & Houghton, 1995; Olson & Caramazza, 1994), speech production (Harley & MacAndrew, 1992; Martin, Dell, & Schwartz, 1994), face recognition (Burton, Young, Bruce, Johnston, & Ellis, 1991; Farah, O'Reilly, & Vecera, 1993), visual object naming (Gordon, 1982; Plaut & Shallice, 1993b), spatial attention (Cohen, Romero, Servan-Schreiber, & Farah, 1994; Humphreys, Freeman, & Müller, 1992), learning and memory (McClelland & Rumelhart, 1986; McClelland, McNaughton, & O'Reilly, 1995), semantic memory (Farah & McClelland, 1991; Horn, Ruppin, Usher, & Hermann, 1993), and control of action and responding (Bapi & Levine, 1990; Cohen & Servan-Schreiber, 1992; Levine & Prueitt, 1989). Although still in its infancy, the relative success of this work suggests that connectionist modeling may provide an appropriate formalism within which to explore how disorders of brain function give rise to disorders of cognition.

The current work attempts to extend the relevance and usefulness of connectionist modeling in neuropsychology to address issues in the rehabilitation of cognitive deficits following brain damage. The focus is on the domain of acquired dyslexia, as this is the area in which the most detailed neuropsychological and computational investigations have been done. The main issues to be addressed are the degree and speed with which behavior can be reestablished as a result of therapy, the extent that recovery due to treatment of particular items generalizes to other materials, and the possible bases on which to select items for treatment so as to maximize this generalization. These issues can be addressed naturally within a connectionist framework because the same learning procedures that support the acquisition of cognitive processes in the normal network can be applied to a damaged network to support the reacquisition of premorbid abilities. The goal of the current work is to demonstrate that an analysis of the correspondence between the nature of recovery in patients and in damaged networks can inform theories of normal cognitive processing as well as suggest specific approaches for improving patient therapy.

The next section summarizes a number of studies on remediation of acquired dyslexia based on cognitive models of normal reading, particularly those studies that relate to reestablishing the process of mapping from the written forms of words (orthography) to their meanings (semantics). Following this, previous preliminary work on the effects of relearning after damage in connectionist networks is described. Most of the simulations described in this paper are based on retraining a network that has been used previously to model impaired reading via meaning in a form of acquired dyslexia known as deep dyslexia (Hinton & Shallice, 1991; Plaut & Shallice, 1993a). The implemented network can be viewed as constituting part of a larger but unimplemented framework for lexical processing (Seidenberg & McClelland, 1989). The first simulation experiment shows that retraining produces rapid recovery and significant generalization after lesions to only some parts of

the network—those implementing a structured mapping—suggesting an explanation for the occurrence and variability of recovery among patients. In a second experiment, retraining on words whose meanings are somewhat atypical of their category yields more generalization than retraining on more typical words, suggesting a counterintuitive strategy for selecting items in patient therapy to maximize recovery. The final experiment investigates the changes in pattern of errors that occur during recovery in deep dyslexia, both at its onset and at its resolution into phonological dyslexia (Friedman, 1996; Glosser & Friedman, 1990). The findings provide constraints on how the nature of recovery of these patients can be explained within the more general lexical framework. The paper concludes with a critical evaluation of the relevance of these demonstrations for patient rehabilitation and a general discussion of the potential for principles of learning and relearning in connectionist networks to provide the basis for a theory of rehabilitation.

COGNITIVE REMEDIATION OF ACQUIRED DYSLEXIA

Cognitive rehabilitation is based on the notion that, by ascribing a patient's deficits to the selective impairment of one (or a few) of a set of functionally separable subsystems involved in carrying out a task, therapy can more effectively focus on the remediation of particular types of representations and processes. In essence, the cognitive analysis enables a more detailed diagnosis of the impairment in an information-processing framework, in order to provide a basis for the design of more appropriate therapeutic procedures.

Most models of word reading (e.g., Buchanan & Besner, 1993; Carr & Pollatsek, 1985; Coltheart, 1978, 1985; Coltheart et al., 1993; Morton & Patterson, 1980; Paap & Noel, 1991; Patterson & Morton, 1985; Reggia, Marsland, & Berndt, 1988; Seidenberg & McClelland, 1989; Shallice & McCarthy, 1985, but see Kay & Marcel, 1981; Van Orden, Pennington, & Stone, 1990, for exceptions) have (at least) two ways of pronouncing written words—that is, of mapping orthography to phonology. The first derives the meaning (semantics) of the word as an intermediate representation; the second bypasses semantics and maps directly from orthography to phonology. Although terminology varies considerably from model to model, we will call the first of these the *semantic* route (or reading via meaning), and the second, the *phonological* route. Some models (e.g., Coltheart et al., 1993; Morton & Patterson, 1980) further subdivide the phonological route into lexical and sublexical components, while others (e.g., Seidenberg & McClelland, 1989; Shallice & McCarthy, 1985) do not. As the focus of the current work is on the nature of impairment and recovery within the semantic route, this latter distinction is not of immediate concern (although see Plaut, McClelland, Seidenberg, & Patterson, in press, for relevant simulations and discussion).

The broad distinction between the semantic and phonological route is supported by the observation that each of these routes can be selectively im-

paired by brain damage. *Surface* dyslexic patients, particularly the ''fluent'' type, appear to have a selective impairment of the semantic route (see Patterson et al., 1985). They can use the phonological route to read words with regular spelling-sound correspondences (e.g., NEAR), as well as word-like pronounceable nonwords (e.g., KEAR), but often mispronounce exception words—particularly those of low frequency—typically giving the ''regularized'' pronunciation (e.g., PEAR read as ''pier''). By contrast, *deep* dyslexic patients appear to have an impairment of the phonological route, in that they are severely impaired at pronouncing nonwords (see Coltheart et al., 1980). They misread some words as well—often giving a semantically related response (e.g., PEAR read as ''apple'')—suggesting some additional partial impairment of the semantic route. *Phonological* dyslexic patients (Beauvois & Derouesné, 1979) also read words much better than nonwords, but do not make semantic errors. Such patients may have impairments that are qualitatively similar but less severe than those of deep dyslexic patients (Friedman, 1996; Glosser & Friedman, 1990).

Most work in reestablishing reading via meaning has involved surface dyslexic patients, as, among acquired dyslexic patients, they typically have the most severe impairment of the semantic route. Below I briefly review the findings of a number of rehabilitation studies that attempt to treat various types of impairments of the semantic route in these (and related) patients. The focus is not so much on how effective the therapy is for items that are treated directly (although this is certainly an important concern), but rather on *generalization*: the extent to which the therapy leads to improvement on untreated but related items. A rehabilitation program can have the broadest impact if it can engender recovery that goes beyond the specific items presented during treatment.

Coltheart and Byng (1989)

Coltheart and Byng (1989, see also Byng & Coltheart, 1986) undertook a remediation study with a surface dyslexic patient, EE, a 40-year-old left-handed postal worker who suffered left temporal-parietal damage from a fall. On the basis of a number of preliminary tests administered about 6 months after the accident, they determined that EE had a specific deficit in deriving semantics from orthography. The most important indication was the occurrence of homophone confusions (i.e., misunderstanding TALE as TAIL). These errors make sense if the patient is unable to map orthography to semantics directly, but rather must map orthography first to phonology (where TALE and TAIL are indistinguishable) and then to semantics.

To improve the patient's ability to associate the written form of words directly with their meanings, Coltheart and Byng designed a study involving words containing the spelling pattern -OUGH (e.g., THROUGH, COUGH, BOUGH), which have highly irregular pronunciations and, thus, are difficult to read

without semantics. EE was retrained on 12 of 24 such words, by having him study the written words augmented with mnemonic pictures for their meaning (e.g., a picture of a tree drawn on the word BOUGH). Prior to therapy, 4 of the treated words were read correctly; after therapy, all 12 were read correctly. In addition, the *untreated* words also improved, from 1 correct prior to therapy, to 7 correct after therapy. Thus, the improvement in the untreated set (6 words) was almost as large as the improvement in the set that was actually treated (8 words). Generalized improvement of untreated words in this domain is surprising because a word and its meaning are arbitrarily related—it is unclear why relearning the meanings of some words should help reestablish performance on other words with unrelated meanings.

A useful measure of generalization is the amount that untreated items improve relative to the amount that they would have improved if they had been treated directly. This measure can be approximated by the ratio of the improvement on untreated items to the improvement on the treated items. Thus, Coltheart and Byng's therapy with EE produced $6/8 = 75\%$ generalization.

In a second study, EE was given the 485 highest-frequency words (Kucera & Francis, 1967) for oral reading. The 54 words he misread were divided in half randomly into treated and untreated sets. EE again learned to read the treated words by studying cards of the written words augmented with mnemonics for their meanings. As a result, his reading performance on the treated words improved from 44 to 100% correct. Once again, the untreated words also improved, from 44 to 85% correct (41% untreated improvement, 56% treated improvement: 73% generalization). This improvement was not due to ''spontaneous recovery'' nor to other nonspecific effects because performance on the words was stable both before and after therapy. A third therapy study, involving the next 388 words in the frequency norms, produced broadly similar results. Overall, Coltheart and Byng found excellent recovery of treated items and substantial generalization to untreated items (also see Weekes & Coltheart, in press, considered under General Discussion).

Scott and Byng (1989)

In a similar study, Scott and Byng (1989) attempted to remediate homophone confusions in another surface dyslexic, JB, a 24-year-old student nurse who suffered a closed-head injury resulting in left temporal damage. The treatment involved selecting the correct homophonic word from six alternatives so as to meaningfully complete each of 136 sentences. Over the course of 29 sessions, performance improved from about 75% to nearly perfect. A second task was administered pre- and post-therapy, in which each of 270 homophones (135 pairs, half of which were treated during therapy) was embedded in both an appropriate and an inappropriate sentence, and JB had

to sort the resulting 540 sentences accordingly. Similar to the Coltheart and Byng study, JB showed therapy-specific improvement for sentences containing both the treated words and, to a lesser but still significant extent, the untreated words. However, Scott and Byng failed to find improvement in JB's *writing* of either set of homophones in sentence contexts, suggesting that generalization occurred within but not between orthographic domains.

Behrmann (1987)

Behrmann (1987) carried out an analogous study on CCM, a 53-year-old surface *dysgraphic* patient with an impairment in writing on the basis of semantics, due to a left temporal-parietal stroke. Given that the nature of the relationship between orthography and semantics is broadly the same in writing as in reading, it seems likely that similar rehabilitation strategies should apply, although the degree to which reading and writing share underlying representations remains a matter of debate (see Behrmann & Bub, 1992; Coltheart & Funnell, 1987; Weekes & Coltheart, in press). Behrmann used picture-matching and sentence-completion tasks to train CCM to produce the appropriate spelling of 25 of 69 homophone pairs (e.g., BREAK/BRAKE) that she initially spelled incorrectly. Therapy improved overall performance from 49 to 67%, but no improvement on the untreated homophone pairs was observed. However, the writing of 75 words with irregular spellings (e.g., COMB) did improve significantly as a result of the therapy. Thus, Behrmann's study produced some type of generalization, but not specifically to other homophonic items.

Hillis (1993)

Hillis (1993) carried out an extensive rehabilitation program with an acquired dyslexic patient, PS, who suffered a closed-head injury resulting in left temporal-parietal damage. Preliminary testing on a variety of lexical tasks (Hillis & Caramazza, 1991) suggested that he had multiple impairments within the reading system: to orthography, to semantics, as well as to the phonological route. A therapy program to reestablish the phonological route, based on training of explicit grapheme–phoneme correspondences (e.g., c→ /k/ after A, O, or U), substantially improved both word and nonword reading although, when retested one year later, only the former remained stable. Furthermore, there was no generalization to the correspondences of untreated graphemes (also see Berndt & Mitchum, 1994; de Partz, 1986; Hillis, 1990).

To reestablish orthographic representations, PS was trained on a lexical decision task involving 50 words that he had failed to read, along with 50 nonwords that each differed from one of the words by a single letter. Over the course of 10 sessions, PS learned to perform perfectly on the lexical

decision task, but showed minimal improvement in his oral reading and comprehension of the words. When subsequently trained to pronounce the words, he improved from 22 to 92% correct, but showed no generalization in pronouncing untreated words.

In a further study, PS was presented with a series of homophones and their definitions, and asked to write the word in a sentence. Over the course of 10 therapy sessions, performance on the treated words in oral reading, comprehension, and spelling from definition improved. For untreated homophones, only oral reading improved, and only very slightly. Specifically, oral reading of treated words improved from 53.3% correct (averaged over 3 pretherapy sessions) to 96.6% correct (averaged over the last 8 therapy sessions), while untreated words improved only from 56.6 to 62.2% correct (12.9% generalization).

A final study involved 60 words whose spellings have multiple plausible pronunciations (e.g., BEAR; -EAR can also be pronounced as in BEER). PS was trained to match half of the words to their written or spoken definitions (among five alternatives), and the other half to written or spoken rhyming words (among five alternatives). For each set of words, when performance on the trained task reached 100% correct, performance on the untrained task was also 100% correct. Furthermore, oral reading of all of the words improved from an average of 68% correct to 100% correct. Thus, there was considerable generalization from trained to untrained *tasks*. Unfortunately, performance on untreated *words* was not measured.

Behrmann and Lieberthal (1989)

Finally, Behrmann and Lieberthal (1989) studied a 57-year-old man, CH, who was globally aphasic following a stroke affecting the left frontal, temporal, and parietal lobes, as well as the internal capsule. CH performed particularly poorly on tasks requiring semantic knowledge; for example, he was able to choose the synonym of a target word only when a distractor was semantically distant (e.g., GLOVE: MITTEN/PLATE vs. MITTEN/SOCK). Performance on written and spoken words was equally impaired.

The therapy study involved a semantic-category sorting task using written and spoken items from six categories: animals, body parts, colors, transport, furniture, and food. Half of the items from three categories (transport, body parts, furniture) were treated in a number of ways, including training CH on written and spoken word/picture associations, and on picking out a target item from an array of distractors when given a semantic cue. The therapy improved overall performance on treated items from 20 to 93% correct. Furthermore, there was significant generalization to untreated transport items (58% generalization)[1] and body parts (81% generalization) but not to un-

[1] Based on a reanalysis of Behrmann and Lieberthal's (1989) data, showing correct performance on treated transport items improved from 8/20 per-therapy to 20/20 post-therapy, while untreated transport items improved from 6/20 pre-therapy to 13/20 (instead of the reported 7/20) post-therapy.

treated furniture items. Among untreated categories, only foods showed significant improvement (from 17.5 to 52.5% correct; 48% generalization). Thus, Behrmann and Lieberthal found significant generalization within some but not all treated and untreated categories.

Taken together, the therapy studies that have attempted to reestablish reading (or writing) via the semantic route have succeeded in improving performance for treated items, and often also for untreated but related items. The degree of recovery and generalization, however, varied considerably across patients, although direct comparison is difficult because different word sets, tasks, and remediate techniques were employed. Why some patients improved while others did not is not entirely clear. Furthermore, even in those patients who did improve and showed generalization, the cause of this generalization—in terms of changes to the underlying cognitive mechanism induced by treatment—is unknown. An explanation of the effects seen in patient therapy in this domain should account not only for the occurrence of generalization in some patients and conditions, but also for its absence in others.

Hillis and Caramazza (1994; Caramazza, 1989; Hillis, 1993) have argued that traditional information-processing models offer little guidance in the design of an appropriate rehabilitation program because they make no contact with the wide variety of factors that can influence whether and to what extent a given patient benefits from therapy. Rather, a theory of rehabilitation must be based on a detailed specification of the representations and processes that perform a given cognitive task normally, how they are impaired by brain damage in a particular patient, and how they are affected by possible interventions.

Connectionist modeling provides a framework for developing computationally explicit models of normal cognitive processes, and damage to such models has been shown to replicate the cognitive impairments of some types of brain-damaged patients. The goal of the current work is to explore whether such models can also provide a basis for understanding the effects of rehabilitation on impaired cognitive processes, and how best to design such rehabilitation. The next section reviews some preliminary investigations of the effects of retraining connectionist networks after damage. These studies demonstrate how retraining a damaged network on only a subset of its premorbid knowledge can induce improvement not only on this subset, but also on the remaining untreated knowledge. The subsequent simulation work attempts to apply the insights from these studies more directly to the domain of rehabilitation of impaired reading via meaning.

RELEARNING AFTER DAMAGE IN CONNECTIONIST NETWORKS

I begin with a brief overview of the nature of processing and learning in connectionist networks, before reviewing studies of relearning after damage. As described briefly in the Introduction, a connectionist network is composed of a large set of simple, neuron-like processing units that interact across

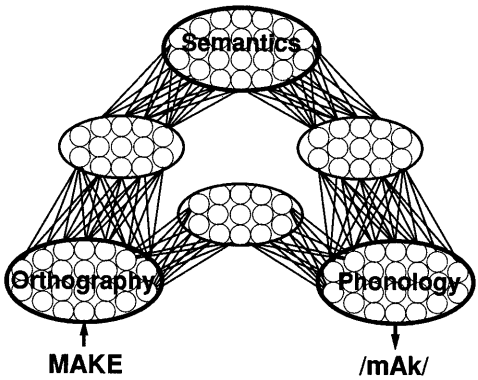

FIG. 1. A framework for lexical processing, including groups of units that represent the orthography, phonology, and semantics of written words, and intermediate or *hidden* groups of units that mediate between these representations (based on Seidenberg & McClelland, 1989, p. 526).

weighted connections. In most networks, units are organized into groups or layers that represent different types of information about the item being processed. For instance, one group might represent the written form of a word, another might represent its spoken form, and yet another might represent its meaning (see Fig. 1, following Seidenberg & McClelland, 1989). Within each group, each item is represented by the combined activity of a number of units, and each unit participates in the representation of a number of items. This type of *distributed* representation (Hinton, McClelland, & Rumelhart, 1986) contrasts with a *local* representation in which each item is represented by the activity of a single, corresponding unit (e.g., Feldman & Ballard, 1982; McClelland & Rumelhart, 1981).

Processing takes the form of simultaneous cooperative and competitive interactions among the units. For example, in pronouncing a written word, input is presented to the network by setting the states of the orthographic units to the pattern of activity that represents the written form of the word. These unit states then affect the states of other units via the positive and negative weights on the connections between them. Units with positive connections tend to support each other, while units with negative weights between them inhibit each other. At some point, the semantic and phonological units begin to change their states in response to information from the orthographic units (via the intermediate units). As a result of the cooperative and competitive interactions, the network as a whole settles gradually into a configuration of active and inactive units that represents the network's interpretation of the orthographic input. Included in this configuration is a pattern of activity over the semantic units representing the meaning of the word, and a pattern over the phonological units representing its pronunciation. In this way, the network computes the pronunciation of a written word, as well

as its meaning, by a parallel settling process governed by the connection weights.

The weights in the network are adjusted gradually on the basis of experience with the written forms, the spoken forms, and the meaning of words. Initially, the weights are set to small random values. As a result, after processing a written word, the patterns of semantic and phonological activity are quite different from the word's correct meaning and pronunciation. These discrepancies are quantified by an error measure, typically the sum of the squared difference between each generated activity and its correct value. An automatic learning procedure (e.g., back-propagation; Rumelhart, Hinton, & Williams, 1986) calculates how each weight in the network contributes to the error for the word. By changing the weights so as to reduce the error, the semantic and phonological patterns will be closer to the correct patterns on the next presentation of the word. As these changes accumulate over the repeated exposure to many words, the weights gradually acquire values that enable the network to settle into the correct semantic and phonological representations when given the orthography of each word.

To understand learning (and relearning) in networks, it may help to think of a multidimensional space with a dimension for each weight. This may be easiest to imagine for a network with only two weights (see Fig. 2). Each point in this space—a plane in two dimensions—defines a set of weights that produces some amount of error if used by the network. If this error is represented along an additional dimension corresponding to height, then the error values of all possible weight sets form an *error surface* in weight space. A good set of weights has low error and corresponds to the bottom of a valley in this surface. At any stage in learning, the network can be thought of as being at the point on the error surface, with a height given by the error for the current set of weights. Possible weight changes consist of movements in different directions along the surface. Changing each weight in proportion to its effect on the error (i.e., the partial derivative of the error with respect to the weight) amounts to moving in the direction of steepest descent, also called gradient descent.

In summary, the type of connectionist network that we are concerned with represents items such as words as distributed patterns of activity over various groups of units. Processing an item involves cooperative and competitive interactions among units causing the network as a whole to gradually settle into a global configuration of active and inactive units constituting its interpretation of the item. This settling process is governed by connection weights that are learned gradually on the basis of exposure to the correct representations of each item using an error-correcting, gradient descent algorithm.

Hinton and Sejnowski (1986)

When items are given distributed representations, similar items—those represented by similar (overlapping) patterns—tend to have similar effects

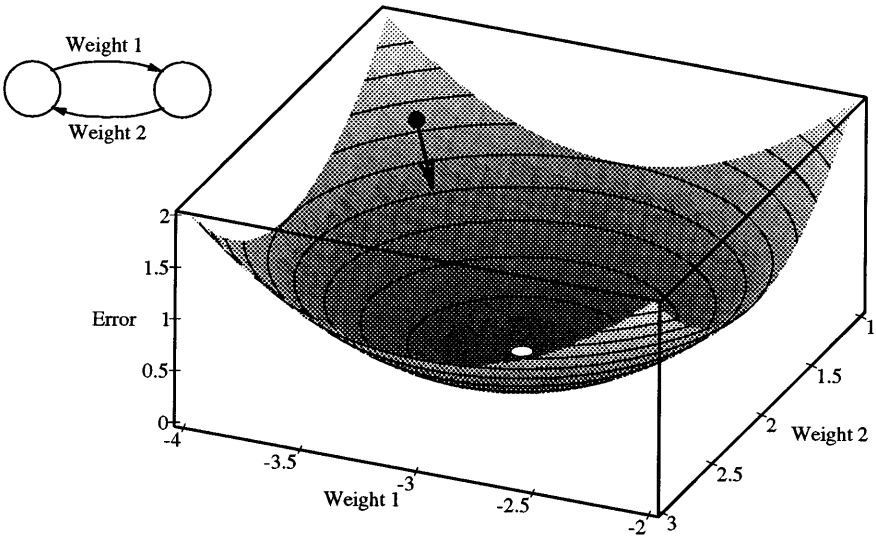

FIG. 2. A possible error surface for a network with only two weights. The height of the surface above each combination of weights indicates the error that would result if those weights were used by the network. The black dot corresponds to the error for the current set of weights; the arrow indicates the optimal direction of weight change—the one causing the steepest decent in error. The learning procedure operates by calculating this direction and taking a small step along it, thereby changing the weights slightly. As this process is repeated, the weights eventually achieve values that produce the minimum amount of error (the white dot; Weight 1 = −3 and Weight 2 = 2 in this example).

on other parts of the system. Consequently, such representations are particularly useful in mapping between domains that are systematically related, such as orthography and phonology. In English, similarly spelled words tend to have similar pronunciations, and the use of distributed representations helps to capture the underlying regularities (Plaut & McClelland, 1993; Plaut et al., in press; Seidenberg & McClelland, 1989). However, for the same reason, distributed representations would seem to be ill suited for mapping between arbitrarily related domains, such as orthography and semantics. There is no inherent or structured relationship between the written form of a (monomorphemic) word and its meaning: orthographic similarity is unrelated to semantic similarity. It is in this situation that local representations—word-specific units—appear to be necessary (see Hinton et al., 1986, for discussion).

Hinton and Sejnowski (1986) challenged this assumption by demonstrating that distributed representations can be effective and efficient in mapping between even arbitrarily related domains. The task they chose was a highly simplified version of the mapping from orthography to semantics: each of 20 three-letter words was to be associated with an arbitrary semantics consisting of a random subset of 30 semantic features. The network used to

accomplish the mapping had three layers of units, corresponding roughly to the orthography-to-semantics portion of Fig. 1. Thirty *grapheme* units, in three groups of 10, represented the three letters of each word. These units were fully connected to 20 *intermediate* units, which in turn were fully connected to 30 *sememe* units, one for each semantic feature. In addition, the sememe units were fully interconnected. The units produced stochastic binary output and all connections were symmetric (i.e., of equal strength in both directions). The network was trained with the Boltzmann Machine learning procedure (Ackley, Hinton, & Sejnowski, 1985) to settle into the correct pattern of activity over the sememe units for each word when the grapheme units for the letters of the word were clamped on.

After training, the undamaged network performed the task almost perfectly, but when single intermediate units were removed, 1.4% of the semantic patterns produced by the network were incorrect. Interestingly, 59% of these incorrect responses were the exact semantics of an alternative word, and these ''word'' errors were more semantically and/or visually similar to the correct word than would be expected by chance. This co-occurrence of visual and semantic similarity in error responses is analogous to aspects of the error pattern produced by deep dyslexic patients (also see Hinton & Shallice, 1991; Plaut & Shallice, 1993a).

In addition to demonstrating visual and semantic influences in the errors produced by the damaged network, Hinton and Sejnowski investigated the behavior of the network in relearning after damage. Specifically, after the network had learned all 20 associations of three-letter strings to semantic patterns, it was damaged in a variety of ways, either by zeroing or adding noise to the weights, or by removing hidden units. After each of these types of damage, when the network was retrained on all of the associations, its performance improved much more quickly than when it was learning the associations originally and had reached the same level of performance (see Fig. 3). Sejnowski and Rosenberg (1987) later replicated this rapid relearning in NETtalk, a feed-forward back-propagation network for mapping orthography to phonology.

An even more interesting effect found by Hinton and Sejnowski was that rapid recovery of all associations occurred even if the network was retrained on only a subset of them. After adding random noise to the weights on connections to and from the hidden units, they retrained the network on only 18 of the 20 associations. As expected, performance on these 18 associations improved rapidly. More surprising was that performance on the two unretrained associations also improved: from 30 to 90% correct (over 50 presentations of each—note that the network is stochastic and thus can produce a variety of responses to repeated presentations of the same stimulus). In a second experiment, performance on a pair of associations with higher error rates improved from 17 to 98% correct with retraining on only the remaining 18 associations. Unfortunately, the effect was not very robust—when only

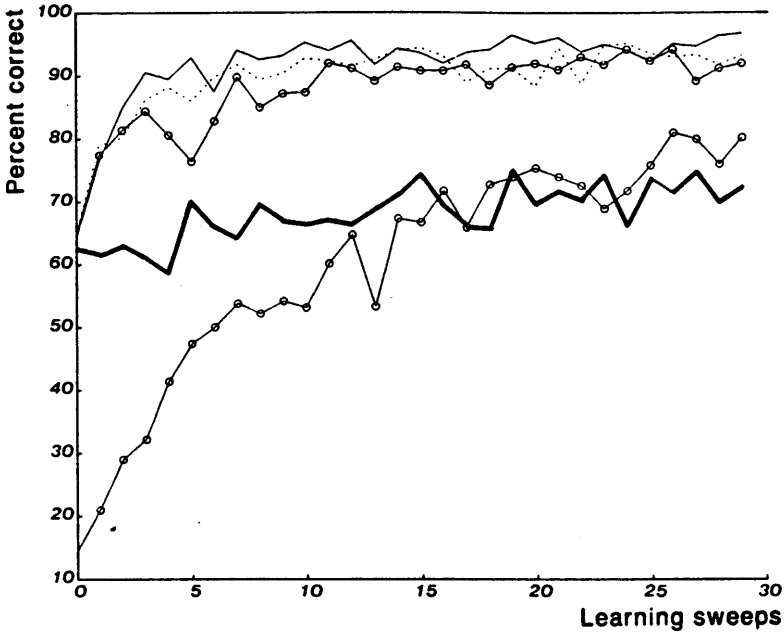

FIG. 3. The recovery of performance of the Hinton and Sejnowski network after various types of damage. The heavy line is a section of the original learning curve after a considerable number of learning sweeps. All the other lines show recovery after damaging the network once learning was complete (99.3% correct). The lines with open circles show the rapid recovery after 20 or 50% of the weights to the hidden units have been set to zero (but allowed to relearn). The dashed line shows recovery after five of the 20 hidden units have been permanently ablated. The remaining solid line is the case when uniform random noise between ± 22 is added to all the connections to the hidden units. In all cases, a successful trial was defined as one in which the network produced *exactly* the correct semantic features when given the grapheme input. (Reprinted from Hinton & Sejnowski, 1986, p. 311.)

15 associations were retrained, performance on the remaining 5 deteriorated slightly. Even so, the demonstration of even modest recovery of unretrained associations in a connectionist network raises the possibility that the unexplained generalization found in some rehabilitation studies with acquired dyslexic patients may occur for similar reasons.

Hinton and Plaut (1987)

Hinton and Plaut (1987) investigated the generalization effect further in a network in which each connection has both a slow weight and a fast weight. The slow weights are like those normally used in connectionist networks—they change slowly and encode all of the long-term knowledge of the network. The fast weights change much more rapidly but continually decay toward zero, so that their values are determined solely by the recent past.

The effective weight on a connection when computing unit activities is the sum of its slow and fast weight. Thus, at any instant, the network's knowledge consists of long-term knowledge (in the slow weights) that captures the inherent regularities in the environment, with a temporary overlay of short-term knowledge (in fast weights) that compensates for particular characteristics of the current context.

One benefit of using fast weights is that they can learn to cancel out the interference in a set of old associations caused by more recent learning. To demonstrate this, Hinton and Plaut built a fully connected feed-forward network with slow and fast weights, which had 10 input units, 100 hidden units, and 10 output units. The network was trained with back-propagation on 100 associations of random binary vectors of length 10, in which each component of each vector had probability 0.5 of being a 1. Although both the slow and fast weight on each connection experience the same pressure to change to reduce the error, most initial learning occurs in the fast weights because they can change more quickly. Their strong tendency to remain small, however, prevents them from solving the task completely by themselves. The slow weights gradually learn under the pressure of the residual error. As the slow weights learn to accomplish more and more of the task, the fast weights can decay further. In this way, knowledge is gradually transferred from the fast to slow weights (from the short-term context to long-term knowledge). Finally, at the end of learning, the network performs the task perfectly, the fast weights are near zero, and all of the knowledge is in the slow weights.

Once the network had learned the 100 associations in this way, Hinton and Plaut trained it on 5 new random associations without further rehearsal of the original 100. Training was continued until all knowledge of the new associations was in the slow weights. As these associations were unrelated to the original ones, the weight changes they induced had the same effect on the original associations as adding random noise to the weights. Thus, performance on the original task was significantly impaired as a result of the interference training (also see McCloskey & Cohen, 1989; Ratcliff, 1990). The network was then retrained on only half of the original 100 associations. Not only did the retrained associations recover quickly, but performance on the remaining unretrained associations improved almost as much: generalization, as measured by the ratio of unretrained to retrained improvement, was 83%. In fact, there was considerable generalization (65%) when only 10% of the original associations were retrained (see Fig. 4).

As the recovery of performance occurs during the first few retraining sweeps, it takes place almost entirely within the fast weights. Nothing in the interaction between fast and slow weights is required for the transfer effect—an analogous network with only slow weights would also exhibit it. The advantage of using both fast and slow weights is that the relearning in the fast weights need not permanently interfere with the new associations—if the fast weights are allowed to decay back to zero, the new knowledge is

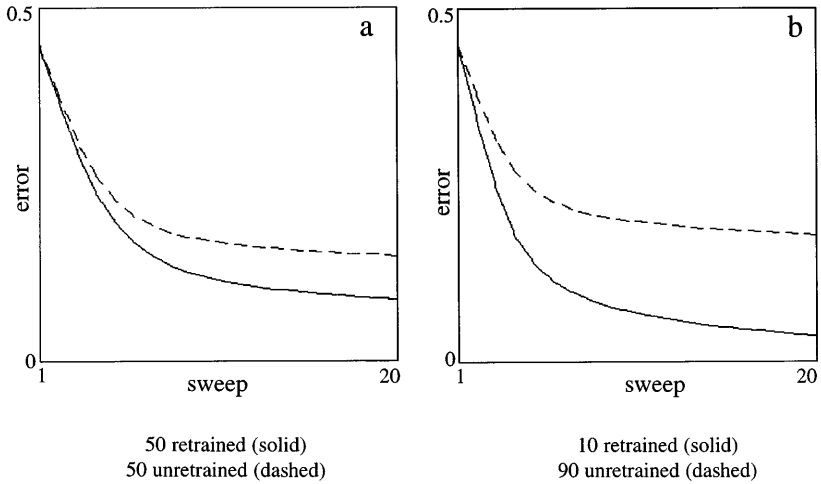

FIG. 4. Average error per output unit for the retrained (solid) and unretrained (dashed) subsets when retraining on (a) 50 or (b) 10 of the original 100 associations. Notice that error rather than correct performance is being plotted, so improved performance is reflected by *decreasing* curves. (Reprinted from Hinton & Plaut, 1987, p. 181.)

restored. However, if retraining is continued, the knowledge will be gradually transferred into the slow weights with less interference to the new associations.

The generalization to unretrained associations found by Hinton and Sejnowski (1986) and Hinton and Plaut (1987) is somewhat paradoxical as there is no systematicity within or between any of the associations. However, during the original learning the weights capture whatever chance regularities there happen to be among the entire set of associations. Most of these regularities still hold for the retrained subset, and so during relearning the weights tend to move back toward values that capture these regularities. Since most of the regularities apply to the unretrained associations as well, they also improve. In other words, because knowledge of the associations is distributed across all of the connections, when the damaged network is retrained on some of the associations (or on other associations that share the same structure), all of the weights are pushed back toward their original values.

The spontaneous recovery of unretrained associations in networks is analogous to the improvements on untreated items found in many rehabilitation studies with acquired dyslexic patients (as summarized in the previous section). Thus, the basis of generalization in networks provides a possible explanation for its occurrence in patients. More generally, the nature of relearning after damage in connectionist networks may provide a useful framework for understanding the effects seen in the remediation of acquired dyslexia. One potentially important caveat, however, is that the generalization experiments

of Hinton and Sejnowski (1986) and of Hinton and Plaut (1987) involved relearning after corrupting weights with noise rather than after permanent damage. The explanations offered, in terms of weight changes back toward the original set of weights, do not strictly apply to a lesioned network, as all of the original weights are no longer available. In this case, relearning must adjust the weights on the remaining connections to new values in order to compensate for the missing connections.

The relevance of these previous retraining simulations is limited further by the fact that the instantiations of the tasks they used bear only a very abstract relationship to the tasks actually performed by acquired dyslexic patients. Although Hinton and Sejnowski modeled their task loosely after the relationship between orthography and semantics, the representations they used captured very little of the structure within either of these domains. Hinton and Plaut demonstrated more impressive generalization results but used a random association task with even less similarity to actual reading for meaning. The implications of the nature of recovery in damaged networks for patient rehabilitation would be far better established in a simulation that corresponded more closely to the tasks carried out by the patients. To this end, the current work investigates the effects of recovery in networks (Hinton & Shallice, 1991; Plaut & Shallice, 1993a) that, when damaged, replicate the qualitative error pattern exhibited by patients with impaired reading via meaning.

Hinton and Shallice (1991)

Based on the previous work by Hinton and Sejnowski (1986) reviewed above, Hinton and Shallice (1991) trained a recurrent back-propagation network to map from the orthography of 40 three- or four-letter words to a simplified representation of their semantics, described in terms of 68 predetermined semantic features. Although their version of the task is still highly simplified, it reflects the essential properties of orthographic and semantic representations: words with similar spellings have similar orthographic representations, words with similar meanings have similar semantic representations, and orthographic similarity is unrelated to semantic similarity (see Plaut & Shallice, 1993a).

After training the network to correctly derive the meanings of all of the words, Hinton and Shallice systematically lesioned it by removing proportions of units or connections, or by adding noise to the weights. They found that the damaged network occasionally settled into a pattern of semantic activity that satisfied the response criteria for a word other than the one presented. These errors were more often semantically and/or visually similar to presented stimuli than would be expected by chance. While the network showed a greater tendency to produce visual errors (e.g., CAT→''cot'') with lesions near orthography and semantic errors (e.g., CAT→''dog'') with lesions near semantics, both types of error occurred for almost all sites of

damage. This pattern of errors is similar to that of patients with deep dyslexia (Coltheart et al., 1980).

More recently, Plaut and Shallice (1993a) have extended these initial findings in two ways. First, they established the generality of the co-occurrence of semantic, visual, and mixed visual-and-semantic errors by showing that it does not depend on specific characteristics of the network architecture, the learning procedure, or the way responses are generated from semantic activity. Second, they extended the approach to account for many of the remaining characteristics of deep dyslexia, including the effects of concreteness/imageability and their interaction with visual errors, the occurrence of visual-then-semantic errors, greater confidence in visual as compared with semantic errors, relatively preserved lexical decision with impaired naming, and the existence of different subvarieties of deep dyslexia.

The replication of the diverse set of symptoms of deep dyslexia through unitary lesions of the network strongly suggests that the underlying computational principles of the network capture important aspects of the process of mapping orthography to semantics in humans. Extending this claim further, relearning in the lesioned network would be expected to produce effects similar to those observed in empirical studies of the rehabilitation of this process in brain-damaged patients. The following experiments test this claim.

EXPERIMENT 1: EFFECT OF LESION LOCATION ON RELEARNING AND GENERALIZATION

The first experiment investigates relearning and generalization after damage in a network that maps orthography to semantics. The issue to be addressed is whether such recovery depends on the nature of the damage to the network—specifically, on the location of damage within the system. The network used is broadly similar to the one developed by Hinton and Shallice (1991). It is actually a replication of one of the networks that Plaut and Shallice (1993a) used—termed the *40–60* network—to demonstrate that Hinton and Shallice's qualitative results do not depend on specific details of the network architecture. The original version of this network could not be used in the present study due to technical details of the training procedure.[2]

[2] The original network was trained with *momentum*, such that each weight change consisted of the newly calculated changes along with a proportion of the previous weight change. As a result, each weight had a tendency to continue to change in the same way as on past updates. While there is nothing wrong in principle with using momentum during retraining, it was not used in the current study to allow a fair comparison between the rates of relearning vs. original learning at the same level of performance. As there are no previous weight changes at the beginning of retraining, momentum would have a different effect in this condition than it would at the corresponding point during original learning, when weight changes would have accumulated over many previous epochs.

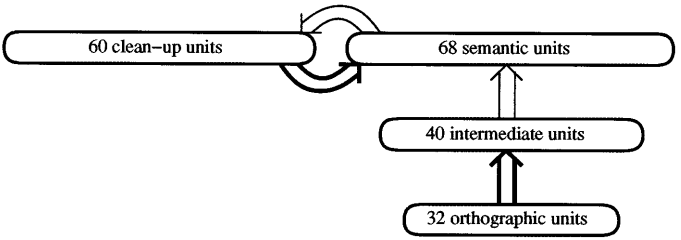

FIG. 5. The architecture of the network used to model impaired reading for meaning, based on the one used by Hinton and Shallice (1991). Arrows represent sets of connections between groups of units—those in bold are lesioned in the present experiment.

In all important respects, however, the two versions of the network are equivalent.

The goal of the experiment is to account for the extent and variability of recovery and generalization found in rehabilitation studies of acquired dyslexic patients (e.g., Behrmann, 1987; Coltheart & Byng, 1989; Scott & Byng, 1989). Note that, as most of the patients studied are surface dyslexic (or dysgraphic), the network is not intended to capture all aspects of their behavior. In particular, such patients are assumed to have an intact phonological route in addition to an impaired semantic route. By contrast, the network has only a semantic route. Nonetheless, given that the phonological route is thought to be normal in surface dyslexic patients, it is assumed that *recovery* in these patients is due primarily to changes in the impaired semantic route. To the extent that this is true—and data below suggest it may be only an approximation—an implementation of an isolated semantic route provides an adequate vehicle for exploring the nature of recovery in these patients.

Methods

Network architecture and task definition. The architecture of the network is shown in Fig. 5. Like the Hinton and Sejnowski (1986) network, it can be viewed as a specific implementation of the orthography-to-semantics portion of Seidenberg and McClelland's (1989) general framework for lexical processing (see Fig. 1). The network has two main pathways: (a) a *direct* pathway, from 32 orthographic units to 68 semantic units via 40 intermediate units, and (b) a *clean-up* pathway, from the semantic units to 40 clean-up units and back to the semantic units. Each set of connections (shown as large arrows in Fig. 5) consists of only a randomly chosen 25% of the possible connections between the two groups of units. The logic of this architecture is that the direct pathway generates initial semantic activity from visual (orthographic) input, while the clean-up pathway iteratively refines this initial activity into the exact correct semantics of the word. At any point in time during this settling process, the activity level (or state) of each unit, ranging between 0.0 and 1.0, is a smooth, nonlinear (logistic) function of its total weighted input from other units.

The basic relearning effects described in this paper do not depend critically on very specific aspects of the network architecture. A network with a rather different architecture, in which the functions of the direct and clean-up pathways are subserved by the same sets of units

(Plaut & Shallice's, 1993a, *80fb* network), produces qualitatively equivalent results (see Plaut, 1991).

The task of the network is to generate the semantics of 40 words from their orthography. The words come from five semantic categories: indoor objects, outdoor objects, animals, body parts, and foods. The meaning of each word is specified in terms of a pattern of activity over 68 semantic units, each of which corresponds to a particular semantic feature. For instance, among the semantic features of CAT are *has-legs, mammal, found-on-farms,* and *does-run.* On average, the semantic representation of a word contains about 15 of the 68 possible semantic features (see Hinton & Shallice, 1991, for details).

In the orthographic representation, each letter within a word is represented in terms of a separate set of eight units corresponding to particular orthographic features (e.g., *contains a vertical stroke, horizontally symmetric*). These features were designed so that visually similar letters (e.g., E and F) are represented by similar (overlapping) patterns. As each word has at most four letters, a total of 32 orthographic units are required (see Plaut & Shallice, 1993a, for details).

The representations over the two remaining groups of units—the intermediate and clean-up units—are not determined by the definition of the task. Rather, the network develops its own representations over these units (by adjusting the weights on their incoming connections) under the pressure of learning to activate the correct semantics of each word when presented with its orthography.

Training procedure. The network was trained in the following way. The connection weights were initialized to small random values. Each word was presented to the network by setting the states of the orthographic units to the appropriate input pattern based on the letters of the word, and by initializing the states of all other units to 0.2. The network was then run for eight iterations, in which each unit updated its state each iteration based on the current states of other units and the weights on connections from them. At the end of these iterations, the network had produced a pattern of activity over all of the units, including the semantic units. As the weights are initially random at the beginning of training, the pattern of semantic activity produced by the word was very different from its correct semantics. An iterative version of the back-propagation learning procedure, known as *back-propagation through time* (see Rumelhart, Hinton, & Williams, 1986; Williams & Peng, 1990, for details) was used to compute the way that each weight in the network should change so as to reduce this difference for the last three iterations. These weight changes were calculated for each word in turn, at which point the accumulated weight changes were carried out (scaled by a learning rate of 0.01) and the procedure was repeated. After 4740 *epochs* (i.e., training presentations of all 40 words), when the network was presented with each word, the resulting activity of each sememe unit over the last three iterations was within 0.2 its correct value for that word, at which point training was considered complete.

Lesioning and retraining procedures. The trained network was lesioned by randomly selecting a specified proportion of the connections between two groups of units and removing them from the network. Each of the four main sets of connections was subjected to lesions of a wide range of severity, in which 5 to 70% of the connections were removed. For each combination of location and severity, 20 instances of lesion (i.e., removal of a specific random subset of connections) were administered.

After a given instance of lesion, the network was presented with each of the 40 words for processing. As a result of the damage, the semantic activity produced by the network would often differ significantly from the correct semantics of the presented word. The network was considered to have responded correctly if the *proximity* (i.e., normalized dot product) of the semantics generated by the network was within 0.8 of the correct semantics of the presented word, and the proximity of the next best word was at least 0.05 further (see Hinton & Shallice, 1991, for details). If the generated semantics satisfied these criteria when compared with the semantics of some word other than the one presented, that word was considered to be the network's response (an error). Otherwise, the network was considered to have failed to respond (an omission). The response criteria can be thought of as substituting for an actual *output*

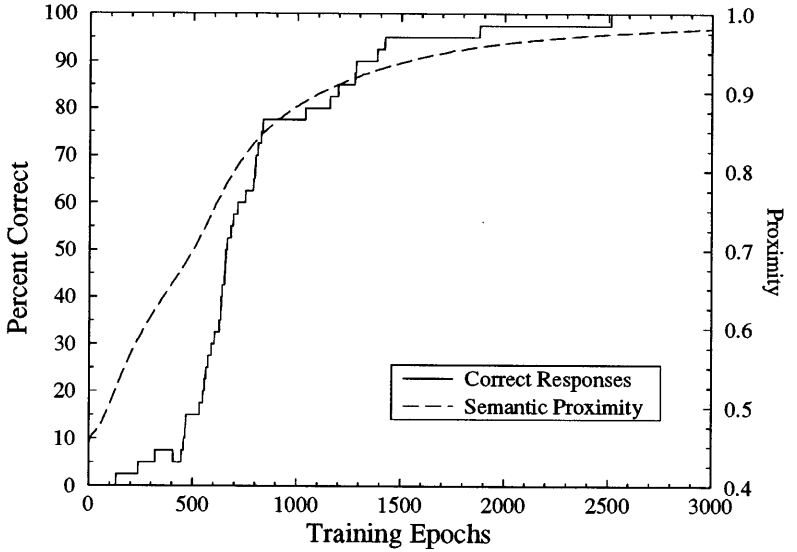

FIG. 6. Improvement in performance of the network when originally learning the task, measured in terms of percent correct and average proximity of the generated and correct semantic patterns.

network that would generate explicit pronunciations on the basis of semantic activity (see Plaut & Shallice, 1993a, for implementations of this process, and for evidence that criteria based on semantic proximity produce qualitatively equivalent behavior).

Once the performance of the lesioned network on all 40 words was determined, half of the correct words and half of the incorrect words were randomly selected and placed in the *treated* set; the remaining words were placed in the *untreated* set. Thus, both the treated and untreated sets contained 20 words and were balanced for correct performance. For the purpose of setting up the treated and untreated sets, explicit errors and omissions were both considered incorrect and were not distinguished.

The lesioned network was then retrained for 50 epochs on the treated words only. Performance was measured at each epoch during relearning separately for the treated and untreated word sets, in terms of the number of words read correctly, and the average proximity of the generated and correct semantics. To ensure that any relearning effects were not simply due to an imbalance in initial performance between the treated and untreated sets, the two sets were exchanged and the retraining was repeated, starting from the same initial set of weights. Thus, each group of words served both as the treated set and the untreated set. Finally, for purposes of comparison, the weights were again reinitialized and the lesioned network was retrained on all 40 words.

Results and Discussion

Figure 6 shows the performance of the network over the course of originally learning the task, in terms of the average proximity of the generated and correct semantics, and the percentage of words read correctly using the response criteria. Notice that the network shows a rapid improvement in correct performance at about 500 epochs, when performance has reached

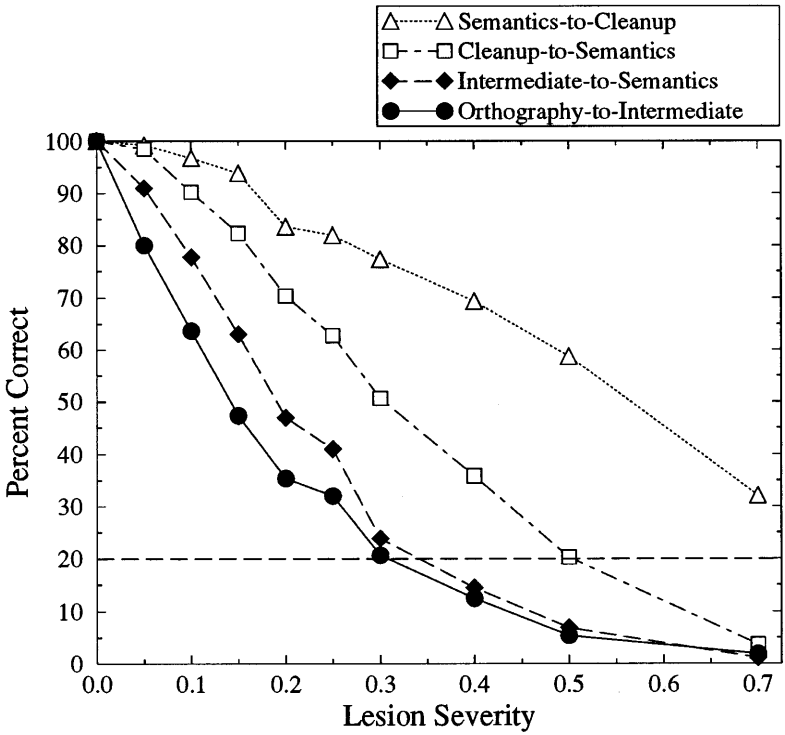

FIG. 7. Correct performance rates after lesions to each main set of connections as a function of lesion severity. Each data point is the average of 20 instances of lesion (i.e., removal of a particular random subset of connections) at that location and severity. The dashed line at 20% correct performance intersects the data points for the two lesion conditions tested in the current relearning study.

around 20%, although it only achieves 100% correct performance at epoch 2513. Also notice that the network continues to learn after this point—as mentioned above, the training criteria are not satisfied until epoch 4740. The semantics generated by the network may be sufficient to distinguish the presented word from the possible alternatives while still being somewhat inaccurate. A more direct reflection of the accuracy of the generated semantics is given by the proximity measure. As Fig. 6 shows, this measure rises more gradually over the course of learning, indicating that improvements in the network's representations during learning do not always translate directly into more accurate overt performance (see Karmiloff-Smith, 1992, for discussion of related phenomena in cognitive development).

Figure 7 shows the correct performance of the fully trained network after lesions to each main set of connections, as a function of lesion severity (percent connections removed). Not surprisingly, performance gradually declines with increasing lesion severity for all lesion locations. The network's perfor-

mance is most disrupted by lesions to the orthography-to-intermediate connections, and least disrupted by semantics-to-cleanup lesions.

Given the available computational resources, the retraining procedure could not be applied to every instance of lesion at every location and severity, as it requires about 250 times more computation than the procedure for measuring performance.[3] Accordingly, as an initial comparison, relearning was tested after lesions at the semantic level (cleanup-to-semantics connections) vs. after lesions near orthography (orthography-to-intermediate connections). These two sets of connections are highlighted in bold in Fig. 5. For each, a severity of lesion was selected which lowered correct performance to near 20%: 50% of cleanup-to-semantics connections (20.2% correct), and 30% of orthography-to-intermediate connections (20.6% correct). This level of performance falls (just) within the range of performance included in the error analyses in previous work (Plaut & Shallice, 1993a), while being sufficiently poor to provide room for significant improvement over the course of retraining.

Cleanup-to-semantics lesions. We first focus on the effects of retraining after 50% lesions of the cleanup-to-semantics connections. Figure 8 shows data from three different training conditions: (1) improvement on all 40 words when the network was originally learning the task and had reached 20% correct performance (epoch 557, see Fig. 6), (2) improvement on all 40 words when retraining on them after damage, and (3) improvement on the treated and untreated word sets when retraining only on the treated set. A comparison of the first two of these conditions (the triangles in Fig. 8) reveals that performance improves much more dramatically when retraining on all of the words after damage then when originally learning them. After 50 epochs of retraining, average correct performance on all 40 words is near perfect (97.9%). In fact, performance increases from 20.2 to 73.5% correct in the first 5 epochs (paired $t_{19} = 24.5$, $p < .001$). By contrast, original learning improved to only 30% correct over the full course of 50 epochs. Notice that this apparently slow original learning is actually occurring during the most rapid rise in performance when learning the task (see Fig. 6). The much more rapid recovery of performance after removal of connections replicates the findings of Hinton and Sejnowski (1986) in retraining after corrupting the weights with random noise.

The more important comparison is the relative improvement on treated vs. untreated word sets (the circles in Fig. 8). When retraining for 50 epochs on the 20 treated words, performance on them improves from 20.2 to 98.4%

[3] For a given lesion, measuring performance requires a single (forward) pass through the network for each of the 40 words. Retraining requires, for each of 50 epochs, 2 passes (forward and backward) for each of the 20 retrained words, 1 (forward) pass for each of the 20 unretrained words, run 2 times with the word sets exchanged, plus 2 (forward and backward) passes for each of the 40 words when retrained together, for a total of 10,000 passes.

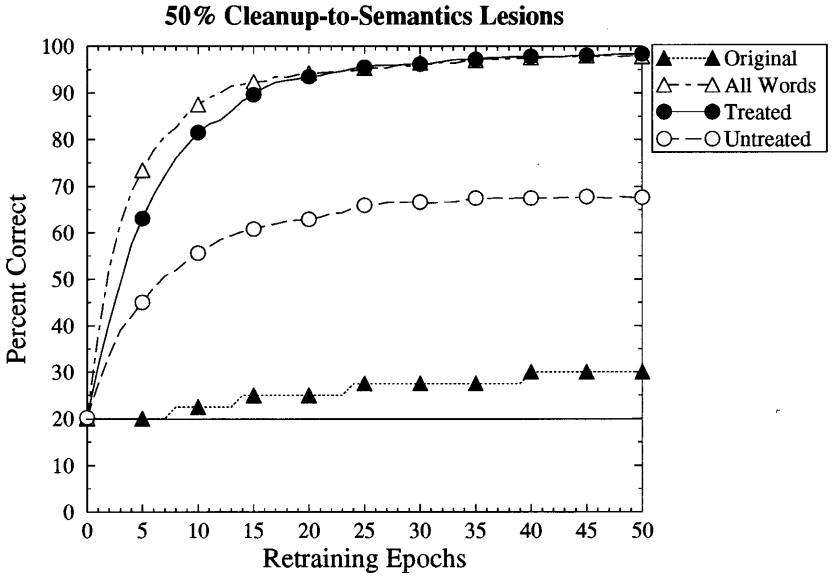

FIG. 8. Improvement in correct performance on the treated and untreated word sets as a function of number of epochs of retraining on the treated word set after removing 50% of cleanup-to-semantics connections. Results are averaged over 20 instances of such lesions and across exchanges of the treated and untreated sets. Also included is the improvement in performance over 50 epochs when retraining on all 40 words, and when the network was learning the task originally and had reached the same overall level of performance (indicated by the solid horizontal line).

correct (paired $t_{39} = 49.2$, $p < .001$). Concurrently, performance on the untreated words improves from 20.2 to 67.6% (paired $t_{39} = 24.0$, $p < .001$). Thus, generalization, as measured by the ratio of improvement on the untreated words to improvement on the treated words, is 60.6%. That is to say, performance on sets of 20 words improved 60.6% as much from training on other words as it did when training on the words themselves. This magnitude of generalization is comparable to that found by Coltheart and Byng (1989) in their rehabilitation study of acquired dyslexic patient EE with impaired access to semantics from orthography.

Orthography-to-intermediate lesions. A very different pattern of results obtains after 30% lesions of the orthography-to-intermediate connections (see Fig. 9). Although relearning all 40 words after damage is still considerably faster than original learning, it is much less effective than after cleanup-to-semantics lesions (53.9% vs. 77.6% improvement, respectively; paired $t_{19} = 7.14$, $p < .001$). Even though the data is averaged across 20 lesion instances, considerable variability in performance over the course of recovery is still apparent. This contrasts sharply with the smooth, rapid relearning of all 40 words after cleanup-to-semantics lesions.

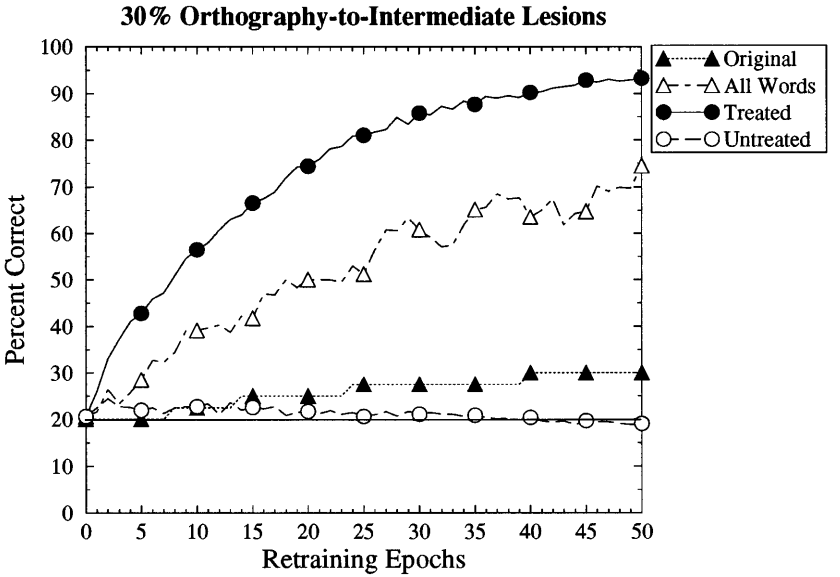

FIG. 9. Improvement in correct performance on the treated and untreated word sets as a function of number of epochs of retraining on the treated word set after removing 30% of orthography-to-intermediate connections.

Perhaps even more striking is the lack of generalization from treated to untreated word sets. While retraining on the treated words improves their performance from 20.6 to 93.3% correct (paired $t_{39} = 79.6, p < .001$), performance on the untreated words deteriorates slightly but not reliably: from 20.6 to 19.1% correct (−2.1% generalization; paired $t_{39} = 1.03$, $p > .3$). Thus, after orthography-to-intermediate lesions, retraining on a subset of the words *interferes* with the performance on other words. This lack of generalization is not due simply to the increased variability in retraining with an excessive learning rate. Retraining with half as large a learning rate (0.005) produces smoother, more gradual relearning curves but still yields no generalization (see Plaut, 1991).

In summary, retaining after lesions at the semantic level (cleanup-to-semantics connections) yields rapid relearning of treated items and substantial generalization to untreated items. By contrast, relearning after lesions near orthography (orthography-to-intermediate connections) produces much slower relearning of treated items and no generalization to untreated items. A clue to the cause of this difference in relearning and generalization comes from comparing the speed of relearning all 40 words vs. only 20 (treated) words after the two types of lesion. After orthography-to-intermediate lesions, relearning 40 words is much slower than learning only 20 words, whereas the opposite is true after cleanup-to-semantics lesions. This makes

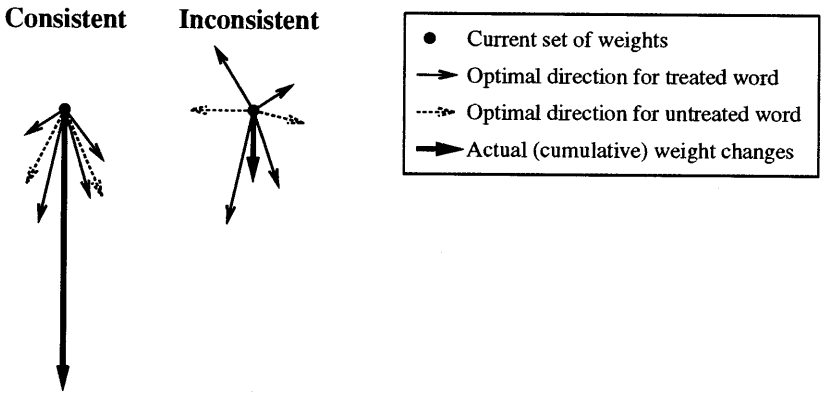

FIG. 10. Depiction of the effect of consistent vs. inconsistent weight changes on the extent of recovery and generalization in relearning. In each condition, the small solid arrows represent directions of weight change induced by treated words; the large solid arrow is the (vector) sum of these smaller arrows, representing the actual weight changes administered to the network. The length of this vector reflects the speed of relearning the treated words. The dotted arrows represent directions of weight change that would be optimal for untreated words if they were trained—to the extent that these point in the same direction as the actual weight-change vector, retraining on the treated words will also improve performance on the untreated words.

sense if the weight changes induced by retrained words after cleanup-to-semantics lesions are more consistent across words than after orthography-to-intermediate lesions. The actual weight changes administered to the network after a retraining epoch are the sum of the weight changes induced by each individual word (scaled by the learning rate). Weight changes that are consistent across retrained words accumulate, resulting in fast learning; weight changes that are inconsistent cancel each other out, resulting in much slower learning.

Another way of thinking about this effect is in terms of the relationships between different directions of movement in weight space (see Fig. 10). For each word, there is a particular direction of weight change that would be optimal if training solely on that word. This direction can be represented as a vector starting from the point corresponding to the current set of weights. Across the entire training set, words tend to have somewhat different optimal directions, given that their input–output requirements differ. The actual weight changes administered to the network is another vector in weight space that is the sum of the vectors for individual words. The cosine of the angle (proximity) between this actual weight-change vector and that for a particular word is an approximate measure of the degree to which the weight change will improve performance on that word. If the vectors for the set of retrained words point in different directions, their average proximity with the vector sum will be low, and will decrease as the number of vectors contributing to

the sum (i.e., retrained words) increases. Thus, average performance when relearning 20 words will be faster than when relearning 40 words. However, if the vectors for words are consistent, the vector sum will have a high proximity with each of them, as well as a large magnitude. In this case, relearning will be rapid overall, with the more words the better.

Greater consistency of the weight-change vectors for treated words after cleanup-to-semantic lesions than after orthography-to-intermediate lesions also explains why generalization to untreated words occurs after the former but not the latter. Just as for treated words, there is a direction of weight change that would be optimal for each untreated word (i.e., that would reduce the error on that word most rapidly; see the dotted arrows in Fig. 10). Although these directions do not affect the learning process (as the words are not presented for retraining), their relationship to the actual weight-change vector determines the degree of generalization—the extent to which retraining on the treated items also improves performance on the untreated items. If the directions for untreated words are similar to those for treated words, they will also be similar to the (vector) sum of the directions for treated words—the actual weight-change direction. Consequently, retraining will tend to help performance on the untreated words. If the directions for untreated words are unrelated (orthogonal) to the actual weight-change direction, retraining on the treated words will have little if any effect on the untreated words.

Thus, the amount of generalization in relearning is determined by the degree of consistency among the optimal weight-change directions for treated and untreated words. But recall that words were assigned to the treated or untreated sets arbitrarily (balancing correct performance). This means that any nonarbitrary relationship among treated words also holds between treated and untreated words. In particular, if the weight-change directions among treated words tend to be consistent, the directions for treated and untreated words also tend to be consistent. Thus, the greater consistency of weight changes across treated words after cleanup-to-semantic lesions than after orthography-to-intermediate lesions implies greater consistency of these weight changes with those that would be optimal for untreated words, and hence, greater generalization.

Why should the weight changes when relearning words be more consistent after cleanup-to-semantics lesions than after orthography-to-intermediate lesions? As described above, the degree of relearning and generalization depends on the consistency of the weight changes (i.e., directions of movement in weight space) that would be optimal for individual words. Although this is typically described in terms of the degree of overlap in the distributed representations of words, it depends more precisely on the regularity or structure in the mapping between input and output patterns. Viewed as an abstract task, there is no systematic structure in mapping from written (monomorphemic) words to their meanings; orthographic similarity is unrelated to seman-

tic similarity. However, when instantiated in a network, the task is broken down by the learning procedure into a number of separate transformations carried out by different parts of the network. These transformations constitute ''subtasks'' that may differ considerably in their degree of structure.[4]

Consider the relative roles of the cleanup-to-semantics connections and the orthography-to-intermediate connections (see Fig. 5). The cleanup-to-semantics connections serve to refine the initial semantic activity generated by the direct pathway into the exact correct semantics of the presented word. Since semantically similar words require similar clean-up, this subtask is highly structured (i.e., similar inputs map to similar outputs). By contrast, the orthography-to-intermediate connections must generate intermediate representations that can be transformed into semantic representations by the intermediate-to-semantics connections. The second half of this process is facilitated to the extent that the intermediate representations for words are semantically organized (i.e., words with similar meanings have similar intermediate representations). Thus, the role of the orthography-to-intermediate connections is to transform input patterns that are orthographically organized into intermediate patterns that are as semantically organized as possible. Because orthographic similarity is unrelated to semantic similarity, there is little structure in this subtask. In general, the findings suggest that the effectiveness of relearning after a lesion reflects the degree of structure in the mapping carried out by the lesioned connections.

Notice that, to the extent that the orthography-to-intermediate connections succeed in generating semantically organized representations, the subtask of the intermediate-to-semantics connections *does* have some structure: similar input (intermediate) patterns tend to correspond to similar output (semantic) patterns. Thus, under the current hypothesis, retraining after intermediate-to-semantics lesions should give rise to an intermediate level of recovery and generalization. Accordingly, additional data were gathered to test this prediction.

Intermediate-to-semantics lesions. As Fig. 7 shows, lesions to 30% of the intermediate-to-semantics connections reduces correct performance to near 20% (23.9% correct). Accordingly, the relearning procedure described above was applied to the 20 instances of intermediate-to-semantics lesions at this level of severity. The results are shown in Fig. 11.

When retraining after intermediate-to-semantics lesions, improvement on the treated words improves at a rate that falls between the rates for the other two lesion locations (see Figs. 8 and 9). For instance, after 10 epochs of

[4] This characterization of the operation of the network in terms of ''subtasks'' should *not* be interpreted as implying that these subtasks correspond to the separate, discrete subcomponent processes in information-processing models (e.g., Morton & Patterson, 1980). In the network, all of the transformations contribute simultaneously, continually, and interactively in producing behavior (see McClelland, 1987; Plaut, 1995, for discussion).

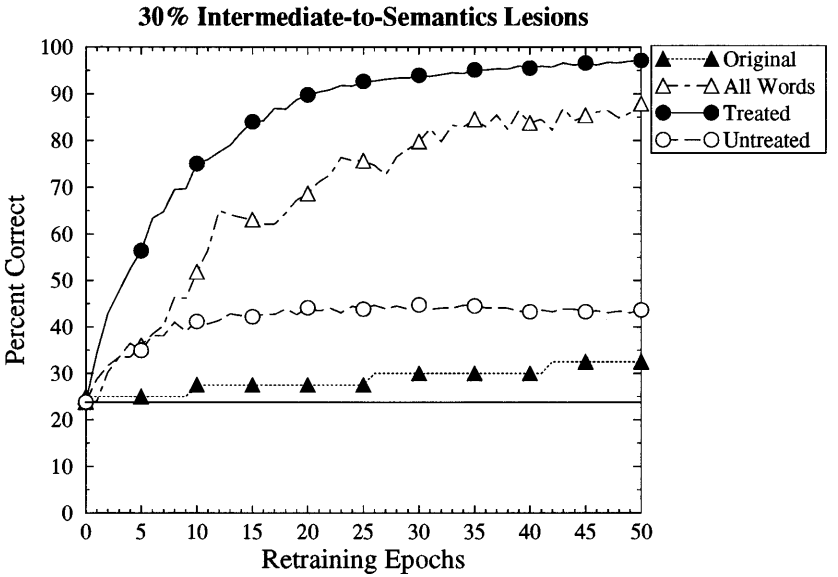

FIG. 11. Improvement in correct performance on the treated and untreated word sets as a function of number of epochs of retraining on the treated word set after removing 30% of intermediate-to-semantics connections.

retraining, correct performance on the treated words is 56.5% after orthography-to-intermediate lesions, 75.0% after intermediate-to-semantics lesions, and 81.5% after cleanup-to-semantics lesions (pairwise comparisons: $t_{78} = 6.53$, $p < .001$; and $t_{78} = 2.99$, $p = .004$, respectively). The same relative ordering holds when retraining on all 40 words (39.1, 51.9, and 87.5% correct, respectively, after 10 epochs of retraining; $t_{38} = 3.29$, $p = .002$; and $t_{38} = 10.1$, $p < .001$, respectively). Finally, correct performance on the untreated words improves 27.0% as much when retraining on the treated words (from 23.9 to 43.6% correct; paired $t_{39} = 10.1$, $p < .001$) as when retraining on the words themselves (from 23.9 to 97.1% correct; paired $t_{39} = 50.9$, $p < .001$). This level of generalization falls between the absence of generalization found after orthography-to-intermediate lesions (−2.1%) and the considerable generalization found after cleanup-to-semantics lesions (60.6%; pairwise comparisons: $t_{78} = 8.48$, $p < .001$; and $t_{78} = 9.95$, $p < .001$, respectively). Thus, as predicted, relearning after lesions to a set of connections that perform a mapping with an intermediate level of structure gives rise to an intermediate degree of relearning and generalization.

The learning curve for the untreated words shows a slight effect of *overlearning*: once performance on the treated set reaches a certain point (at about epoch 30), continued retraining on this set begins to degrade the previous recovery of performance on the untreated set. This effect is also present

when retraining after orthography-to-intermediate lesions, beginning at an earlier point in recovery (about epoch 10; see Fig. 9), although the initial improvement on untreated words is small to begin with. Even though overlearning is not a serious problem in the current context, after only 50 epochs of retraining, it would become increasingly deleterious if training were continued indefinitely. More generally, overlearning is a standard problem when the information in the available training examples underconstrains the parameters of an optimization procedure, such as learning in a connectionist network. An operational approach for preventing overlearning, known as *cross-validation* (Morgan & Bourlard, 1990), is to observe performance on a set of examples drawn from the same task as the trained examples but not actually used directly in learning. Training is halted when performance on the *untrained* set peaks. Cross-validation may be a useful technique in patient therapy if the goal is to maximize overall performance within an entire domain rather than performance solely on the treated items.

Distribution of weight changes. An important issue that remains to be addressed concerns the nature of the changes in the network that underlie recovery. Traditional theorizing in cognitive neuropsychology often assumes that the effects of damage on the operation of the cognitive system are ''local,'' in that only the damaged component is affected—the rest of the system continues to operate normally (see Farah, 1994, and the accompanying commentaries). A fully adequate consideration of the effects of damage must incorporate reasonable assumptions about the nature of the changes in the system that occur after the damage. The current analysis considers whether changes in the functionality of a network during retraining are restricted to the locations of damage. Specifically, is the effect of retraining to reestablish the mapping at the lesioned location by adjusting only the weights on the remaining connections at that location, or does it produce changes throughout the network that compensate for the effects of damage? A similar question arises in the interpretation of recovery in patients: Are improvements in performance due to the reestablishment of the normal function of the damaged location, or are other parts of the system adjusting to compensate for the damage? Note that this is a somewhat different issue from whether, and under what conditions, the *goal* of therapy should be to teach alternative, compensatory strategies for coping with cognitive deficits, rather than to attempt to remediate the impaired process itself (see Wilson, 1989, for discussion).

To investigate the basis of recovery in the network, the amount of change in the weights at each location in the network was calculated when retraining after each type of lesion. Figure 12 shows, for each location within the network, the average amount that each weight at that location changed when retraining on the 20 treated words, as a function of the location of lesion. The most general finding is that, regardless of lesion location, *all* of the remaining connections in the network, including those far removed from the

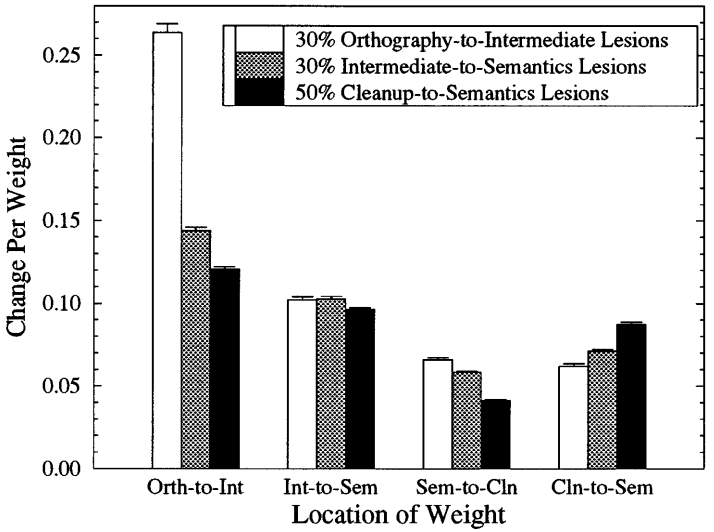

FIG. 12. The average weight change per connection for different sets of connections when relearning the treated words after various types of lesion.

lesion, adjust their weights to some degree to compensate for the effects of the damage. Thus, the network is not simply reestablishing the impaired mapping; ''normal'' portions of the network are changing their function somewhat to improve the performance of the network as a whole.

Overall, the amount of weight changes during relearning depends on the location of lesion ($F_{2,117} = 158.6$, $p < .001$). Weights change an average of 0.123 after 30% orthography-to-intermediate lesions, but only 0.086 after 50% cleanup-to-semantics lesions. Furthermore, different sets of weights change more than others ($F_{3,351} = 3698$, $p < .001$). Orthography-to-intermediate weights change most (mean 0.176) while cleanup-to-semantics weights change least (mean 0.055). There is, in addition, a significant interaction of the effects of lesion location and weight location on the amount of weight change ($F_{6,351} = 683.0$, $p < .001$). Thus, the location of lesion does affect the distribution of weight changes throughout the network that occur during relearning. In general, the form of this interaction is that the location of weight changes is biased toward the location of lesion. For instance, as can be seen in Fig. 12, when relearning after orthography-to-intermediate lesions, the weights on the remaining connections at the lesion site change much more than they do when relearning after the other lesions. This is also true after cleanup-to-semantics lesions, although not to the same extent. Thus, the lesioned location does take on a disproportionate amount of the weight changes during relearning, but the remaining portions of the network also compensate to reduce the behavioral impairment due to damage.

Summary

A network was trained to map from the written forms of words to their meanings. When retrained after damage, the degree to which the network relearned the treated words, and showed generalization to untreated words, depended on the location of the damage. For lesions at the level of semantics, the network showed rapid relearning and substantial generalization; for lesions near orthography, the network showed much slower relearning and no generalization. Lesions to an intermediate location produced intermediate results.

The findings suggest that the effectiveness of relearning after a lesion reflects the degree of structure in the mapping carried out by the lesioned connections. However, relearning does not simply reestablish the function of the damaged location. Although there is a general bias toward changing weights at the location of lesion, the undamaged parts of the network also change their weights to varying degrees to compensate for the effects of damage.

As reviewed earlier, studies of cognitive rehabilitation of acquired dyslexics in the domain of reading (or writing) via meaning have demonstrated considerable relearning of treated items and (often) improvement on untreated but related items. Relearning after lesions to a network that operates in the same domain results in similar qualitative effects, although the magnitude of the effects depends on the particular location of damage. Thus at a general level, the cause of rapid relearning and generalization in the network—distributed representations and structure in subtasks—may provide an explanation for the nature of recovery in these patients.

At a more specific level, the finding that the extent of relearning depends on the location of damage may provide an explanation for why only some patients show substantial recovery and generalization. The simulation results suggest that a patient with a functional impairment close to or within semantics should show considerable generalization, while one with an impairment close to orthography should show little or none. Conversely, the degree of generalization observed in a patient can be used to predict the fine-grained location of their functional impairment *within* the semantic route. These implications are developed more fully under General Discussion.

EXPERIMENT 2: SELECTING TREATED ITEMS TO MAXIMIZE GENERALIZATION

Ideally, an understanding of the impairment in a particular patient should lead to the design of a rehabilitation strategy that maximizes recovery. A potential benefit of connectionist modeling in neuropsychological rehabilitation is that it can provide a framework for investigating the relative effectiveness of alternative rehabilitation strategies. The first experiment investigated the effect of lesion location and severity, which are not under the control of

the therapist. One aspect of a retraining simulation that is under experimental control, and that might influence the nature of recovery, is the selection of items for treatment. This selection was done randomly in the first experiment, subject to the constraint that correct performance on the treated and untreated word sets was balanced. It seems reasonable to suppose, however, that retraining on particular subsets of words might lead to greater recovery. Maximizing generalization is particularly important in this regard, as improvement on only the treated items may be of limited usefulness to the patient (but see Hillis, 1993).

The first experiment demonstrated that the amount of generalization observed when retraining after a lesion is strongly influenced by the regularity or structure of the mapping carried out by the lesioned part of the network. In a network that maps orthography to semantics, this structure corresponds to the degree to which word representations over a set of units are semantically organized. The findings suggests a strategy of selecting treated words on the basis of the nature of their semantic representations. Specifically, a set of treated words should produce good generalization if, collectively, they provide a good estimate of the relevant semantic structure of the entire set of words.

An important aspect of the structure of semantic representations, at least of nouns, is that they are organized into categories. Furthermore, relative to this category structure, a critical semantic variable is typicality—how close a concept is to the central tendency of its category (Rosch, 1975). For instance, a robin is highly typical among birds, an eagle is less typical, and a penguin is highly atypical. The current experiment tests the hypothesis that the degree of generalization produced by retraining is influenced by the relative typicality of the treated words.

The question is, is it better to retrain on typical or atypical words? A natural intuition is that relearning the central tendency of a category—that is, retraining on typical words—should lead to the greatest generalization to other words in the category. The results of the experiment, however, show the opposite: retraining on words that are somewhat atypical of their semantic category leads to greater generalization than retraining on more typical words. The reason, put briefly, is that atypical words collectively convey more information on the overall structure of the category—specifically, on how semantic properties can *vary* across category members—while still providing a good approximation of the central tendency of the category.

Unfortunately, the training set in Experiment 1 contains only eight words in each of five categories, which is too limited to allow a reasonable investigation of this effect (see Plaut, 1991, for results and discussion). Accordingly, the current experiment involves training a new network on a more artificial version of the task of reading via meaning, in which it is possible to have greater control over the relationships among the semantic representations of words.

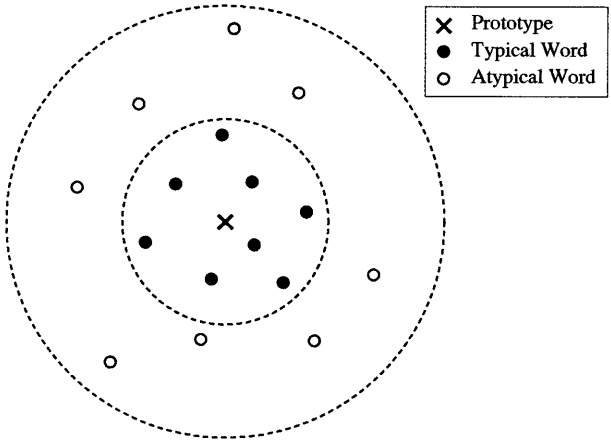

FIG. 13. A depiction of the relationship in semantic space between the prototype of a category and typical vs. atypical exemplars in that category.

Method

Network architecture and task definition. The architecture of the network has the same general structure as the network from the first experiment (see Fig. 5). The direct pathway maps orthography to semantics via 40 intermediate units, and the semantic units reciprocally interact with 60 clean-up units. The differences are that, in the current network, there are only 20 orthographic units (compared with 32), 50 semantic units (compared with 68), and 50% of the possible connections between groups of units are included (compared with 25%).

The task that the network is trained on does not directly correspond to mapping from written words to their meanings. Rather, it is an artificial task that captures some of the important statistical properties of the real task—namely, that orthographic similarity is unrelated to semantic similarity, and that semantic representations have a category structure.

The training set consists of 100 artificial ''words.'' Each word was arbitrarily assigned an orthographic representation consisting of a random binary pattern over 20 orthographic features, such that each feature has a probability $p = 0.2$ of being active. To generate semantic representations for the words, a single semantic prototype was created by randomly setting each of 50 semantic features to be active with probability $p = 0.2$. The representations for individual words were then generated by randomly changing some of the features of the prototype (Chauvin, 1988). The degree of typicality of a word is reflected in the number of features that its representation differs from the prototype—typical words share most of the features of the prototype, while atypical words share far fewer. To implement this, two sets of 50 word meanings were generated from the prototype using different levels of random distortion. The *typical* set consisted of instances produced by a small distortion of the prototype—each semantic feature had a probability $d = 0.1$ of being randomly regenerated (with $p = 0.2$). The *atypical* set consisted of instances generated using a large distortion ($d = 0.5$). Geometrically, if the prototype corresponds to a particular point in the space of semantic representations, the typical words are points that are near the prototype, while the atypical words are farther away (see Fig. 13).

Training procedure. As in Experiment 1, the network was trained with back-propagation through time to generate the correct semantic representation of each word over the last three of eight iterations when presented with its orthographic representation. A somewhat lower

learning rate (0.002) was used to compensate for the greater number of words in the training set. Also, as no specific comparison between original vs. retrained learning is needed, weight changes were subject to momentum (0.9) to speed the convergence of learning. Training was halted when the network succeeded in activating the semantic units to within 0.1 of their correct values over the last three iterations.

Lesioning and retraining procedures. After training, the network was lesioned by removing some randomly-selected proportion of the connections between two groups of units. The results reported below are based on lesions of 25% of the intermediate-to-semantics connections. This lesion location was selected because it produced an intermediate amount generalization in the first experiment (27.0%), providing a clear opportunity for the composition of the treated set to have either a positive or a negative impact on generalization.

The retraining procedure was designed to test the amount of generalization in retraining as a function of the typicality of both the treated and untreated sets. After each lesion, performance on all 100 words was measured. A presented word was considered correct if the semantics generated by the network had a higher proximity (normalized dot product) to the correct semantics for the word than to the semantics for any other word. Based on this initial performance, the typical and atypical word sets were each divided in half, balancing for correct performance. The lesioned network was then retrained for 50 epochs, either on half of the typical words or on half of the atypical words (25 words). During retraining, improvement in correct performance was measured on this treated set as well as on two untreated sets: the remaining words of the same type (typical or atypical) and all of the words of the other type. Each half of each group in turn served as the treated set for retraining (reinitializing the weight each time). In this way, the retraining procedure was able to measure the generalization to both typical and atypical words when retraining on either typical or atypical words.

Results and Discussion

Seventy instances of 25% lesions to the intermediate-to-semantics connections reduces overall performance to 35.6% correct on average. Interestingly, lesions impair performance on typical words more than on atypical words (23.8% vs. 47.5% correct, respectively; $F_{1,69} = 355.7$, $p < .001$). This is largely due to the use of a best-match procedure to measure correct performance—typical words have more close competitors than atypical words. The network has to be more accurate in generating the semantics of typical words to distinguish them from other typical words, whereas atypical words are easier to distinguish. On the other hand, the network relearns typical words better than atypical words. Retraining on typical words for 50 epochs improves their performance 65.7% (in absolute terms), whereas performance on atypical words improve only 51.6% ($F_{1,69} = 105.9$, $p < .001$).

The more important issue is the nature of recovery among untreated words. Overall, typical words account for most of the improvement among untreated words (15.4% vs. 1.4%; $F_{1,69} = 113.7$, $p < .001$). These numbers are misleading, however, as there is a strong interaction with the type of the treated set ($F_{1,69} = 104.6$, $p < .001$). Improvement among untreated typical words is substantial regardless of whether the treated set is typical (12.5%) or atypical (18.4%). Untreated atypical words, however, improve only when the treated set is also atypical (10.0%); performance on these words actually decreases when retraining on typical words (−7.2%). This interaction is also clear in

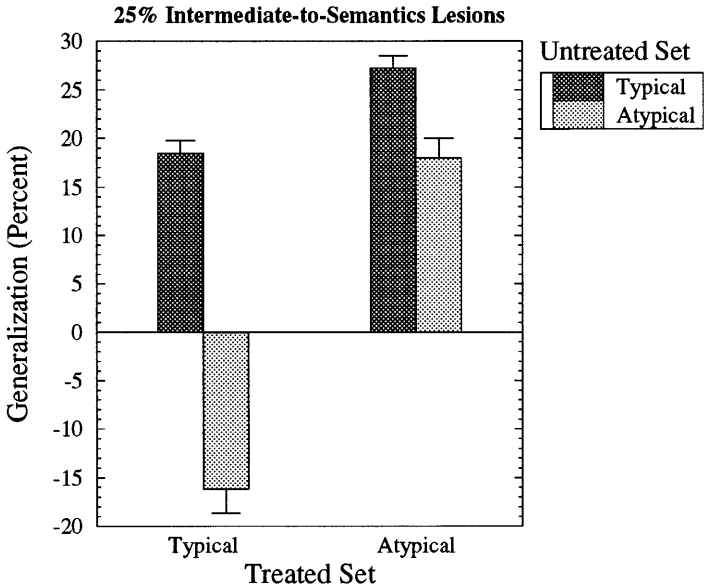

FIG. 14. Generalization from retraining after lesions of 25% of the intermediate-to-semantics connections, as a function of the semantic typicality of the treated and untreated sets.

the average rates of generalization (i.e., ratio of untreated to treated improvement in correct performance) as a function of the semantic typicality of the treated and untreated sets (see Fig. 14). Thus, over all types of untreated words, retraining on atypical words produces greater generalization than does retraining on typical words (22.6% vs. 1.1%, $F_{1,69} = 337.4$, $p < .001$).

Why does retraining on atypical words produce generalization to typical words, but not *vice versa*? Earlier it was argued that generalization depends on the extent to which the treated words provide a good estimate of the relevant (semantic) structure for the entire words set. Typical words accurately estimate the central tendency of a category, but provide little information about the ways in which category members can vary. By contrast, each atypical word indicates a number of ways (i.e., semantic features) in which members can differ from the prototype and yet still belong to the category. Collectively, the semantic representations of atypical words cover more of the features needed by the entire set of words than do the representations of typical words. Thus, retraining on atypical words provides much clearer information about variation within the category. At the same time, the average affects of retraining on atypical words provides a reasonable estimate of the central tendency of the category, and thus, produces generalization to typical words (see, e.g., Posner & Keele, 1968, for similar findings in human category learning). By analogy, among a random cluster of points, the average of the outliers may be quite near the central points, but the average of

the central points is still quite far from the outliers (see Fig. 13). This effect is diminished as the dimensionality of the representations is increased. It is enhanced, however, to the extent that the distribution of atypical words is evenly distributed in all directions from the prototype. Individual words may be quite different from the prototype along some dimensions, but as long as there are other words that collectively are equally different in the opposite direction, the average for the atypical words will still be close to the typical words. Thus, retraining on atypical words can produce generalization among both typical and atypical words.

Summary

A network was trained on an artificial version of the task of reading via meaning, in which words varied in the typicality of their semantic representations. After damage, retraining on words with atypical semantics produced greater overall generalization than did retraining on more typical words. More specifically, retraining on atypical words improves performance on all types of untreated words, whereas retraining on typical words generalized only to other typical words; performance on atypical words deteriorated in this condition. This finding arises because atypical words provide a better estimate of both the central tendency and the variation within the category along each semantic dimension, whereas typical words provide information only about the central tendency. In order for the effect to hold, however, the argument implies that the atypical words cannot be *too* atypical—they must collectively cover the range of variation in the category while still being balanced around its central tendency. The implications of these results for the design of patient rehabilitation are elaborated under General Discussion.

EXPERIMENT 3: CHANGES IN ERROR PATTERN DURING RELEARNING

The first two experiments demonstrate that retraining a damaged connectionist network that maps orthography to semantics can give rise to similar improvements in correct performance on treated and untreated words as have been found in rehabilitation studies with acquired dyslexic patients. Correct performance, however, is a rather coarse characterization of reading behavior. Much of the progress in the study of acquired dyslexia within cognitive neuropsychology has come from detailed analysis of the patterns of errors produced by patients—that is, how various psycholinguistic variables, such as word frequency, orthographic neighborhood, spelling-sound consistency, part-of-speech, and concreteness or imageability, affect the likelihood and nature of incorrect responses. For example, some of the strongest empirical support for the separation of semantic and phonological routes in word reading comes from Marshall and Newcomb's (1966, 1973) original distinction between the error patterns in surface vs. deep dyslexia. Furthermore, many

of the claims for the usefulness of connectionist modeling in neuropsychology have been based on the ability of these networks to provide natural accounts for otherwise counterintuitive combinations of symptoms, including diverse error patterns (Hinton & Shallice, 1991; Plaut & Shallice, 1993a). Thus, error patterns have played a central role in the development of neuropsychological theories and in evaluating connectionist models of impaired cognitive processes. For this reason, it is important to determine the extent that the current connectionist approach to rehabilitation of acquired dyslexia can account for the changes in the patterns of errors produced by patients over the course of recovery.

The current experiment investigates the effects of retraining on the pattern of errors produced after damage to the network developed in Experiment 1. It might seem most natural to compare these effects with the nature of the errors produced by the patients in the rehabilitation studies reviewed earlier—primarily surface dyslexic and dysgraphic patients. A number of considerations, however, make such comparisons inappropriate. Recall that the orthography-to-semantics network constitutes only part of the full framework for lexical processing, as shown in Fig. 1. That framework has both a semantic and a phonological route for pronouncing written words. The surface dyslexic patients in the rehabilitation studies (e.g., Coltheart & Byng, 1989) are assumed to have damage primarily to the semantic route—specifically, in accessing semantics from orthography. The remaining parts of the system are thought to be relatively intact. Given this, and the fact that the therapy was designed specifically to reestablish the orthography-to-semantics pathway, the nature of recovery in these patients is assumed to reflect the properties of relearning in this pathway alone. Thus, it makes sense to model this recovery process by retraining a damaged network that only maps orthography to semantics.

Notice, however, that this argument does not apply when attempting to account for the nature of the patients' error responses. This is because, unlike the network, the patients also have a phonological route available to them for pronouncing words. In fact, most theories ascribe the central characteristics of surface dyslexia—regularization of low-frequency exception words—to the operation of the phonological route, either in normal operation when isolated from semantics (Coltheart et al., 1993; Plaut, Behrmann, Patterson & McClelland, 1993; Plaut et al., in press) or after partial damage (Patterson et al., 1990; Shallice & McCarthy, 1985). For this reason, it would be inappropriate to use an implementation of an isolated semantic route to model the pattern of errors in surface dyslexia and its change over the course of rehabilitation.

Rather, it makes more sense to use relearning in the network to model changes in the error pattern of patients who read solely (or at least primarily) by the semantic route: deep and phonological dyslexic patients. Deep dyslexic patients make a wide range of types of errors in oral reading (see

Coltheart et al., 1980): semantic (e.g., CAT→''dog''), visual (e.g., CAT→ ''mat''), mixed visual-and-semantic (e.g., CAT→''rat''), visual-then-semantic (e.g., CAT→''floor'' via *mat*), morphological (e.g., CAT→''cats''), as well as unrelated errors (e.g., CAT→''mug''). They make virtually no nonword responses to words (literal paraphasias), although they may frequently fail to respond at all (omissions). Note that morphological errors can be considered special cases of mixed visual-and-semantic errors (Funnell, 1987). Phonological dyslexic patients, on the other hand, do not make semantic errors, although they can be quite similar to deep dyslexic patients in other respects. In fact, Glosser and Friedman (1990; Friedman, 1996, also see Newcombe & Marshall, 1980) argue that deep and phonological dyslexic patients fall on a continuum of severity of impairment, with deep dyslexia at the most severe end.

Most rehabilitation studies with deep/phonological dyslexic patients, understandably, have focused on reestablishing the severely impaired phonological route (e.g., Berndt, 1991; Berndt & Mitchum, 1994; de Partz, 1986; Hillis, 1990). As stated above, such changes fall outside the purview of the current implemented network. Thus, we will confine ourselves to a consideration of changes in the patterns of errors that occur without specific therapeutic intervention.

Two specific phenomena are addressed. The first concerns the nature of the onset of deep dyslexia. There is some evidence that the characteristics of deep dyslexia, including the occurrence of semantic errors, may emerge only after some initial period of recovery. Earlier on, the patient is likely to be globally aphasic, making very few overt responses. This seems to have been true of GR (Marshall & Newcombe, 1966) and KF (Shallice & Warrington, 1980), although only tentative conclusions are warranted because, as is generally true, there is only very limited clinical information available on the patients' initial post-morbid behavior.

The second phenomenon to be addressed is the observation that deep dyslexia can eventually resolve into phonological dyslexia. Friedman (1996, also see Klein, Behrmann, & Doctor, 1994) has argued that the symptoms in deep dyslexia resolve in a particular order over the course of recovery. The occurrence of semantic errors is the first symptom to resolve, followed by the concreteness effect, then the part-of-speech effect, then the visual and morphological errors, and only lastly, the impaired nonword reading. A similar pattern of recovery has been documented in deep dysphasic patients, who make semantic errors in repetition (Martin, Dell, & Schwartz, 1994; Martin, Saffran, & Dell, 1996).

The following experiment tests whether relearning after lesions to a network that maps orthography to semantics can account for changes in the error pattern of deep dyslexia at its onset and at its resolution into phonological dyslexia. The form of the network embodies an implicit assumption that the recovery in the patients under these two conditions does not involve

improvements in the phonological route. If the network is unable to account for the data, it may be that this assumption is incorrect and that the phonological route makes a significant contribution to the changing error patterns of the patients.

Method

The current experiment uses the same trained network, lesioning procedure, and retraining procedure as the first experiment. As reported there, the lesioned network produces a response if the proximity between the generated semantics and the defined semantics for some word is at least 0.8, and the next-best match is at least 0.05 worse. If these criteria are not satisfied by the semantics of any word, the network is considered to have failed to respond—it produces an omission. Otherwise, the best-matching word is taken as the response. If this word matches the one presented as input, the network responds correctly; otherwise it produces an explicit error. Notice that the network can be incorrect either by producing an incorrect response or by failing to respond at all.

Error responses were categorized in terms of their visual and semantic similarity with the stimulus word. Following Hinton and Shallice (1991), a response was considered to be visually similar to a stimulus if they overlap in at least one letter, and semantically similar if they come from the same semantic category. Typically, in neuropsychological testing, words must share 50% of the letters ''in some semblance of correct order'' (Morton & Patterson, 1980, p. 103) to be considered visually similar. A somewhat more lenient criterion was adopted for the network because most of the words it was trained on contain only three letters. As a consequence, the chance rate of visual similarity across all possible stimulus-response pairs is exaggerated for the network compared with the patients. To compensate for this, only the relative rates of different error types will be considered. With regard to semantic similarity, the semantic representations of words from the same category have, on average, higher proximity than words from different categories (Plaut and Shallice, 1993a).

Given these definitions for visual and semantic similarity, all possible errors can be categorized into one of four types: visual (but not semantic), semantic (but not visual), mixed visual-and-semantic, and unrelated (neither visual nor semantic). The chance rates of these error types, across all pairings of stimuli and responses from the word set, are 29.9% visual errors, 11.8% semantic errors, 6.2% mixed errors, and 52.2% unrelated errors.

Results and Discussion

Figure 15 shows the distributions of error rates produced by the network after 20 instances of 30% orthography-to-intermediate lesions, 30% intermediate-to-semantics lesions, and 50% cleanup-to-semantics lesions, before any retraining. Also shown for comparison purposes is the chance distribution of errors. Only the relative rates of this distribution are informative. As it shows, over half of the network's error responses would be unrelated if its responses were generated randomly. After each type of lesion, however, the rates of visual errors and semantic errors, relative to the rates of related errors, are greater than predicted by the chance distribution. Thus, visual and semantic similarity are having a significant effect on the errors produced by the network. Furthermore, the rates of mixed visual-and-semantic errors after each type of lesion are greater than would be expected if visual errors and semantic errors were caused by two independent processes (Shallice &

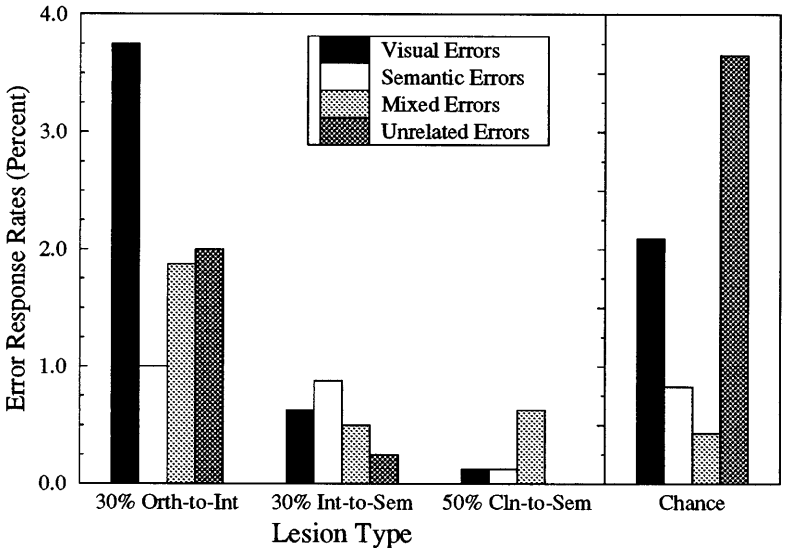

FIG. 15. Rates of visual, semantic, mixed visual-and-semantic, and unrelated errors, after 30% orthography-to-intermediate lesions, 30% intermediate-to-semantics lesions, and 50% cleanup-to-semantics lesions. Results are averaged over 20 instances of each lesion type. These absolute error rates must be judged relative to the ''Chance'' distribution of error rates that would arise if responses were chosen randomly from the word set. The absolute height of the Chance distribution is set arbitrarily—only the relative rates are informative.

McGill, 1978). Thus, the network shows a particular bias toward producing mixed visual-and-semantic errors when damaged. These findings replicate those of Hinton and Shalice (1991) and correspond to the basic error pattern of deep dyslexia.

It should be pointed that the absolute error rates produced by the network are significantly lower than those of most deep dyslexic patients. This discrepancy may be due in part to inadequacies of the use of response criteria as a substitute for an actual output network, and in part to limitations in the scale of the simulation—specifically, the limited number of visual and semantic competitors for each word (see Plaut & Shallice, 1993a, for discussion). In spite of this limitation in the quantitative adequacy of the network, it is still useful to study the qualitative changes in the distribution of error types when it is retrained after damage.

Onset of deep dyslexia. The first issue to be addressed is the gradual onset of the deep dyslexic error pattern from an initial state in which very few overt responses are produced. To model this situation, the network was retrained after a very severe type of lesion: 70% of intermediate-to-semantics connections. This lesion condition was selected because it produces the greatest impairment of all those studied, reducing correct performance to

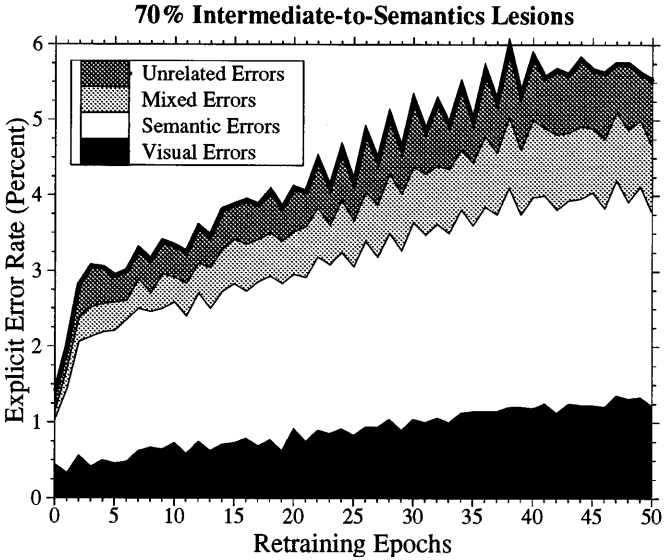

FIG. 16. The distribution of error types over all words during relearning on 20 words after lesions of 70% of the intermediate-to-semantics connections. The dark line indicates the overall error rate which is broken down into visual, semantic, mixed visual-and-semantic, and unrelated errors as shown below it.

1.2% on average (see Fig. 7). In addition, the network produces overt errors to only 1.4% of word presentations; a full 98.5% of presentations yield no response.

To approximate the nature of recovery without specific intervention, the network was retrained on only half of the 40 words, although improvement in performance and changes in error pattern were calculated across all of the words. Furthermore, given the relatively low absolute error rates produced after lesions, results are averaged over a total of 60 instances of lesions as well as exchanges of the treated and untreated words.

After 50 epochs of retraining on the treated words, overall performance improved from 1.2 to 18.2% correct on average (paired $t_{119} = 27.5$, $p < .001$). Figure 16 shows the distribution of error types over the course of this retraining. In the figure, the error types are stacked on top of each other; the top line indicates the total error rate which is broken down into the four types of errors as shown below it. The most important effect is that the overall explicit error rate *increases* as a result of retraining: from 1.4 to 5.5% of word presentations at epoch 50 (paired $t_{119} = 10.9$, $p < .001$) and as high as 6.0% at epoch 38. Thus, one consequence of relearning is to make the network more likely to produce any type of overt response, either correct or incorrect, rather than producing an omission.

Among error responses, the greatest proportional increases are among er-

rors with a semantic component. Mixed visual-and-semantic errors increase to 8.4 times their initial rate (from 0.104 to 0.875%; paired $t_{119} = 5.13$, $p < .001$); the rate of semantic errors increases 4.1 times (from 0.625% to 2.56%; paired $t_{119} = 7.91$, $p < .001$). By contrast, visual errors increase to only 2.8 times their initial rates (from 0.438 to 1.23%; paired $t_{119} = 4.13$, $p < .001$) and unrelated errors increase only 3.5 times (from 0.250 to 0.875%; paired $t_{119} = 3.91$, $p < .001$; increase in semantic vs. unrelated errors; paired $t_{119} = 4.48$, $p < .001$). Thus, the error pattern that emerges after retraining corresponds to the standard pattern found in deep dyslexia: a substantial number of semantic and mixed visual-and-semantic errors that co-occur with visual errors.

To test the generality of the results, the same procedure was applied to the network after 70% orthography-to-intermediate lesions. As in Experiment 1, the network has more difficulty recovering after these lesions than after 70% intermediate-to-semantics lesions. Fifty epochs of retraining on the treated words improved performance from 1.38% to only 10.5% correct on all 40 words (paired $t_{119} = 19.0$, $p < .001$). In contrast to the above results, the concurrent increase in explicit error rates is not reliable (from 9.21 to 10.6%; paired $t_{119} = 1.58$, $p = .116$). However, if error types are considered separately, there are reliable increases in the rates of semantic errors (from 1.63 to 2.17%; paired $t_{119} = 2.01$, $p = .047$) and mixed visual-and-semantic errors (from 0.83 to 1.56%; paired $t_{119} = 3.20$, $p = .002$). By contrast, visual error rates do not increase reliably (from 3.50 to 4.27%; paired $t_{119} = 1.62$, $p = .107$) and unrelated errors show a trend toward decreasing (from 3.25 to 2.58%; paired $t_{119} = 1.67$, $p = .097$). Thus, retraining after 70% orthography-to-intermediate lesions produces the same general effects on the pattern of errors as retraining after intermediate-to-semantics lesions of equivalent severity, except that the overall error rate does not increase.

In summary, retraining after severe intermediate-to-semantics lesions changes the error pattern of the network from one in which virtually all word presentations produce omissions, to one in which correct performance has improved considerably and the rates of explicit errors have increased, particularly among those semantically related to the stimulus. Broadly similar results obtain after orthography-to-semantic lesions. This relearning after severe lesions approximates the onset of deep dyslexia from an initial global aphasia.

Resolution of deep dyslexia into phonological dyslexia. The second issue investigated is the observation that deep dyslexia occasionally resolves into phonological dyslexia. For the present purposes, this transition can be operationalized in terms of the finding that, over the course of recovery, semantic errors are the first type of error to drop out. The resulting pattern of behavior, involving visual, mixed, and unrelated errors along with a reasonable level of correct performance, approximates phonological dyslexia in the current context.

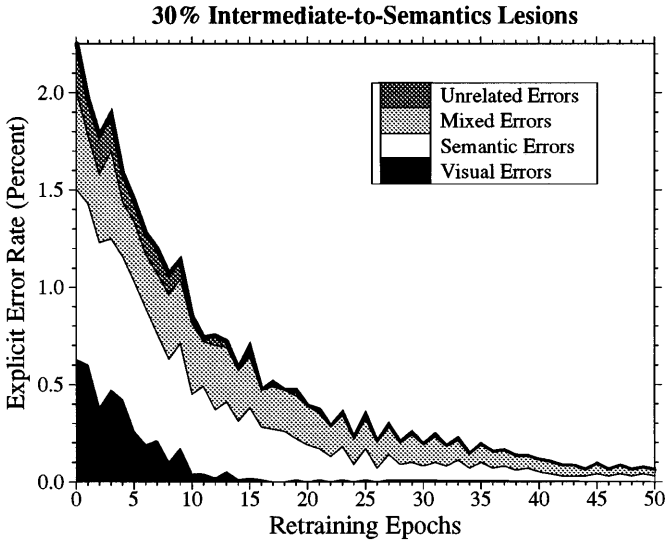

FIG. 17. The distribution of error types over all words during relearning on 20 words after lesions of 30% of the intermediate-to-semantics connections. The dark line indicates the overall error rate which is broken down into visual, semantic, mixed visual-and-semantic, and unrelated errors as shown below it.

To investigate whether the network exhibits a similar transition in recovery, it was retrained after 30% lesions to either the orthography-to-intermediate or intermediate-to-semantics connections. These locations and severities were chosen because they were studied in detail in Experiment 1 and they give rise to substantial rates of explicit errors (see Fig. 15). To ensure that overall correct performance improves to a reasonable level, all 40 words were presented to the network during retraining. Again, results were averaged over 60 instances of lesions to provide reasonable estimates of the rates of each type of error.

Considering 30% intermediate-to-semantics lesions first, over the course of 50 epochs of retraining, performance on the 40 words improves from 24.5 to 92.7% correct (paired $t_{59} = 54.0$, $p < .001$).[5] Concurrently, explicit error responses are virtually eliminated, dropping from 2.25 to 0.06% of all word presentations (paired $t_{59} = 7.82$, $p < .001$). Figure 17 shows the distribution of error types over the course of this relearning. Rather than semantic errors being the first to drop out, visual and unrelated errors are eliminated earliest, between epochs 15 and 30. Semantic and mixed visual-and-semantic errors are eliminated only at the very end of retraining. Thus, the changes in the pattern of errors produced by the network in recovery to near normal levels

[5] This numbers are slightly different than those found in Experiment 1 (see Fig. 11) because different specific lesion instances were administered.

of correct performance fails to replicate the observed transition from deep to phonological dyslexia in patients.

Similar results obtained when retraining after 30% orthography-to-intermediate lesions. Overall correct performance improves from 19.1 to 75.3% (paired $t_{59} = 51.0$, $p < .001$) while explicit error rates drop from 8.42 to 1.33% (paired $t_{59} = 13.0$, $p < .001$). Over the course of this recovery, the rate of visual errors drops much faster than that of semantic errors. For instance, after 20 epochs of retraining, the visual error rate has dropped 73.8% of its original value on average (from 3.67 to 0.96%; paired $t_{59} = 7.67$, $p < .001$). By contrast, the semantic error rate has dropped only 38.3% (from 1.96 to 1.21%; paired $t_{59} = 2.33$, $p = .023$; visual vs. semantic: paired $t_{59} = 4.96$, $p < .001$). Thus, it is not the case that semantic errors are eliminated before visual errors during retraining.

This discrepancy between the network and the patients can be understood if recovery in the patients involves more than relearning in the semantic route alone. In particular, the findings suggest that, within the current approach, the transition from deep dyslexia to phonological dyslexia must also involve some improvement in the operation of the phonological route (or in phonology itself). Such improvement would lead naturally to a greater reduction in semantic errors relative to other types of error because even partial correct phonological information about the stimulus would be sufficient to rule out most semantic errors (Newcombe & Marshall, 1980). For example, any information about the pronunciation of CAT is sufficient to disqualify ''dog'' as an appropriate response. By contrast, visual and mixed visual-and-semantic errors would be more difficult to detect on the basis of partial operation of the phonological route, as words that are visually similar also tend to be phonologically similar.

One clear indicator of the operation of the phonological route is the ability to read pronounceable nonwords, as such items cannot be read via semantics. Thus, the above explanation is supported by the observation that, for many deep/phonological patients, as their rates of semantic errors dropped to near zero, their nonword reading performance improved. For example, on initial testing, patient GR (Glosser & Friedman, 1990) made 11% semantic errors and read correctly 5% (1/20) of nonwords. Seven months later, he made no purely semantic errors (although 3% were visual-and-semantic); concurrently, his nonword reading had improved to 44% (22/50). Similar results have been found with a number of other patients (e.g., DV, Glosser & Friedman, 1990; EG, Laine, Niemi, & Marttila, 1990; see Friedman, 1996, for further discussion). Consideration of a possible exception (RL, Klein et al., 1994) will be postponed to the General Discussion.

Summary

The current experiment tests whether a network that maps orthography to semantics can account for changes in the pattern of errors made by patients

over the course of recovery. Two specific phenomena are considered. The first is that the deep dyslexic error pattern often emerges only after some recovery from an initial state in which very few overt response are made. In the network, retraining after severe lesions produces a similar effect: rates of correct and error responses increase relative to omissions, with the resulting error pattern exhibiting the main characteristics of deep dyslexia.

The second phenomenon addressed in the current experiment is the finding that deep dyslexia occasionally resolves into phonological dyslexia, as patients gradually stop making semantic errors. However, retraining the network fails to produce the change in error pattern from deep to phonological dyslexia. Specifically, in the network, visual errors drop out before semantic errors, which is the opposite of what is found in the patients. The limitations of the network on its own lead to the interpretation that recovery in these patients involves relearning in the phonological route as well as in the semantic route.

GENERAL DISCUSSION

Theoretical analyses of cognitive impairments following brain damage should lead to the design of more effective strategies for rehabilitation. Within cognitive neuropsychology, such impairments have traditionally been described in terms of lesions to one or more components within a ''box-and-arrow'' information-processing diagram of the cognitive system. However, as a number of authors have pointed out (e.g., Basso, 1989; Caramazza, 1989; Hillis, 1993; Wilson & Patterson, 1990), identifying which components are damaged in a particular patient is only the *start* of a specification of how best to remediate the patient's cognitive abilities. As Hillis (1993, p. 24) puts it,

> A theory concerning the normal cognitive processes underlying language tasks, and an hypothesis about each patient's level of breakdown, *are* necessary, but not sufficient, for specification of a rational treatment programme. What is needed in addition is a theory of intervention—how the damaged system is modified in response to a specific intervention. . . . The essential components of a theory of intervention would include: (1) detailed analysis of both the pre-therapy and the post-therapy damaged states of the cognitive system in each treated patient and the relationship between these two states; (2) articulation of how the change from one state to the other occurred . . . and (3) determination of the characteristics of the patient and the patient's brain damage that are relevant to treatment outcomes.

Unfortunately, traditional theorizing within cognitive neuropsychology seems to provide little guidance in these matters. Relatively few attempts have been made to specify the actual representations and processes that underlie each component's operation (Seidenberg, 1988). Furthermore, the box-and-arrow framework provides only a very coarse characterization of the effects of brain damage: a component may be spared or eliminated, but the possibility of partial damage is ill defined (Allport, 1985; Wilson & Pat-

terson, 1990). Finally, little attention is paid typically to how the system learns, either normally or after damage.

Connectionist modeling offers specific hypotheses about the nature of the representations and computations that underlie cognitive processes, as well as how these processes are learned through experience and how they are affected by brain damage. The current work attempts to extend the relevance of connectionist modeling in neuropsychology one step further, to contribute to a theory of rehabilitation—as outlined by Hillis above—based on analyses of relearning in damaged networks. To this end, three simulation experiments were carried out in the domain of reading via meaning. The first two address a central issue in rehabilitation studies: What factors influence the degree of recovery and generalization observed in a patient? The factors investigated were the specific location of lesion within the system, and the nature of the set of items presented during treatment. The third experiment attempted to provide further constraints on the nature of recovery by comparing the changes in the patterns of errors produced by patients and by the network when relearning.

Relearning, Generalization, and Task Structure

Attempts at cognitive rehabilitation of the mapping between orthography and semantics (e.g., Behrmann, 1987; Coltheart & Byng, 1989; Scott & Byng, 1989) have resulted in considerable improvement in performance on treated words, as well as significant generalization to untreated but related words. Why should this occur? In general, there is little understanding of the underlying mechanisms by which cognitive functions recover, either spontaneously or as a direct result of therapeutic intervention. Generalization in the domain of reading via meaning is particularly puzzling as there is no systematic relationship between the written or spoken forms of words and their meanings. Unfortunately, the degree and breadth of recovery and generalization can vary considerably across patients. Some patients show generalization in some categories but not others (e.g., CH; Behrmann & Lieberthal, 1989); some learn the treated items well but show no generalization to untreated items (e.g., PS; Hillis, 1993); still others have difficulty learning the treated items themselves.

The findings in Experiment 1, together with previous connectionist demonstrations (Hinton & Plaut, 1987; Hinton & Sejnowski, 1986), provide an explanation. In a network, the orthography and semantics of each word are represented as distributed patterns of activity such that words that are similar in each domain are assigned similar (overlapping) patterns. Consequently, the system's knowledge of the relationship between orthography and semantics for every word is encoded in the same set of weights. Although there is no systematic structure in the task as a whole, the network learns weights that capture whatever local regularities happen to exist among the words.

When the network is retrained after damage, the weight changes induced by treated words tend to reestablish all of the relevant regularities among words. As a result, retraining also tends to improve performance on untreated words. In this way, the principles that give rise to the effects of relearning in connectionist networks may provide insight into the nature of recovery and generalization in patients.

The explanation suggests further that the degree of relearning and generalization should depend on factors which influence how well retraining captures the relevant structure of the task. Experiment 1 showed that the degree of relearning and generalization after damage in the network depends considerably on the location of damage: lesions at the level of semantics gave rise to rapid relearning and considerable generalization; lesions near orthography produced much poorer relearning and no generalization; lesions at an intermediate location produced intermediate generalization. These findings correspond directly to the degree of structure in the subtasks performed by each part of the network. The initial connections from orthography map representations that are orthographically organized to ones that are partially semantically organized. Given that orthographic similarity is unrelated to semantic similarity, there is no structure in this subtask. However, to the extent that the resulting representations are semantically organized, generating semantics from them has some structure. Interactions within the clean-up system have the greatest degree of structure, however, as representations at this level are highly semantically organized.

The network results may help explain the variability in recovery observed in patients. Specifically, patients who show considerable generalization would be predicted to have lesions near or within semantics, whereas patients who show no generalization should have more peripheral impairments. This prediction has recently been challenged by Weekes and Coltheart (in press) on the basis of a rehabilitation study with a surface dyslexic patient, NW. They argued that NW has an orthographic rather than semantic impairment, and yet, using a therapy program similar to that of Coltheart and Byng (1989), they demonstrated significant generalization in his reading (42.2%, averaging across all words and across all pre- and post-treatment tests). The evidence for the orthographic impairment was that NW was unable to distinguish word/pseudohomophone pairs (e.g., TURTLE/TERTLE) in visual lexical decision. However, lexical decision would seem to be a particularly poor choice of task to identify an orthographic deficit, as the critical distinction between words and nonwords is that words have semantics whereas nonwords do not. A more informative approach might have been to demonstrate an impairment in cross-case matching or orthographic priming. The evidence for a lack of semantic impairment was that NW was better at defining spoken than written words. However, his auditory comprehension may not have been fully intact: while NW was able to define correctly 37/40 spoken words from the PALPA (Psycholinguistic Assessments of Language Processing in

Aphasia), he was correct on only 17/50 spoken words from the NART (National Adult Reading Test; although some of the words on this test may have been unknown to NW premorbidly, and Weeks and Coltheart give no control data). Perhaps this patient's behavior can best be explained by a lesion intermediate between orthography and semantics—in the network, such lesions yield moderate levels of generalization (27.0%). The quantitative difference between this and the patient's generalization is difficult to interpret given that different word sets were used and that the scale of the implementation is vastly smaller than the actual word-reading system.

A final issue concerns the nature of the changes within the system that underlie recovery. A common assumption in cognitive neuropsychology is that the effects of damage, and the resulting recovery, is local to the location of damage. In the extreme form, only a single processing component is affected while all others remain unchanged. Results from the first experiment showed that relearning involves weight changes throughout the damaged network, even for connections far removed from the lesion. Thus, recovery in the network violates the locality assumption and calls into question its appropriateness for patients (cf. Farah, 1994). Certain locations within the network tend to change their weights more than others, regardless of lesion location. Nonetheless, the location of damage does significantly affect the distribution of weight changes during relearning: the weights on connections that remain at the lesioned location tend to change more than when the lesion is elsewhere. Thus, recovery in the network can be viewed as a combination of ''restoration'' of the damaged part of the system and ''compensation'' by the remaining parts.

Designing Treatment to Maximize Generalization

Given that, at least under some conditions, treatment can induce improvement in untreated items, one would like to know how to design treatment to maximize this generalization. A potential benefit of connectionist modeling in neuropsychological rehabilitation is that it provides a framework for investigating the relative effectiveness of alternative rehabilitation strategies. Unfortunately, the range of strategies that are currently available for retraining a network is far more limited than that available to a therapist. In particular, the only clear option in retraining that corresponds to a choice available to therapists is the selection of items for treatment. Nevertheless, with respect to this specific issue, principles of connectionist relearning may provide useful insights.

The findings from the first experiment suggest that generalization depends on the extent to which retraining captures the relevant structure of the task. In other words, retraining will give rise to generalization if the effects of retraining on the treated words provide a good estimate of the task structure that is relevant to the entire set of words. In reading via meaning, much

of the relevant structure is semantic, and an important aspect of semantic representations is typicality—how similar the semantic representation of a word is to the central tendency of its category. Accordingly, Experiment 2 tested the effect of semantic typicality of treated words on the degree of generalization to untreated words.

A second network, analogous to the one used in the first experiment, was trained on a more artificial version of the task of mapping orthography to semantics, to allow greater control of the semantic relationships among words. In retraining after damage, retraining on words typical of their semantic category yielded generalization only to untreated words that were also typical; performance on untreated atypical words deteriorated. By contrast, retraining on atypical words produced generalization both to typical and atypical untreated words. Thus, overall, treated items that are somewhat atypical of their semantic category gave rise to greater generalization than more typical treated items.

These findings make sense given the adequacy with which sets of typical vs. atypical words can approximate the semantic structure of the entire word set. The average effect of retraining on typical words provides a good estimate of the changes needed by other typical words; they are, however, quite different from the changes needed by atypical words. The atypical words, by contrast, provide better information about semantic dimensions on which category members vary, which supports generalization to other atypical words. In addition, the atypical words collectively provide a good approximation of the central tendency of the category, so that untreated typical words also improve. Thus, the findings from Experiment 2 also support the view that generalization depends on the extent to which retraining approximates the overall structure of the task.

Caution is warranted, however, in considering the implications of these findings for the design of patient therapy. First, according to the explanation offered, the atypical words cannot be *highly* atypical. In particular, the treated set should cover the full range of variation within the category in a balanced way along the relevant semantic dimensions, so that its average effects are close to the central tendency of the category. Furthermore, the simulation involved typical and atypical words from only a single semantic category. Further work is required to verify that similar results obtain when retraining on atypical exemplars from multiple categories simultaneously. Nonetheless, the findings suggest that connectionist modeling can provide useful insights into the relative efficacy of alternative treatment strategies.

Changes in Error Pattern during Recovery

In relating the behavior of brain-damaged patients to models of normal cognitive processing, cognitive neuropsychology relies heavily, not only on relative levels of correct performance, but also on different patterns of errors

produced by the patients. Experiment 3 investigated the extent to which the changes in the error pattern produced by a network over the course of retraining corresponds to the analogous error patterns of recovering patients.

Two specific phenomena were addressed, relating to the onset of deep dyslexia, and to its occasional resolution into phonological dyslexia. The first is that patients do not typically exhibit the symptoms of deep dyslexia immediately following the neurological insult, but after some amount of recovery from a state of global aphasia, in which very few overt responses are produced. Retraining the network from the first experiment after a severe lesion produces an analogous effect. Initially after the lesion, virtually all word presentation produce omissions. After some amount of retraining, the rates of both correct and error responses increase, ultimately giving rise to the deep dyslexic error pattern.

The second phenomenon that was investigated is the observation that, over the course of recovery, deep dyslexia may resolve into phonological dyslexia (Friedman, 1996; Glosser & Friedman, 1990). As the main distinction between these two types of patients is that only deep dyslexic patients make semantic errors, the relevant pattern is that this is the first type of error to drop out in recovery. By contrast, if the network is retrained after a moderate lesion, visual and unrelated errors are eliminated first, and semantic errors drop out only near the end of retraining, when performance is near perfect. Thus, the behavior of the network fails to correspond to the observed pattern of behavior of the analogous patients. A possible explanation for this finding is that the underlying recovery that gives rise to the elimination of semantic errors does not occur within the semantic route (as implemented by the network) but, rather, within the phonological route. Consistent with this interpretation, most patients who cease making semantic errors simultaneously improve in their ability to read pronounceable nonwords, which must involve the phonological route (see Friedman, 1996, for details).

A possible exception to this pattern has recently been described by Klein et al. (1994). On initial testing, their patient, RL, produced 33% (43/131) semantic errors but could read only 3% (1/32) nonwords. On a second testing, RL made only 10 explicit errors, none of which were purely semantically related (one was mixed visual-and-semantic). At that point, he could read 13% (4/32) nonwords. The apparent improvement in nonword reading over the two testing sessions was significant only at the $p = .1$ level. However, it should be kept in mind that nonword reading is a stringent test of the operation of the phonological route. All of the phonemes of a nonword must be produced correctly for it to be counted as correct. By contrast, only partial phonological information, perhaps even a single phoneme, may be sufficient to edit out most semantic errors (Newcombe & Marshall, 1980). In fact, there is indirect evidence that RL had some partial operation of the phonological route at the time of the second testing: he showed a significant advantage in reading pseudohomophones (e.g., BRANE) relative to orthographically-

matched nonwords (e.g., FRANE; 8/20 vs. 2/20 correct, respectively; $p < .05$). A similar explanation would seem to apply to another deep dyslexic patient, LR (Berndt & Mitchum, 1994; Mitchum & Berndt, 1991), who showed no improvement in nonword reading over the course of recovery but had learned the correct sounds of single letter graphemes and could segment initial phonemes from words.

Thus, while the network on its own fails to account for the resolution of deep to phonological dyslexia, its performance is consistent with a more general account, in which the phonological route also contributes to the nature of the recovery in these patients.

Conclusions

Connectionist modeling provides a useful framework for exploring the nature of normal and impaired cognitive processes. The current work uses principles of connectionist learning (and relearning) to contribute to an understanding of the nature of recovery in brain-damaged patients. A general finding that emerges from the simulation experiments is that the relative structure of the tasks performed by different parts of the system, and the extent to which the items selected for treatment approximate this structure, play important roles in determining the degree of recovery and generalization produced by retraining after damage.

It must be kept in mind, however, that the current findings relate to patient therapy only in the most general way. The simulations address only a particular neuropsychological domain: impaired reading via meaning. Furthermore, the version of the task of mapping orthography to semantics is of a vastly smaller scale than the actual task performed by patients. Nonetheless, despite these limitations, the principles that emerge as central to understanding the nature of relearning and generalization in the networks may provide the foundations for understanding the nature of recovery in patients.

Much more could be learned from a detailed attempt to model the pattern of recovery of a particular patient (see Martin, Saffran, & Dell, 1996, for a recent attempt). The current work makes predictions primarily about the influence of semantic variables, such as typicality, on relearning and generalization. Hence, the most appropriate type of patient with which to test these predictions would be one with semantic deficits. Research is currently ongoing to develop a more elaborate version of the semantic system in which the effects of a wider range of semantic variables could be investigated than is possible in the current simulations.

A few years ago, Wilson and Patterson (1990, p. 256) suggested that connectionist or parallel distributed processing (PDP) models might provide a valuable framework in which to explore rehabilitation strategies.

> Clearly, we would not try to argue that all of the many issues in design of rehabilitation programmes are about to be solved by PDP models. We merely suggest that

> much might be learned by simulating, within working computational models (and without any ethical considerations), various forms of damage followed by various regimes of re-learning. . . . Perhaps such models will one day become part of the standard set of tools to be used in rehabilitation research.

Of course, the ultimate test of the adequacy of a connectionist approach to cognitive rehabilitation is the extent to which the hypotheses it generates lead to improved therapy for patients. While such a goal is beyond the scope of the current work, hopefully it has moved us a step closer toward achieving it.

REFERENCES

Ackley, D. H., Hinton, G. E., & Sejnowski, T. J. 1985. A learning algorithm for Boltzmann Machines. *Cognitive Science,* **9**(2), 147–169.

Allport, D. A. 1985. Distributed memory, modular systems and dysphasia. In S. K. Newman & R. Epstein (Eds.), *Current perspectives in dysphasia.* Edinburgh: Churchill Livingstone.

Bapi, R. S., & Levine, D. S. 1990. Networks modeling the involvement of the frontal lobes in learning and performance of flexible movement sequences. In *Proceedings of the 12th Annual Conference of the Cognitive Science Society.* Hillsdale, NJ: Erlbaum. Pp. 915–922.

Basso, A. 1989. Spontaneous recovery and language rehabilitation. In X. Seron & G. Deloche (Eds.), *Cognitive approaches in neuropsychological rehabilitation.* Hillsdale, NJ: Erlbaum. Pp. 17–37.

Beauvois, M.-F., & Derouesné, J. 1979. Phonological alexia: Three dissociations. *Journal of Neurology, Neurosurgery and Psychiatry,* **42,** 1115–1124.

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

Behrmann, M., & Bub, D. 1992. Surface dyslexia and dysgraphia: Dual routes, a single lexicon. *Cognitive Neuropsychology,* **9**(3), 209–258.

Behrmann, M., & Lieberthal, T. 1989. Category-specific treatment of a lexical semantic deficit: A single case study of global aphasia. *British Journal of Communication Disorders,* **24,** 281–299.

Berndt, R. S. 1991. Sentence processing in aphasia. In M. T. Sarno (Ed.), *Acquired aphasia.* San Diego, CA: Academic Press. Pp. 223–269.

Berndt, R. S., & Mitchum, C. C. 1994. Approaches to the rehabilitation of ''phonological assembly'': Elaborating the model of non-lexical reading. In M. J. Riddoch & G. W. Humphreys (Eds.), *Cognitive neuropsychology and cognitive rehabilitation.* Hillsdale, NJ: Erlbaum. Pp. 503–526.

Buchanan, L., & Besner, D. 1993. Reading aloud: Evidence for the use of a whole word nonsemantic pathway. *Canadian Journal of Experimental Psychology,* **47**(2), 133–152.

Burton, M. A., Young, A. W., Bruce, V., Johnston, R. A., & Ellis, A. W. 1991. Understanding covert recognition. *Cognition,* **39**(2), 129–166.

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

Byng, S. & Coltheart, M. 1986. Aphasia therapy research: Methodological requirements and illustrative results. In E. Hjelmquist & L. G. Nilsson (Eds.), *Communication and handicap.* Amsterdam: Elsevier Science Publishers. Pp. 191–213.

Caramazza, A. 1989. Cognitive neuropsychology and rehabilitation: An unfulfilled promise? In X. Seron & G. Deloche (Eds.), *Cognitive approaches in neuropsychological rehabilitation.* Hillsdale, NJ: Erlbaum. Pp. 383–398.

Carr, T. H., & Pollatsek, A. 1985. Recognizing printed words: A look at current models. In

D. Besner, T. G. Waller, & G. E. MacKinnon (Eds.), *Reading research: Advances in theory and practice, vol. 5.* New York, NY: Academic Press.

Chauvin, Y. 1988. *Symbol acquisition in humans and neural (PDP) networks.* PhD thesis, University of California, San Diego.

Cohen, J. D., Romero, R. D., Servan-Schreiber, D., & Farah, M. J. 1994. Disengaging from the disengage function: The relation of macrostructure to microstructure in parietal attentional deficits. *Journal of Cognitive Neuroscience,* **6,** 377–387.

Cohen, J. D., & Servan-Schreiber, D. 1992. Context, cortex, and dopamine: A connectionist approach to behavior and biology in schizophrenia. *Psychological Review,* **99**(1), 45–77.

Coltheart, M. 1978. Lexical access in simple reading tasks. In G. Underwood (Ed.), *Strategies of information processing.* New York: Academic Press.

Coltheart, M. 1985. Cognitive neuropsychology and the study of reading. In M. I. Posner & O. S. M. Marin (Eds.), *Attention and performance XI.* Hillsdale, NJ: Erlbaum. Pp. 3–37.

Coltheart, M., & Byng, S. 1989. A treatment for surface dyslexia. In X. Seron & G. Deloche (Eds.), *Cognitive approaches in neuropsychological rehabilitation.* Hillsdale, NJ: Lawrence Erlbaum Associates. Pp. 159–174.

Coltheart, M., Curtis, B., Atkins, P., & Haller, M. 1993. Models of reading aloud: Dual-route and parallel-distributed-processing approaches. *Psychological Review,* **100**(4), 589–608.

Coltheart, M., & Funnell, E. 1987. Reading writing: One lexicon or two? In D. A. Allport, D. G. MacKay, W. Printz, & E. Scheerer (Eds.), *Language perception and production: Shared mechanisms in listening, speaking, reading and writing.* New York: Academic Press. Pp. 313–339.

Coltheart, M., Patterson, K. E., & Marshall, J. C. (Eds.) 1980. *Deep dyslexia.* London: Routledge & Kegan Paul.

Coltheart, M., Sartori, G., & Job, R. (Eds.) 1987. *The cognitive neuropsychology of language.* Hillsdale, NJ: Erlbaum.

de Partz, M.-P. 1986. Re-education of a deep dyslexic patient: Rationale of the methods and results. *Cognitive Neuropsychology,* **3**(2), 149–177.

Farah, M. J. 1994. Neuropsychological inference with an interactive brain: A critique of the locality assumption. *Behavioral and Brain Sciences,* **17,** 43–104.

Farah, M. J., & McClelland, J. L. 1991. A computational model of semantic memory impairment: Modality-specificity and emergent category-specificity. *Journal of Experimental Psychology: General,* **120**(4), 339–357.

Farah, M. J., O'Reilly, R. C., & Vecera, S. P. 1993. Dissociated overt and covert recognition as an emergent property of a lesioned neural network. *Psychological Review,* **100**(4), 571–588.

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

Friedman, R. B. 1996. Phonological dyslexia is a continuum (with deep dyslexia as its endpoint). *Brain and Language,* **52,** 114–128.

Funnell, E. 1987. Morphological errors in acquired dyslexia: A case of mistaken identity. *Quarterly Journal of Experimental Psychology,* **39A,** 497–539.

Glosser, G., & Friedman, R. B. 1990. The continuum of deep/phonological alexia. *Cortex,* **26,** 343–359.

Gordon, B. 1982. Confrontation naming: Computational model and disconnection simulation. In M. A. Arbib, D. Caplan, & J. C. Marshall (Eds.), *Neural models of language processes.* New York: Academic Press.

Harley, T. A., & MacAndrew, S. B. G. 1992. Modelling paraphasias in normal and aphasic speech. In *Proceedings of the 14th Annual Conference of the Cognitive Science Society.* Hillsdale, NJ: Erlbaum. Pp. 378–383.

Hillis, A. E. 1990. Effects of separate treatments for distinct impairments within the naming process. In T. Prescott (Ed.), *Clinical aphasiology, 1989.* Austin, TX: Pro-Ed. Pp. 255–265.

Hillis, A. E. 1993. The role of models of language processing in rehabilitation of language impairments. *Aphasiology,* **7**(1), 5–26.

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

Hillis, A. E., & Caramazza, A. 1994. Theories of lexical processing and theories of rehabilitation. In M. J. Riddoch & G. W. Humphreys (Eds.), *Cognitive neuropsychology and cognitive rehabilitation.* Hillsdale, NJ: Erlbaum. Pp. 449–484.

Hinton, G. E., McClelland, J. L., & Rumelhart, D. E. 1986. Distributed representations. In D. E. Rumelhart, J. L. McClelland, & the PDP research group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition. Volume 1: Foundations.* Cambridge, MA: MIT Press. Pp. 77–109.

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

Hinton, G. E., & Sejnowski, T. J. 1986. Learning and relearning in Boltzmann Machines. In D. E. Rumelhart, J. L. McClelland, & the PDP research group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition. Volume 1: Foundations.* Cambridge, MA: MIT Press. Pp. 282–317.

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

Horn, D., Ruppin, E., Usher, M., & Herrmann, M. 1993. Neural network modeling of memory deterioration in Alzheimer's Disease. *Neural Computation,* **5**(5), 736–749.

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

Humphreys, G. W., Freeman, T., & Müller, H. J. 1992. Lesioning a connectionist model of visual search: Selective effects on distractor grouping. *Canadian Journal of Psychology,* **46,** 417–460.

Karmiloff-Smith, A. 1992. *Beyond modularity: A developmental perspective on cognitive science.* Cambridge, MA: MIT Press.

Kay, J., & Marcel, A. J. 1981. One process, not two, in reading aloud: Lexical analogies do the work of nonlexical rules. *Quarterly Journal of Experimental Psychology,* **33A,** 397–414.

Klein, D., Behrmann, M., & Doctor, E. 1994. The evolution of deep dyslexia: Evidence for the spontaneous recovery of the semantic reading route. *Cognitive Neuropsychology,* **11,** 571–611.

Kucera, H., & Francis, W. N. 1967. *Computational analysis of present-day American English.* Providence, RI: Brown Univ. Press.

Laine, M., Niemi, J., & Marttila, R. 1990. Changing error patterns during reading recovery: A case study. *Journal of Neurolinguistics,* **5,** 75–81.

Levine, D. S., & Prueitt, P. S. 1989. Modeling some effects of frontal lobe damage—Novelty and perseveration. *Neural Networks,* **2,** 103–116.

Margolin, D. (Ed.) 1992. *Cognitive neuropsychology in clinical practice.* Oxford: Oxford Univ. Press.

Marshall, J. C., & Newcombe, F. 1966. Syntactic and semantic errors in paralexia. *Neuropsychologia,* **4,** 169–176.

Marshall, J. C., & Newcombe, F. 1973. Patterns of paralexia: A psycholinguistic approach. *Journal of Psycholinguistic Research,* **2,** 175–199.

Martin, N., Dell, G. S., & Schwartz, M. F. 1994. Origins of paraphasias in deep dysphasia: Testing the consequences of a decay impairment to an interactive spreading activation model of lexical retrieval. *Brain and Language,* **47,** 609–660.

Martin, N., Saffran, E. M., & Dell, G. S. 1996. Recovery in deep dysphasia: Evidence for a relation between auditory-verbal-STM capacity and lexical errors in repetition. *Brain and Language,* **52,** 83–113.

McClelland, J. L. 1987. The case for interactionism in language processing. In M. Coltheart (Ed.), *Attention and performance XII: The psychology of reading.* Hillsdale, NJ: Erlbaum. Pp. 3–36.

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

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

McClelland, J. L., & Rumelhart, D. E. 1986. Amnesia and distributed memory. In J. L. McClelland, D. E. Rumelhart, & the PDP research group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition. Volume 2: Psychological and biological models.* Cambridge, MA: MIT Press. Pp. 503–528.

McClelland, J. L., Rumelhart, D. E., & the PDP Research Group (Eds.) 1986. *Parallel distributed processing: Explorations in the microstructure of cognition. Volume 2: Psychological and biological models.* Cambridge, MA: MIT Press.

McCloskey, M., & Cohen, N. J. 1989. Catastrophic interference in connectionist networks: The sequential learning problem. In G. H. Bower (Ed.), *The psychology of learning and motivation.* New York: Academic Press.

Mitchum, C. C., & Berndt, R. S. 1990. Aphasia rehabilitation: An approach to diagnosis and treatment of disorders of language production. In M. G. Eisenberg (Ed.), *Advances in clinical rehabilitation.* New York: Springer-Verlag.

Mitchum, C. C., & Berndt, R. S. 1991. Diagnosis and treatment of the non-lexical route in acquired dyslexia: An illustration of the cognitive neuropsychological approach. *Journal of Neurolinguistics,* **6**(2), 103–137.

Morgan, N., & Bourlard, H. 1990. Generalization and parameter estimation in feedforward nets: Some experiments. In D. S. Touretzky (Ed.), *Advances in neural information processing systems 2.* San Mateo, CA: Morgan Kaufmann. Pp. 630–637.

Morton, J., & Patterson, K. 1980. A new concept at an interpretation, Or, an attempt at a new interpretation. In M. Coltheart, K. E. Patterson, & J. C. Marshall (Eds.), *Deep dyslexia.* London: Routledge & Kegan Paul. Pp. 91–118.

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

Newcombe, F., & Marshall, J. C. 1980. Transcoding and lexical stabilization in deep dyslexia. In M. Coltheart, K. E. Patterson, & J. C. Marshall (Eds.), *Deep dyslexia.* London: Routledge & Kegan Paul. Pp. 176–188.

Olson, A., & Caramazza, A. 1994. Representation and connectionist models: The NETspell experience. In G. D. A. Brown & N. C. Ellis (Eds.), *Handbook of spelling: Theory, process and intervention.* New York: Wiley. Pp. 337–363.

Paap, K. R., & Noel, R. W. 1991. Dual route models of print to sound: Still a good horse race. *Psychological Research,* **53,** 13–24.

Patterson, K. E. 1994. Reading, writing and rehabilitation: A reckoning. In M. J. Riddoch & G. W. Humphreys (Eds.), *Cognitive neuropsychology and cognitive rehabilitation.* Hillsdale, NJ: Erlbaum.

Patterson, K. E., Coltheart, M., & Marshall, J. C. (Eds.) 1985. *Surface dyslexia.* Hillsdale, NJ: Erlbaum.

Patterson, K. E., & Morton, J. 1985. From orthography to phonology: An attempt at an old interpretation. In K. E. Patterson, M. Coltheart, & J. C. Marshall (Eds.), *Surface dyslexia.* Hillsdale, NJ: Erlbaum. Pp. 335–359.

Patterson, K. E., Seidenberg, M. S., & McClelland, J. L. 1990. Connections and disconnections: Acquired dyslexia in a computational model of reading processes. In R. G. M. Morris (Ed.), *Parallel distributed processing: Implications for psychology and neuroscience.* London: Oxford Univ. Press, 131–181.

Plaut, D. C. 1991. *Connectionist neuropsychology: The breakdown and recovery of behavior*

in lesioned attractor networks. PhD thesis, School of Computer Science, Carnegie Mellon University. Available as Technical Report CMU-CS-91-185.

Plaut, D. C. 1995. Double dissociation without modularity: Evidence from connectionist neuropsychology. *Journal of Clinical and Experimental Neuropsychology,* **17,** 291–321.

Plaut, D. C., Behrmann, M., Patterson, K. E., & McClelland, J. L. 1993. Impaired oral reading in surface dyslexia: Detailed comparison of a patient and a connectionist network [Abstract 540]. *Psychonomic Society Bulletin,* **31,** 400.

Plaut, D. C., & McClelland, J. L. 1993. Generalization with componential attractors: Word and nonword reading in an attractor network. In *Proceedings of the 15th Annual Conference of the Cognitive Science Society.* Hillsdale, NJ: Erlbaum. Pp. 824–829.

Plaut, D. C., McClelland, J. L., Seidenberg, M. S., & Patterson, K. E. in press. Understanding normal and impaired word reading: Computational principles in quasi-regular domains. *Psychological Review.*

Plaut, D. C., & Shallice, T. 1993a. Deep dyslexia: A case study of connectionist neuropsychology. *Cognitive Neuropsychology,* **10**(5), 377–500.

Plaut, D. C., & Shallice, T. 1993b. Perseverative and semantic influences on visual object naming errors in optic aphasia: A connectionist account. *Journal of Cognitive Neuroscience,* **5**(1), 89–117.

Posner, M. I., & Keele, S. W. 1968. On the genesis of abstract ideas. *Journal of Experimental Psychology,* **77,** 353–363.

Quinlan, P. 1991. *Connectionism and psychology: A psychological perspective on new connectionist research.* Chicago: University of Chicago Press.

Ratcliff, R. 1990. Connectionist models of recognition memory: Constraints imposed by learning and forgetting functions. *Psychological Review,* **97,** 285–308.

Reggia, J. A., Marsland, P. M., & Berndt, R. S. 1988. Competitive dynamics in a dual-route connectionist model of print-to-sound transformation. *Complex Systems,* **2,** 509–547.

Riddoch, M. J., & Humphreys, G. W. (Eds.) 1994. *Cognitive neuropsychology and cognitive rehabilitation.* Hillsdale, NJ: Erlbaum.

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

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. 1986. Learning representations by back-propagating errors. *Nature,* **323**(9), 533–536.

Rumelhart, D. E., McClelland, J. L., & the PDP Research Group (Eds.) 1986. *Parallel distributed processing: Explorations in the microstructure of cognition. Volume 1: Foundations.* Cambridge, MA: MIT Press.

Scott, C., & Byng, S. 1989. Computer assisted remediation of a homophone comprehension disorder in surface dyslexia. *Aphasiology,* **3,** 301–320.

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

Seidenberg, M. S., & McClelland, J. L. 1989. A distributed, developmental model of word recognition and naming. *Psychological Review,* **96,** 523–568.

Sejnowski, T. J., & Rosenberg, C. R. 1987. Parallel networks that learn to pronounce English text. *Complex Systems,* **1,** 145–168.

Seron, X., & Deloche, G. (Eds.) 1989. *Cognitive approaches in neuropsychological rehabilitation.* Hillsdale, NJ: Erlbaum.

Shallice, T., Glasspool, D. W., & Houghton, G. 1995. Can neuropsychological evidence inform connectionist modelling? Analyses of spelling. *Language and Cognitive Processes,* **10,** 195–225.

Shallice, T., & McCarthy, R. 1985. Phonological reading: From patterns of impairment to possible procedures. In K. E. Patterson, M. Coltheart, & J. C. Marshall (Eds.), *Surface dyslexia.* Hillsdale, NJ: Erlbaum. Pp. 361–398.

Shallice, T., & McGill, J. 1978. The origins of mixed errors. In J. Requin (Ed.), *Attention and performance VII.* Hillsdale, NJ: Erlbaum. Pp. 193–208.

Shallice, T., & Warrington, E. K. 1980. Single and multiple component central dyslexic syndromes. In M. Coltheart, K. E. Patterson, & J. C. Marshall (Eds.), *Deep dyslexia*. London: Routledge & Kegan Paul. Pp. 119–145.

Van Orden, G. C., Pennington, B. F., & Stone, G. O. 1990. Word identification in reading and the promise of subsymbolic psycholinguistics. *Psychological Review*, **97**(4), 488–522.

Weekes, B., & Coltheart, M. in press. Surface dyslexia and surface dysgraphia: Treatment studies and their theoretical implications. *Cognitive Neuropsychology.*

Williams, R. J., & Peng, J. 1990. An efficient gradient-based algorithm for on-line training of recurrent network trajectories. *Neural Computation*, **2**(4), 490–501.

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

Wilson, B. A. 1989. Models of cognitive rehabilitation. In R. L. Wood & P. Eames (Eds.), *Models of brain injury rehabilitation.* London: Chapman & Hall. Pp. 117–141.