Psychological Review
1991, Vol. 98, No. 1, 74–95

0033-295X/91/$3.00

Lesioning an Attractor Network: Investigations of Acquired Dyslexia

Geoffrey E. Hinton
Department of Computer Science, University of Toronto
Toronto, Ontario, Canada

Tim Shallice
Medical Research Council, Applied Psychology Unit
Cambridge, England

A recurrent connectionist network was trained to output semantic feature vectors when presented with letter strings. When damaged, the network exhibited characteristics that resembled several of the phenomena found in deep dyslexia and semantic-access dyslexia. Damaged networks sometimes settled to the semantic vectors for semantically similar but visually dissimilar words. With severe damage, a forced-choice decision between categories was possible even when the choice of the particular semantic vector within the category was not possible. The damaged networks typically exhibited many mixed visual and semantic errors in which the output corresponded to a word that was both visually and semantically similar. Surprisingly, damage near the *output* sometimes caused pure visual errors. Indeed, the characteristic error pattern of deep dyslexia occurred with damage to virtually any part of the network.

A connectionist network consists of a large number of relatively simple neuronlike processing elements that interact in parallel by means of weighted connections. The connection weights encode the long-term knowledge of the network. Some networks are organized into layers, with no feedback connections from later layers to earlier ones, but other networks are more complex. They have feedback connections and can exhibit resonant or attractor states: Under the influence of an external input vector, the network settles into a stable state that represents an interpretation of that input.

Connectionist models are becoming increasingly popular within psychology for various reasons. Early models showed that a set of simple pairwise associations between patterns of activity could be stored by modifying the weights. Each weight is involved in storing many associations, and each association is stored by many weight changes (Anderson, Silverstein, Ritz, & Jones, 1977; Kohonen, 1977; Willshaw, Buneman, & Longuet-Higgins, 1969). Later, it was shown that structured propositional information could be represented as distributed patterns of activity, and these patterns could be made into stable attractor states by suitable weight modifications (Hinton, 1981).

Newer learning procedures are now capable, in principle, of learning appropriate distributed representations. These new learning procedures operate in networks that contain internal, hidden units that are not part of the input or output (Ackley, Hinton, & Sejnowski, 1985; Rumelhart, Hinton, & Williams, 1986). The networks construct their own internal representations in the hidden units, and this enables them to solve tasks that are too difficult for networks that lack hidden units. These networks can discover implicit semantic features (Hinton, 1986); solve computationally difficult problems, such as correctly pronouncing English text (Seidenberg & McClelland, 1989; Sejnowski & Rosenberg, 1986); and perform many other tasks.

One of the main arguments in favor of connectionist models is that the most effective ways of performing computations in these networks are likely to resemble the most effective ways of performing computations in the brain because the hardware is similar. One piece of evidence that is often offered for the broad similarity between brains and connectionist models is that, like brains, connectionist networks frequently degrade gracefully when they are damaged. This crude, quantitative argument would be far more compelling if specific qualitative effects of damaging a connectionist network could be shown to resemble specific qualitative effects of brain damage. Our aim in this article is to demonstrate that some specific neuropsychological phenomena that are intuitively surprising when viewed within a conventional information-processing framework become natural and unsurprising when viewed within a connectionist framework.

The effects with which we are concerned occur in forms of acquired dyslexia in which the patient cannot obtain the phonological representation of a written word without first accessing a semantic representation. Thus, in so-called deep dyslexia (Coltheart, Patterson, & Marshall, 1980; Marshall & Newcombe, 1966), a patient who is shown the word *peach* printed on a card and asked to read it can say "apricot." In the input or central forms of deep dyslexia, this effect cannot be reduced to a problem in selecting the incorrect name (Newcombe & Marshall, 1980a; Shallice & Warrington, 1980); the patient misunderstands the word that has been presented. This is a puzzle for any straightforward information-processing model that postulates a lexicon containing discrete entries that can be accessed from the visual form of the word. The entry for *peach* must still be present because it is required for mapping from the visual

This research was supported by Grant 87-2-36 from the Alfred P. Sloan Foundation. Geoffrey E. Hinton is the Noranda fellow of the Canadian Institute for Advanced Research.

We thank Ian Nimmo-Smith for his advice and Alfonso Caramazza, Mark Seidenberg, and an anonymous referee for many helpful comments.

Tim Shallice is now at the Psychology Department, University College, London.

Correspondence concerning this article should be addressed to Geoffrey E. Hinton, Department of Computer Science, University of Toronto, 10 Kings College Road, Toronto, Ontario, Canada M5S 1A4.

form of *peach* to the meaning, and the error clearly depends on the meaning of the word *peach*.

How would a connectionist model account for this phenomenon? It would be straightforward if the space of semantic representations was completely filled by regions corresponding to the meanings of individual words. Then any excessive noise, at least at the later stages of processing, would be liable to give rise to semantic errors. However, it seems implausible that the meanings of individual words are represented in such a fashion. An entity halfway, conceptually, between a prototypic rhinoceros and a prototypic unicorn is not likely to be within the domain of any word in the lexicon. If not all conceivable semantic representations are equally acceptable as nameable entities, it is useful to build attractors corresponding to each familiar, nameable concept. Then, even if the input to the semantic system is noisy, the state of the network will be more likely to move toward one of its learned representations; it will automatically clean up its input. Normally, the representation toward which it will move will be that corresponding to the input, but if the system is damaged, it can easily move to a nearby attractor, which will presumably correspond to the meaning of a related word. This provides a simple explanation of the peach–apricot phenomenon.

Another puzzling aspect of the reading of all deep-dyslexic patients so far described is that the errors they make are not only semantic; there is also a visual component to the errors. In its most obvious form, there is a simple co-occurrence of the so-called visual errors (e.g., *mat* → "rat") with the semantic errors, which has virtually always been observed in dyslexic patients who produce some semantic errors (Coltheart, Patterson, & Marshall, 1987). In the few patients in whom it has been investigated, there has also been a higher rate of mixed errors such as *cat* → "rat," which are similar both in visual form and in meaning, than would be expected from the rate of the two types of errors in isolation (Shallice & Coughlan, 1980; Shallice & McGill, 1978). Intuitively, there seems to be no reason why damage to later parts of the system should cause the error corpus to have a visual component, so it is surprising that this should occur for virtually all the relevant patients. In fact, our simulations show a similar phenomenon: Damage to later, semantic parts of our connectionist network leads not only to semantic errors but also to visual and to mixed errors. No extra mechanism or tuning was required to produce this effect, and it took us some time to understand why it was occurring.

To understand the effect, it is necessary to escape from the view that the visual form of a word acts as a purely arbitrary pointer to the meaning. In a connectionist network, similar inputs tend to cause similar outputs, and generally a lot of training and large weights are required to make very similar inputs give very different outputs. Now, if each meaning has a large basin of attraction, the network is free to make the visual form of the word point to any location within this basin, so the network will, if it can, choose to make visually similar words point to nearby points in semantic space (Figure 1).[1] Damage that moves the boundaries of the basins of attraction in semantic space will then have a tendency to cause mixed errors or even purely visual ones. This would be a fairly minor effect in the two-dimensional semantic space shown in Figure 1 if there are many targets because each basin of attraction must be a continuous region. If, however, the meanings are patterns of activity over a large number of units (68 in our simulation), the effect becomes much more pronounced (see Appendix A).

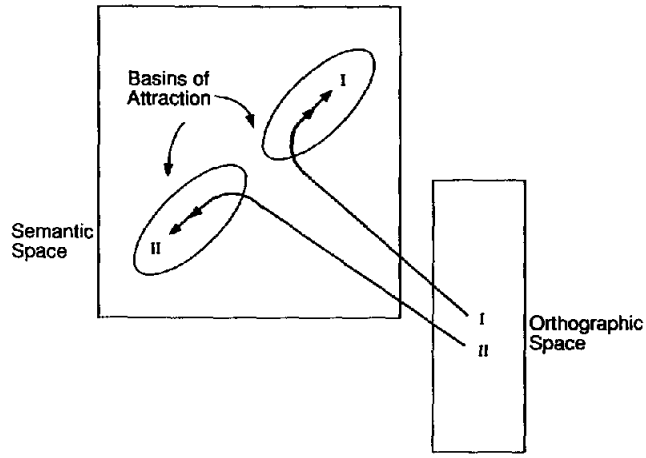

Figure 1. A two-dimensional representation of high-dimensional basins of attraction illustrating how visually similar words can point to nearby parts of semantic space even though the meanings of the words are far apart in semantic space. (The bottom-up orthographic input does not need to point exactly to the meanings because the attractors can clean up the effect of the bottom-up input. By choosing appropriately shaped basins of attraction, the network can allow visually similar words such as I and II to point to nearby parts of semantic space, even though their meanings are far apart. Thus, small changes in the basins of attraction in semantic space can cause the network to erroneously settle on the meaning of a visually similar word.)

Modeling Impaired Reading to Meaning

The Neuropsychological Domain: Impairments in Reading to Meaning

The empirical phenomena described in the introduction come from the acquired dyslexias, a group of disorders that have been intensively studied by the single-case method in recent years. For certain patients having an acquired dyslexia with an impairment at a more central locus, reading aloud appears not to involve semantic mediation (Bub, Cancelliere, & Kertesz, 1985; McCarthy & Warrington, 1986; Schwartz, Saffran, & Marin, 1980; Shallice, Warrington, & McCarthy, 1983).

In complementary patients, such as the deep dyslexics mentioned in the introduction, however, the accessing of semantic information does not seem to involve the attaining of any phonological representations. The patients make semantic errors in reading aloud (Marshall & Newcombe, 1966). They are poor at reading pronouceable nonwords aloud and carrying out rhyme judgements on written words (Patterson & Marcel, 1977). Also, the probability of a word's being read is strongly influenced by semantic variables, such as imageability, but not by the regularity of its spelling-to-sound mapping (Patterson, 1981).

[1] The idea that the visual form points to a particular point in semantic space is a simplification. It would be more accurate to think of each visual input as contributing a different potential function that is added to the semantic potential function, the minima of which are the word meanings.

A standard way to explain the existence of this contrasting pattern of disorders is by assuming a functional architecture in which there are two (or more) reading routes, and one (or more) is impaired in each set of disorders (Marshall & Newcombe, 1973). Recently, the suggestion has been made on the basis of experiments using word-to-category matching that the different types of transformation of the orthographic input should not be conceived of as independently operating processes, as in horse-race models, but instead should be considered as a global connectionist mechanism reflecting the covariance between orthographic and all other linguistic features (syntactic, semantic, and phonological) in its associative weights (Van Orden, 1987). Although we are sympathetic to this as a possible eventual way of modeling the reading system, any physical separation of the material underpinning of, say, phonological and semantic processing would make it appropriate to model each of the two classes of disorder in terms of separate complementary transformations of the orthographic representations. In any case, this approach seems an appropriate starting point for the development of any more complex overall model of disorders of the reading process. The use of the term *route* is intended to refer to these abstractions from a potentially more complex total process.

In recent articles, Sejnowski and Rosenberg (1986) and Seidenberg and McClelland (1989) have attempted to provide a connectionist model for spelling-to-sound translation, and Patterson, Seidenberg, and McClelland (1989) have interpreted neuropsychological evidence about the first group of disorders in terms of one of the models. In this article, we are concerned with the modeling of the complementary process, by which a semantic representation is accessed from an orthographic one without phonological mediation, and with how that process might break down if lesioned. Certain aspects of the performance of the second set of patients flow directly from assuming that the two types of process are carried out by separate procedures and that one—the phonological transformation—does not operate in these patients. The inability to carry out rhyme judgments on the written word and to read pronounceable nonwords are transparent phenomena of this type. Other aspects of the behavior of the patients require a more detailed account of how the semantic transformation might be operating. These phenomena are addressed by the present model.

In the preceding paragraphs, we refer to *patients* in a loose fashion and imply that many patients have a set of characteristics in common and that this pattern occurs consistently when one or two individual characteristics occur. This is the strong-syndrome approach, which has been criticized as a general theoretical approach within neuropsychology (Caramazza, 1984; Schwartz, 1984), particularly for deep dyslexia (Shallice & Warrington, 1980). On the other hand, the starkly contrasting unique-case approach (Caramazza, 1986) is also controversial (Shallice, 1988). As a compromise, we use the strong-syndrome approach in the introduction, but in the Discussion section, we refer specifically to patients whose behavior does not fit the claimed general pattern. Moreover, phenomena investigated in only a small subset of patients are labeled by *patient*.

We adopt this procedure for three reasons. First, relevant evidence is available on a large number of patients. For alphabetic languages alone, 16 cases were reviewed by Coltheart (1980a), 2 of which were rejected by Shallice (1988) as insufficiently detailed; 8 more were reviewed by Kremin (1982); and another 10 are referred to in a recent review by Coltheart et al. (1987). In what is essentially a modeling project, it would be inappropriate to review the empirical information on such a large set of patients. Second, in this domain, recent reviewers, even when rejecting the strong-syndrome approach at the metatheoretical level, have commented on how close an approximation the approach provides to the empirical situation for certain aspects of the behavior of the group of patients they considered, namely patients who made semantic errors in reading aloud (Coltheart et al., 1987). Third, our theoretical model makes the prediction that qualitative similarities should occur in the behavior of patients even with quite wide differences in the functional location of the lesion. We must therefore refer to behavior in a wider sample than the individual patient.

The most basic effect we are considering is the semantic error. Researchers now generally accept that for such errors to occur, the semantic route must be damaged in some way.[2] Several accounts have been given of what that damage might consist of, such as those provided by Caramazza and Hillis (1990), Coltheart (1980b), Howard (1985), and Marshall and Newcombe (1966). The issue is complicated by the way that there appear to be different loci across patients for the impairment to the semantic route (see Shallice, 1988, for review). For some patients, such as those described by Caramazza and Hillis (1990) and Patterson (1979), the primary impairment to the semantic route is held to lie in the accessing of output phonology from a semantic representation. In most patients, though, in whom semantic errors in reading have been observed, the problem must lie at an earlier stage of the process because matching the written word to one of a set of pictures produces similar semantic errors to reading aloud (e.g., Coslett, Rothi, & Heilman, 1985; Friedman & Perlman, 1982; Newcombe & Marshall, 1980a). Some of this group of patients, moreover, have a difficulty in accessing semantics, which is much more severe for visual than for auditory input (e.g., Sartori, Bruno, Serena, & Bardin, 1984; Shallice & Coughlan, 1980; Shallice & Warrington, 1980), which suggests that their problem lies on the input side of the overall process. In this article, we are concerned only with reading impairments up to the level of semantic representations, so patients whose semantic errors arise from damage to some level of the speech production process would be excluded.

Despite the differences in pattern of performance across reading tasks, almost all the patients we refer to earlier as being in the second group had a qualitatively similar error pattern. They made semantic errors (e.g., *cat* → "mice"); visual errors (e.g., *patent* → "patient"); mixed visual and semantic errors (e.g., *last* → "late"); and derivational errors (e.g., *bake* → "baker"; e.g., see Coltheart et al., 1980).[3]

Why should visual errors co-occur with semantic ones? The straightforward explanation is that the visual errors arise from an additional impairment to a visual lexicon, logogen, or word-

[2] See Newcombe and Marshall (1980b) for a contrary view and Morton and Patterson (1980), Nolan and Caramazza (1982), and Shallice (1988) for criticisms.

[3] Exceptions are considered in the Discussion section.

form system (Gordon, Goodman-Schulman, & Caramazza, 1987; Patterson, 1978). The strongest argument for this position concerns the consistency with which particular types of errors occur for particular words, if words are presented again on another occasion. It was argued by Gordon et al. that if an error arose from the lack of an orthographic entry, then any future error on that word would also be likely to be visual, and for similar but not identical reasons, a future error on a word giving rise to a semantic error would be likely to be semantic. Gordon et al. found such an effect in Patient FM. This argument is considered in the Discussion section. However, on the two-impairment position, it would be natural to expect that the variables that determine on which words visual errors occur should relate to visual aspects of words and not to semantic ones. In fact, it is words in those semantic and syntactic classes which patients find most difficult to read that produce the highest rate of visual errors (Patient GR in Barry & Richardson, 1990; Patient PD in Coltheart, 1980c; Patient FM in Gordon et al., 1987; Patients KF and PS in Shallice & Warrington, 1980). The type of explanation that has been offered for this finding is that candidate lexical outputs of orthographic analysis are passed to higher systems in parallel (Gordon et al., 1987; Shallice & Warrington, 1980). However, no formal model of such a way of generating visual errors has been put forward, and therefore, whether two separate impairments would be required on such a model is unclear.

A second difficulty for explaining visual and semantic errors as stemming from impairments at two separate stages of the reading process concerns mixed visual and semantic errors. If the visual errors and the semantic errors arise independently, then one can estimate the upper bound for the number of errors that are similar on both dimensions (see Shallice & McGill, 1978). In the two patients in whom it has been investigated, the number of mixed errors exceeds the upper bound (Patient KF in Shallice & McGill, 1978; Patient PS in Shallice & Coughlan, 1980). The approach of having candidate orthographic outputs in parallel would seem to be able to give an account of this phenomenon too. This account is, however, on the purely verbal level. The model we put forward is one way of making the proposal formal.[4]

A final, nontransparent problem in this domain that we want to address is posed by certain phenomena observed in two patients (Patient JE in Rapp & Caramazza, 1989; Patient AR in Warrington & Shallice, 1979). The patients were unable to read or identify many words for which they were able to perform at a very-much-above-chance level in a forced-choice category or attribute judgment task. This effect could not be attributed to a mere problem of producing the name when reading. For instance, Patient AR could not give an appropriate mime for a written word stimulus that he could not read aloud, although he could produce one if presented with a spoken word.

Similar effects occur in what has been thought to be a separate group of acquired-dyslexic patients in whom phonological mediation is not possible. These are certain patients who are not in general capable of explicit identification of words in reading except by most unusual procedures. This is the case in the original form of dyslexia isolated by Dejerine (1892)—variously called *pure alexia*, *agnosic alexia*, or *word-form alexia*—in which patients are in general not aphasic and attempt to read by the so-called letter-by-letter procedure, laboriously reconstructing the word from the sounds of its constituent letters. More recently, research has shown that when words are exposed for too brief an interval for the letter-by-letter strategy to be used, certain patients cannot name or identify more than a small proportion of them but can perform, say, a categorical decision about them at well-above-chance levels (see Coslett & Saffran, 1989; Shallice & Saffran, 1986; see also Landis, Regard, & Serrat, 1980).

Again, this is a phenomenon that is not transparently explicable from the characteristics of the functional architecture. Various suggestions have been made for how to account for it (e.g., see Howard, 1985; Humphreys, Riddoch, & Quinlan, 1988; Rapp & Caramazza, 1989; Shallice & Saffran, 1986; Warrington & Shallice, 1979).[5] As yet, no explanation for this finding in these patients has been generally accepted.

In this article, we consider three phenomena—the occurrence of semantic errors in patients whose impairments lie in accessing semantic representations, the co-occurrence of such errors with visual errors, and the relative sparing of categorization performance by contrast with explicit identification in another group of patients who partly overlap the previous group. Our aim is not specifically to contrast the explanation for the phenomena that our approach provides with alternative explanations in the literature. It is to show that a connectionist model of the domain produces as interrelated effects the three phenomena when other current approaches seem to require several independent assumptions to explain them.

A Previous Connectionist Model of Acquired Dyslexia

One of the earliest simulations of the effects of damage to a multilayer network, that of Hinton and Sejnowski (1986), is the one most directly relevant to this article. This study was designed to show that an arbitrary mapping between two virtually independent domains could be learned with the use of distributed representations. The domains used were the orthographic and semantic ones. After the network had successfully learned the mapping, it was lesioned and showed behavior somewhat similar to that occurring in the acquired-dyslexic disorders considered herein. The network they investigated consisted of three layers of units. The grapheme group contained 30 units, which represented the letters in three-letter words. The sememe group, which also contained 30 units, represented the semantic features of a word. There were no direct connections between the grapheme and sememe units. Instead, there was an intermediate layer of 20 units, each of which was connected to all the units in both the grapheme and sememe groups. Unlike the network we describe later, all the connections were symmetri-

[4] Alternative suggestions with some similarities to the explanation offered in this article but that were not based on simulations were given by Morton and Patterson (1980) and Shallice and Warrington (1975).

[5] The least interesting possibility is that the categorization task narrows the range of possible words and allows the patient to guess the word from identifying one or two letters (see Rapp & Caramazza, 1989). This is not a very plausible account for the patients studied by Coslett and Saffran (1989), who used stimuli to control for this possibility, or for Patient AR (Warrington & Shallice, 1979), who was very poor at identifying letters. The possibility was also explicitly tested and rejected by Shallice and Saffran (1986).

cal, and the units were stochastic binary processors. A binary input pattern was clamped on the grapheme units, and the network was allowed to settle for a while before a binary output pattern was read off from the sememe units. During the settling process, units that are not clamped compute the total input they are receiving from other active units and make repeated stochastic decisions about whether to be on or off. After a while, the network reaches the equivalent of thermal equilibrium, which means that the probability of finding it in any particular global state remains constant even though the units continue to change states.

The Boltzmann machine learning procedure (Hinton & Sejnowski, 1986) was used to train the network to associate 20 patterns of activity in the grapheme units (representing 20 short words) with 20 patterns of activity in the sememe units (representing the meanings of words). The patterns used to represent meanings were chosen at random. After prolonged training (5,000 sweeps through the entire training set), the network was able to select the semantic representation, which was exactly correct more than 99.9% of the time provided it was allowed to settle to equilibrium slowly.

The network was then damaged either by adding noise to all the weights or by setting a percentage of the weights to zero or by removing units in the intermediate layer. For example, when 20% of the weights were set to zero, the performance of the network dropped to 64%. However, relearning was extremely rapid. Within three sweeps through the training set, it reached 90%. By comparison, during the original learning when performance had reached roughly the same level of 64% correct, 30 more sweeps increased it by less than 10%. It is noteworthy that neurological patients, too, often show rapid improvement in performance after a lesion occurs, although this is not always found. Why this improvement occurs is not understood (Geschwind, 1985). Moreover, Coltheart and Byng (1989) have recently shown in an acquired-dyslexic patient that retraining reading on one group of words benefited not only that group of words but also performance on a second, untrained set, which is an effect equivalent to that observed with the lesioned network.

For our purpose the most directly relevant investigation that Hinton and Sejnowski (1986) carried out was an analysis of the effects of removing single units in the intermediate layer. The network error rate then increased from less than 0.1% to 1.4%, and 59% of these errors were the precise meaning of an alternative word. An analysis of the whole-word errors showed them to be both semantically and visually significantly more similar to the correct word than a word of the set selected by chance. Clearly, lesioning only a single unit in a network and reducing its performance to 98.6% is not an adequate simulation of the way a lesion causes a neurological syndrome. However, the way that visual and semantic errors co-occur when only a single layer of the network is lesioned suggested that a more detailed investigation of the effects of damage in such a network would be worthwhile.

In this article, we describe a more systematic study of the effect of damage in a related network that uses nonstochastic units and a more efficient training procedure. Our aim in the investigation is not to produce a complete model of the reading-to-meaning process. This would require the use of a large set of words representative of the full English language, both orthographically and semantically, and a relatively complete representation of their underlying semantics. The first of these requirements would prove computationally very demanding, and the second is not possible with the present understanding of semantics.

Instead, our aim is more limited. It is to explore the behavior of a network that maps from orthographic representations to semantic features when it is subject to different forms of damage. If its properties are similar to those observed in acquired dyslexia, this will provide a hypothesis for the origin of these characteristics in patients.

The Network

To simulate any empirical domain in a connectionist fashion, many design decisions have to be made about the network. This section describes the detailed specification of the network and why we made the particular choices. In the Discussion section, we consider the more general issue of what aspects of the design decisions were critical for the effects obtained. Most important, we claim, is that the network builds attractors.

The Units

Many different types of units have been used in connectionist models. These include linear units, deterministic binary threshold units, stochastic binary threshold units, and units with output that is a real-valued, deterministic, nonlinear function of the total input received. The outputs of units in this last class are often interpreted as approximations to the firing rates of neurons. For the learning rule used, it is normal to use units of this type with output y related to their total input x by the logistic function:

$$y_i = \frac{1}{1 + e^{-x_j}}. \tag{1}$$

The total input to a unit includes a threshold term and a weighted sum of the activities of other, connected units:

$$x_j = -\theta_j + \sum_i y_i w_{ji}, \tag{2}$$

where y_i is the state of the ith unit, w_{ji} is the weight on the connection from the ith to the jth unit, and θ_j is the threshold of the jth unit. The threshold term can be eliminated by giving every unit an extra input connection, the activity level of which is fixed at 1. The weight on this special connection is the negative of the threshold. It is called the *bias*, and it can be learned in just the same way as the other weights.

Representation of the Input

The network maps from the visual form of a word to its meaning. We assume that the primitive components are letters and that their positions are represented relative to a reference frame based on the word itself. Each input unit in the network therefore represents the conjunction of a letter identity and a position within the word, so the input units of our network correspond to the letter-level units used by McClelland and Rumelhart (1981). Psychological evidence compatible with the existence of such units in humans can be obtained from the

Table 1
Words Used in the Model

Indoor objects	Animals	Body parts	Foods	Outdoor objects
Bed	Bug	Back	Bun	Bog
Can	Cat	Bone	Ham	Dew
Cot	Cow	Gut	Hock	Dune
Cup	Dog	Hip	Lime	Log
Gem	Hawk	Leg	Nut	Mud
Mat	Pig	Lip	Pop	Park
Mug	Ram	Pore	Pork	Rock
Pan	Rat	Rib	Rum	Tor

study of migration errors in pattern masking (e.g., Mozer, 1983) and from the preservation of word length in errors made by neglect-dyslexic patients (see Ellis, Flude, & Young, 1987). However, we do not see this design decision as critical to the effects obtained. To keep the network small, we restrict the input to three- or four-letter words that use only the letters {*b, c, d, g, h, l, m, n, p, r, t*} in the first position, {*a, e, i, o, u*} in the second position, {*b, c, d, g, k, m, p, r, t, w*} in the third position, and {*e, k*} in the fourth position. There are therefore 28 input units. These are called the *grapheme units*.

Representation of the Meaning of a Word

The simplest way to represent the meaning of a word in a connectionist network is to use binary or real-valued semantic features and to dedicate a single sememe unit to each semantic feature. The meaning of the word is then a pattern of activity across the sememe units. This way of representing meaning appears to be very different from semantic networks (e.g., Collins & Loftus, 1975) or frames (Minsky, 1975), which encode relationships between entities, with special emphasis on the "IS-A" relationship between a class and its members. Fortunately, these more sophisticated representations can be implemented with sememe units provided that an individual unit is used to represent the conjunction of a role and some significant property of its filler (Derthick, 1987; Hinton, 1981). For example, the representation of *president* might contain an active unit that represents the conjunction of the role "has-job" and the filler "important." Notice that this is just another example of the method we use for representing the input. Each active input unit represents a binding between a role (i.e., a spatial position within the word) and a filler (i.e., a letter identity).

To reduce the computational load, we used a restricted set of 40 words, all of three or four letters and falling into five concrete categories: indoor objects, animals, parts of the body, food, and outdoor objects (see Table 1). The complete set of features used for the words is shown in Appendix B. The use of several categories enabled us to mimic the category-selection tasks used in semantic-access dyslexia.

We also assumed that identification of a word did not require that all its features be fully activated. Instead, if the network settled to a semantic representation sufficiently close to the ideal, the word was considered to be accessed. This approach is related to that used in the probabilistic feature models of semantic memory in psychological theorizing (see Smith & Medin, 1981).

Layers and Connections

The simplest network for associating the input vectors with the desired output vectors would have direct connections from grapheme to sememe units. Unfortunately, there are strong limitations on the computational abilities of such a simple network. It is not, in general, capable of representing a set of arbitrary associations between input and output vectors (Hinton, 1989). For example, in a network with two input units that are directly connected to one output unit, it is impossible to find any set of weights on the connections that represents the set of four associations {11 → 1, 10 → 0, 01 → 0, 00 → 1}. In general, it is necessary to introduce one or more layers of nonlinear hidden units between the input and output of the network (Ackley et al., 1985). These hidden units detect higher order combinations of activities among the units to which they are connected. We use a network that contains only one layer of hidden units (called *intermediate units*) between the graphemes and sememes.

Some of the phenomena we describe in this article can be observed, to some degree, in a simple layered net in which the grapheme units completely determine the activities of the intermediate units, and these in turn completely determine the activities of the sememe units. Our simulation, however, is based on the assumption that the semantic space contains attractors. There are many ways to realize that possibility. It is clearly more appropriate if the attractors are not handcrafted and the system builds them itself. The simplest way to enable that to happen is to introduce recurrent connections in a network model so that the process of accessing a word's meaning corresponds to settling to a stable state.

In determining the number of hidden units and the pattern of connections, we were influenced by three considerations. First, the time it takes the network to learn, given the algorithm we used, increases rapidly with the number of connections, so on the workstation we were using it was difficult to experiment with networks containing more than a few thousand connections.

Second, one needs a sufficient number of connections to store all 40 associations of word forms with word meanings. If we make the simplifying assumption that the sememes are independent variables, the information H in a single association is given by

$$H = -\sum_i p_i \log_2 p_i + (1 - p_i)\log_2(1 - p_i), \qquad (3)$$

where p_i is the probability (measured over the whole set of associations) of an individual sememe's being active. Assuming that all of the p_i are 0.22 (which is very approximately true for our simulations), the information in the whole set of 40 associations is given by

$$H = -40 \times 68(0.22 \log_2 0.22 + 0.78 \log_2 0.78) = 2{,}068. \qquad (4)$$

A good rule of thumb for the storage capacity of a network that uses the backpropagation learning procedure is two bits per weight, so the network should contain at least 1,034 connections. The network we used contained about 3,300 (including biases).

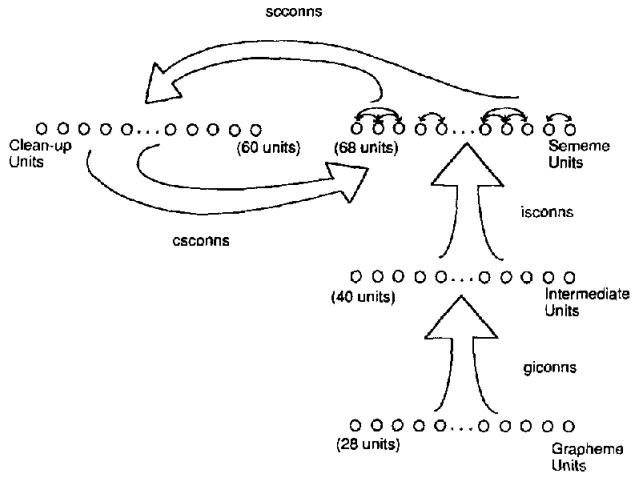

Figure 2. The groups of units and their connectivity. (The connections between groups are named by using the first letters of each group name [*scconns, csconns, isconns, giconns*]. In addition to the intergroup connections, there are direct connections between pairs of highly related sememe units. The highly related sets of sememes are defined in Appendix B.)

The third consideration is to encourage the network to build strong attractors. We achieved this by making the bottom-up inputs to the sememe units somewhat impoverished and providing the potential for rich interactions to be developed between the sememe units. The groups of units and their connectivity are shown in Figure 2. Connections between any two sets of units *a* and *b* are labeled *abconns*. We chose to use a probability of .25 for including each potential connection between a unit in one group and a unit in another connected group. The existence of direct connections between the sememe units allows the network to develop lateral inhibitory interactions between the activation of rival sememe units. For our task, however, there are potentially 4,624 such connections. So, instead of allowing all possible direct connections between sememes, we restricted such connections to all those within small subsets of sememes that correspond to different values on a dimension (see Appendix B). For sememes within a subset, there are typically strong negative correlations, and for those in different subsets, there is typically much less correlation. Significant interactions between sememes in different subsets can be implemented, if necessary, by the cleanup units, which can detect particular combinations of activity in the sememe units and "infer" that other sememe units should be active. Many such inferences potentially occur in parallel, and to avoid any implication that they correspond to conscious, deliberate inference, we call them *microinferences* (Hinton, 1981). The weights on the connections to and from the cleanup units were learned in the same fashion as all the others in the network.

Running the Network

The network is run for seven iterations, while the input units are clamped to a state that represents the current input word. The remaining units start off with activity levels of 0.2. To further encourage the network to build robust attractors, the network is trained to produce the correct activity pattern over the sememe units for the last three iterations. Figure 3 shows the seven successive states of activity for all the units when an undamaged network is presented with a word that it has learned.

The Learning Procedure

The network was trained using the iterative version of the backpropagation training procedure explained in Rumelhart et al. (1986). We do not believe that a literal implementation of this procedure is a good model for learning in the brain. The procedure is simply one of the many known ways of learning by gradient descent in a neural network. Other methods, such as "mean field learning" (Hinton, 1989), may be less unrealistic. For the research described herein, we simply wanted an efficient method of constructing networks that worked, and we are not concerned with the veridicality of the learning process.

The heart of the backpropagation procedure is just an efficient method of computing, for a given graphemic input vector, how small changes in the weights would affect the errors in the final activities of each sememe unit. The aim is to change the weights in the direction that reduces these errors. In the batch version of backpropagation, we sweep through all 40 training cases, computing the total derivative with respect to each weight of the error E for all sememes in all training cases. We then change each weight by an amount proportional to its total error derivative:

$$\Delta w_{ji} = -\epsilon \frac{\partial E}{\partial w_{ji}}. \tag{5}$$

This learning procedure can be pictured by imagining a multidimensional weight space that has an axis for each weight and one extra axis (called *height*) that corresponds to the total error measure. For each combination of weights, the network will have a certain error that can be represented by the height of a point in weight space. These points form a surface called the *error surface*. The learning procedure consists of moving the point that represents the weights down the error surface in the direction of steepest descent. This simple, gradient-descent procedure can be accelerated by adding to each weight change a fraction α of the previous weight change:

$$\Delta w_{ji}(t) = -\epsilon \frac{\partial E}{\partial w_{ji}}(t) + \alpha \Delta w_{ji}(t-1), \tag{6}$$

where t is incremented by 1 on each sweep through the 40 training cases. This momentum method speeds up the gradient descent along the bottoms of ravines in the error surface without causing divergent oscillations across the ravines.

Most simulations that use the backpropagation learning procedure assume that the appropriate error measure is the squared distance between the desired output vector and the output vector actually produced by the network. However, when the output units can be interpreted as representing discrete binary decisions (as they can in our network), it is more appropriate to use a different error measure, called the *cross-entropy* (Hinton, 1989).

$$E = -\sum_{j,c} d_{j,c} \log_2(y_{j,c}) + (1 - d_{j,c}) \log_2(1 - y_{j,c}), \tag{7}$$

where $d_{j,c}$ is the desired probability of output unit j in case c and $y_{j,c}$ is its actual probability.

This error measure can be understood as follows: We imag-

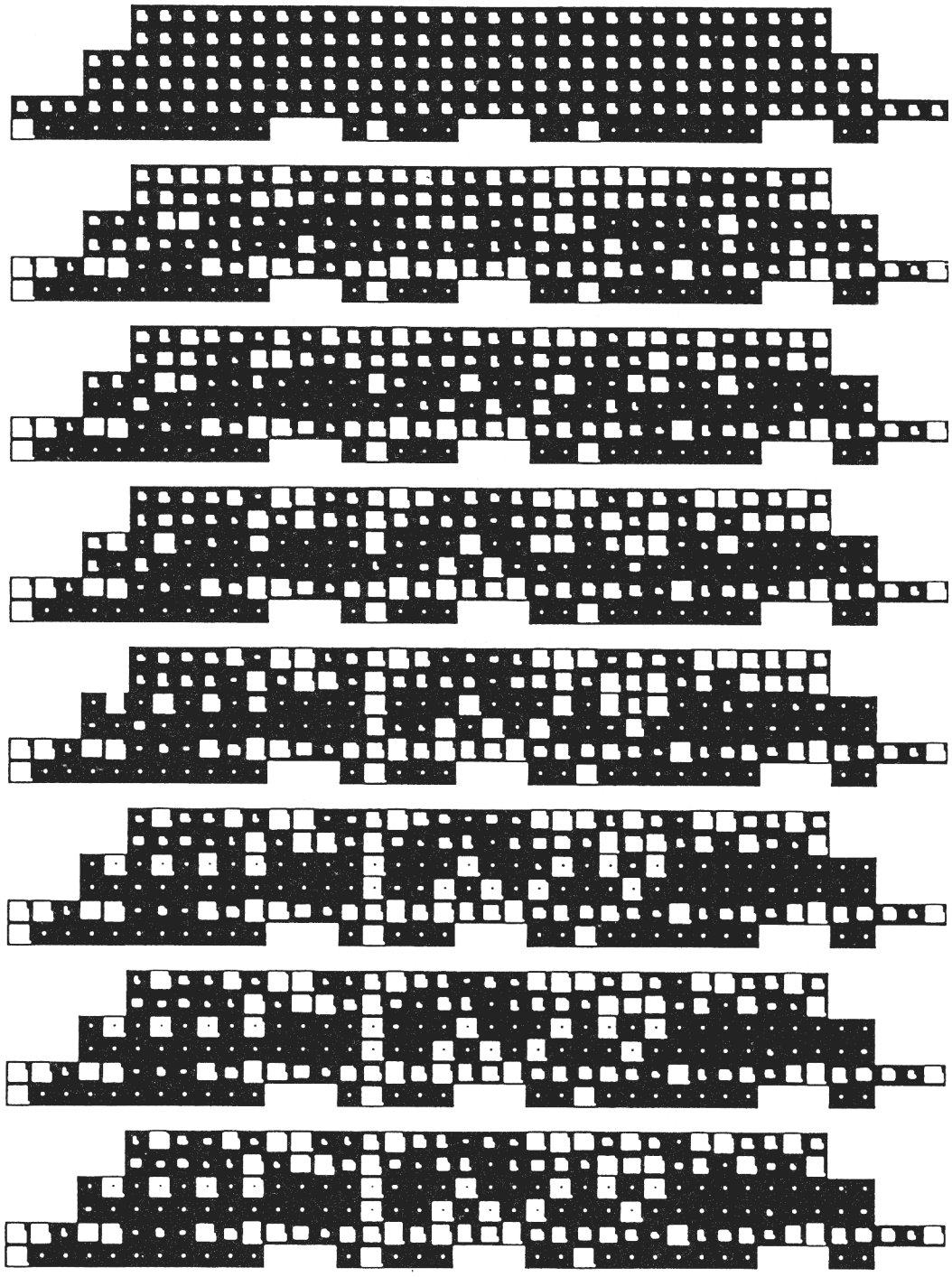

Figure 3. How the activations of all units in a network change as the network settles into a stable state with a fixed input. (The white blobs represent activation levels [with a small white dot representing zero]. The top panel shows the initial state of the network, and the lowest three panels show the final three time slices during which all sememe units that are not in the correct state receive error signals. The sememe units that should be active are indicated by small black dots. The top two rows within each diagram show the activation levels of the 60 cleanup units, the next two rows show the 68 sememe units, the fifth row shows the 40 intermediate units, and the bottom row shows the 28 grapheme units divided into four sets corresponding to the four-letter positions. Throughout the settling, the grapheme units are clamped in a fixed, binary state representing the input word *bed*. The activations are for an undamaged network after learning is complete.)

ine that the real-valued output vector produced by the network is stochastically converted into a binary vector by treating the real values as the probabilities that individual components have value 1 and assuming independence between these stochastic choices. We then compute the log probability that this binary vector exactly matches the desired vector. The negative of this log probability is the error measure.

Starting with small random weights and biases that are chosen from a uniform random distribution between −0.3 and 0.3, the learning requires about 1,000 sweeps through the training set with $\epsilon = 0.0005$. For the first 10 sweeps, α is set at 0.5, and after this it is set at 0.95.[6] The network was considered to have learned when the output of all sememes for all words was within 0.1 of the desired value over the last three iterations.

That the network develops attractors can be seen by inspection of Figure 3. The input remains on constantly from Time 0. Its direct effect reaches the sememes by Time 2. If one then compares their state at Time 2 with their state at Times 5, 6, or 7, one can see major differences for over 10 units for that word alone. The input from the intermediate units remains constant over the Period 2 to 7, as do the weights, so the change in the sememe activities must be caused by the interactions among them via direct connections and cleanup units. Indeed, the activation of the cleanup units themselves becomes much sharper between Times 1 or 2 and Times 5, 6, or 7.

The input and output weights learned by some of the intermediate units are shown in Figure 4. Intermediate units are not dedicated to particular words. Instead, each intermediate unit is activated by many different words and influences the activations of many different sememes. The third unit from the bottom, for example, is strongly activated by the letters *g* or *n* in the first position within a word and is strongly inhibited by *d* in the third position. So it is most active for the words *gem*, *gut*, and *nut* and least active for the words *bed* and *mud*. This intermediate unit strongly activates Sememes 44 and 59 and strongly inhibits Sememe 25. *Gem*, *gut*, and *nut* each have one or both of Sememes 44 and 59 but not Sememe 25. *Bed* and *mud* both have Sememe 25 but not 44 or 59. Note, however, that it is sometimes misleading to interpret units in isolation, because each unit has learned to have the optimal marginal effect given the current behavior of all the other units.

Network A and Network B

The network that we used for the main experiments on the effects of damage (Network A) was actually learned using a somewhat ad hoc procedure in which the parameter ϵ was manually adjusted as learning proceeded. The values used were all close to 0.0005. This was done in a vain attempt to make the learning take less than 56 hr on a Symbolics lisp machine. However, the subsequent simulation (Network B) using a different set of initial random weights maintained an epsilon value of 0.0005 after the first 10 sweeps. The behavior of the second network when lesioned showed qualitatively the same type of behavior as the first, with one exception, to be discussed later.

The Effects of Lesions: Results

Three alternative procedures were used to simulate the effect of a lesion on the network. First, each set of connections was taken in turn, and a specific proportion of their weights were set to 0. We use the notation *disconnect*(*csconns*, 0.3) to mean that a randomly chosen 30% of the connections from the *cleanup* units to the *sememe* units were set to 0. At each level of severity of the disconnection, the original, undamaged network was randomly damaged 10 different times so that we could see how much the effects of damage depended on which specific connections were removed. Second, for each set of connections, noise was added to the weight on each connection, with its value drawn independently from a uniform distribution between $-n$ and n; several different values for n were used to mimic different degrees of unreliability of neural connections. Again, for each value of n, random noise was added 10 different times to the original, undamaged network. We use the notation *noise*(*csconns*, 0.4) to mean that every *csconns* connection was given added noise uniformly distributed between −0.4 and 0.4. Finally, for the two sets of hidden units, the intermediate and the cleanup, a specific number of units were removed. As before, the number of units removed was varied, and for each value of the number, 10 randomly selected sets of units were eliminated. We use the notation *ablate*(*intermediate*, 7) to mean that 7 intermediate units were removed.

When a network has been lesioned, the mean value of the activation of the sememes over the last three iterations for each input word will differ from the stored-meaning vector of the word (see Figure 5). As a summary statistic, we defined the proximity r_{wm} of the actual-meaning vector, $\hat{\mathbf{s}}_w$, obtained with input word w to the stored-meaning vector, $\mathbf{s}_m$, for each word m as the cosine of the angle between the actual and the stored-meaning vectors in the 68-dimensional space of the sememes:[7]

$$r_{wm} = \frac{\hat{\mathbf{s}}_w \cdot \mathbf{s}_m}{\|\hat{\mathbf{s}}_w\| \cdot \|\mathbf{s}_m\|}. \tag{8}$$

In keeping with the general principles of probabilistic feature models (Smith & Medin, 1981), we assumed that r for some target word does not have to be 1 for satisfactory semantic access to be achieved. What value should r then take for it to be accepted that access to the semantic representation of some target word has occurred? A plausible lower bound can be obtained from the median proximity between a word and its nearest neighbor in semantic space; this is 0.76. Given the geometrical properties of 68-dimensional spaces, the a priori probability of obtaining a proximity greater than r declines very rapidly as r moves toward 1. For an initial comparison between the different conditions, we adopted a threshold value of $r = 0.8$.[8]

---

[6] A small alpha value was used for the initial phase because the error surface contains a big initial ravine. The error surface slopes down steeply until the weights have reached values that yield an optimal guessing strategy for the sememes (ignoring the graphemic input). The network then settles to the same activity pattern over the sememes regardless of the input vector. If an alpha value near 1 is used in this initial phase, the network may drive some of the weights to a very large positive or negative value and may take a very long time to recover from this. This behavior is often mistakenly diagnosed as indicating a local minimum.

[7] This proximity measure was chosen because a unit in a connectionist net computes a total input that is a scalar product of incoming activities with weights. So, two incoming activity vectors that have a cosine near 1 will tend to have similar effects on any recipient unit. By comparison with a Euclidean distance measure, proximity is more sensitive to changes toward other possible stored-meaning vectors rather than ones that just generally reduce sensitivity.

[8] The stored-meaning vectors correspond to particular vertices of a 68-dimensional hypercube. Each of these vertices has between 12 and

Where a nonunitary value of r is used, the system needs to be capable of discriminating between the correct meaning and other meanings that also have high proximity to the actual vector; otherwise, it would not be able to drive a plausible output system effectively. We therefore added an additional criterion—the gap criterion—that the proximity between the actual meaning vector and that of the closest meaning must be at least 0.05 greater than the proximity of the actual vector and that of the next closest target. Later in this section, we consider to what extent our conclusions depend on the choice of these criteria.

Overall Effects

Using these criteria, we examined the absolute levels of performance of lesioned systems, the quantity and nature of errors, and their consistency over multiple trials. This was carried out for a wide range of lesions for Network A and a rather more restricted range for Network B. Tables 2 and 3 show how the probability of correct identification varies with lesion parameters when the threshold criteria of 0.8 and 0.05 are used. It is apparent that lesions affecting the input to the semantic system have greater effect than more distant lesions; because there are more *isconns* than *giconns*, this cannot be due to greater information carried per connection. In addition, disconnections in the cleanup circuit (those connections involving the *cleanup* system) have less effect than disconnections in the direct route, but the cleanup circuit is just as sensitive as the direct route to added noise. In other words, removing some of the cleanup effect is much less disruptive than adding erroneous cleanup. The effect of ablating *intermediate* units or *cleanup* units is equivalent to disconnecting the same proportion of their connections to the semantic system (*isconns* and *csconns*, respectively).

The effects of lesions on the two networks are qualitatively very similar. There is, however, a quantitative difference, with Network B being absolutely about 0.05 to 0.15 the less impaired by a lesion, except for lesions to *csconns*, where the mean difference is about 0.2. In general, the networks behave in a slightly different quantitative fashion but similar qualitative fashion. Therefore, the results for Network B are reported only where qualitative differences exist.

Errors

Noncorrect responses were divided into *omissions*, where one or more of the criteria are not satisfied for the closest target, and *errors*, where the criteria are both satisfied but with respect to some other word meaning. Errors were in turn divided into four types: semantic (S) errors; words semantically similar to the target but not visually similar; visual (V) errors, words visually similar to the target but not semantically similar; mixed (M) errors, words both semantically and visually similar to the target; and other (O) errors.

For simplicity, only responses that were words in the same semantic category were treated as semantic errors, and words with at least one letter in common in the same position in the word were considered visual errors. The use of this criterion for semantic errors will on rare occasions exclude a response that is semantically fairly close to a target (e.g., *mug* → "pop") and so reduce the number of observed semantic errors. It will become clear that this is not a problem.

The most obvious result of the error analysis is that all types of error occur with all types of lesion (see Table 4).[9] There is one exception—disconnecting the *scconns*—that produces very few errors. The likelihood of the observed error types' occurring by chance can be assessed by comparing the incidence rates with that of the other errors. In all cases, the incidence of a given type of error is a number of times greater than would be expected by chance. For lesion sites other than *scconns*, the ratio of semantic to other errors is at least 8 times the chance value; for the ratio of mixed to other errors, it is at least 36 times; and for the ratio of visual to other, at least 3 times the chance value.

In addition, assuming independence, the expected rate of mixed errors M can be predicted from the rates V and S of the visual and the semantic errors. Shallice and McGill (1978) showed that the following relation holds:

$$M < \frac{s}{1-s} V + \frac{v}{1-v} S, \tag{9}$$

where v and s are the a priori probabilities that a randomly selected input–output pair would be considered to be visually and semantically similar, respectively. By this formula, the incidence of mixed visual and semantic errors is higher than would be expected if visual and semantic errors arose independently for almost all lesion sites.[10]

It is possible that the comparison among the effects of different sites for lesions might be complicated by an effect of lesion severity on type of error in certain cases. Even if this were so, the proportion of different types of error varies for different lesion sites. Consider networks that have disconnections in *giconns* or *isconns*. Their ratio of semantic errors to visual errors differs significantly (a) if one matches for degree of lesion size

21 positive coordinate values of 1, with the remaining coordinate values being 0. The expected number of coordinates with a value of 1 is 15.2. After mild or moderate lesions, the actual-meaning vectors remain close to vertices of the hypercube (see Figure 5). A suitable approximation to the a priori distribution of proximity values can therefore be obtained by considering the proximities between vertices of the hypercube randomly selected to have a number of positive coordinates which lies within a certain range. To obtain an upper bound estimate of the a priori proximities for high values of r, consider the distribution of proximities between vertices that have the same number of positive coordinates (15). If two such vertices are selected by chance, then proximity depends on the number of positive coordinates in common; this is given by the hypergeometric distribution. The probability that there will be 11 positive dimensions in common (proximity = 0.73) is 0.9×10^{-6}. The probability that they have 12 or more positive dimensions in common (proximity = 0.8) is 0.24×10^{-7}, that is, less than 0.03 of the former value. Thus, a small increase in proximity leads to a very large decrease in the probability of a value that high being obtained by chance.

[9] This does not apply for visual errors occurring with some lesions to the cleanup circuit in Network B. However, the high rates of mixed errors for cleanup circuit lesions in Network B—generally more than 50%—indicate that for Network B, too, graphemic similarity is having an effect at the part of the system most distant from the graphemic input.

[10] This does not apply to *cleanup* ablations in Network B, but too few lesions were made to obtain a sufficiently large corpus of errors.

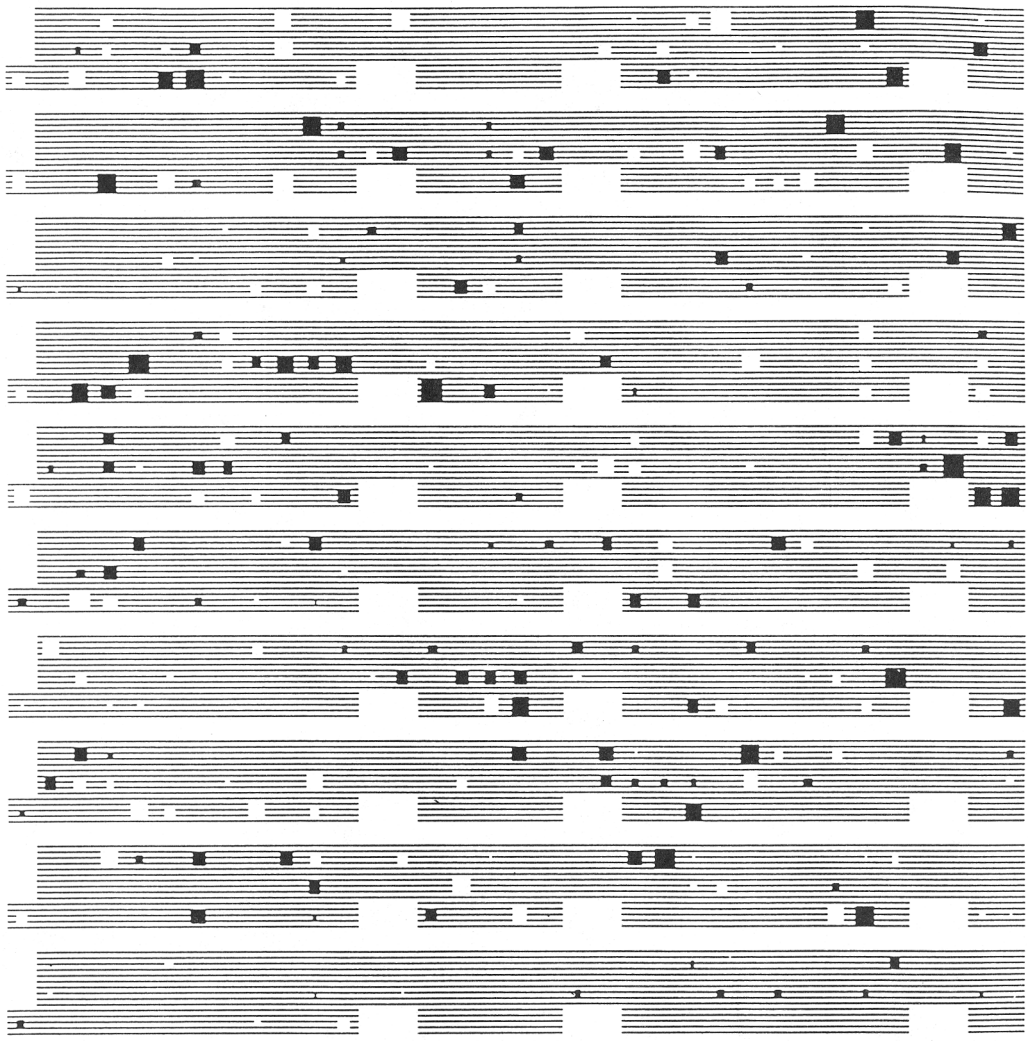

Figure 4. The weights on the incoming and outgoing connections of the first 10 intermediate units. (The white blobs indicate positive weights, and the black blobs negative weights, with the area of the blob representing the magnitude of the weight. The bottom row of each of the 10 panels represents the weights on the incoming connections from the four groups of graphemic units. The top two rows of each panel represent the weights on the connections to the 68 sememe units. The large inhibitory weight near the right of the top row of the top panel has a magnitude of −5.38. The bias of each unit is represented by the leftmost weight in the bottom row of each panel.)

(combining 0.1, 0.15, 0.2, and 0.25), $\chi^2(1, N = 61) = 10.71$, $p < .01$, or (b) if one matches for a rate of correct responses: combining *disconnect*(*giconns*, 0.2, 0.25, and 0.3), mean correct = 43.7%, and *disconnect*(*isconns*, 0.1 and 0.15), mean correct = 43%, $\chi^2(1, N = 52) = 13.7$, $p < .01$. Similar effects occurred when noise was added and also in Network B. Apparently lesions occurring earlier in the primary circuit are more prone to give visual rather than semantic errors when compared with lesions later in that circuit.

The Criteria

The quantitative values given in Tables 2, 3, and 4 are all dependent on the choice of criterion. The effect of making two particular lesions was therefore examined in detail to assess how changing the criteria would affect the results. The two chosen were *disconnect*(*giconns*, 0.3) and *disconnect*(*isconns*, 0.15), in that they gave comparable percentage correct results (35.3% and 36.5%) and were the pair of disconnections that produced the greatest contrast in error type.

Figure 6 shows for *disconnect*(*giconns*, 0.3) that there is a wide spread for the values of the proximity and gap when the correct word is the closest target. Proximity values range from about 0.5 to 0.99, and gaps range from 0 to about 0.4. The most critical point is that where an incorrect word is the closest stored-meaning vector, there is a similar range in both variables. Thus, the *disconnect*(*giconns*, 0.3) contains a visual error with proximity of 0.95 and gap of 0.42, and the *disconnect*(*isconns*, 0.15) contains a semantic error of proximity 0.96. That the error candidate distributions have similar ranges to the correct candidate distribution means that whatever values, within reason, are chosen for the two criteria, errors of at least

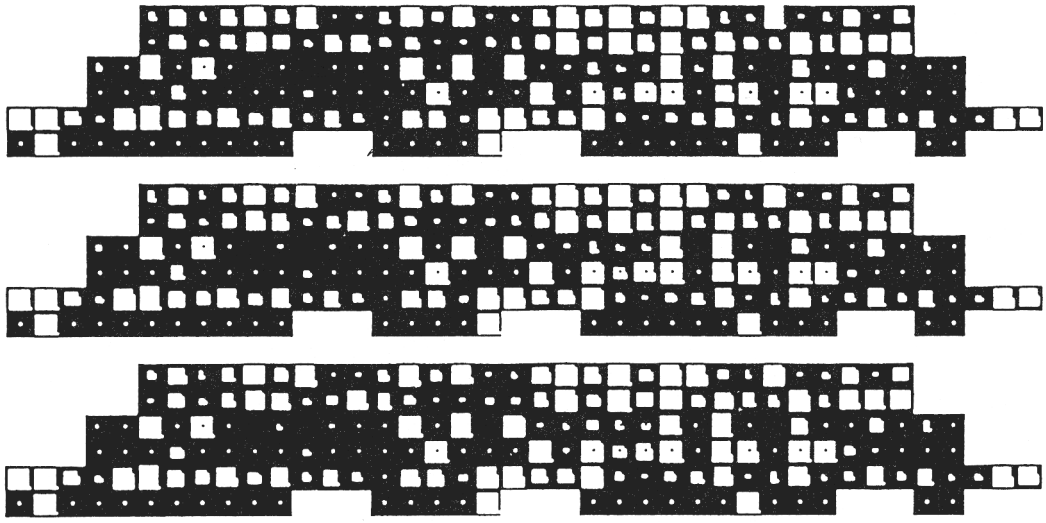

Figure 5. The activation values of all units in the network for the critical last three time slices for the lesioned network *disconnect*(*giconns*, 0.3), where 30% of the connections between grapheme and intermediate units were set to 0, with the input word *cup*. (The graphic conventions used are the same as for Figure 3.)

these two types will be observed. Clearly, the number of errors made will change as the thresholds vary, but the actual existence of errors of these two types will not.[11]

Below-Threshold Information

A second consequence of the broad range of proximity and gap values attained when the correct word is the best candidate is that there will be many trials when the closest target is the correct word but insufficient information is available to drive the response system. As pointed out earlier, the proximity criterion can hardly be placed below 0.76, which is the average proximity of a word to its nearest neighbor. Yet, many of the correct best candidate values achieved are below this level.

To assess on what percentage of trials there is useful below-threshold information available to the system, two types of tests were carried out for trials on which an explicit response would not be made. The first was five-alternative, between-categories forced choice. The proximities of the obtained value to the centroids in semantic space of each of the five categories were compared, and the closest category was chosen. The second was an eight-alternative, within-category forced choice. The closest of the eight category members was selected. Unfortunately, during the simulation, we only saved information about the proximity of the output to the six closest targets and to the centroids of the categories, and we only saved this information for targets with a proximity closer than 0.4. This means that occasionally, when all category members were far from the obtained value, no information was available as to which was the closest in the within-category test. In this case, each possible response was chosen on 12.5% of occasions.

Performance on these forced-choice tests was assessed for several types of disconnection for all trials on which the two criteria were not achieved. Lesions to the cleanup circuit led to high levels of performance on both forced-choice measures. A complete lesioning of *scconns*, which depresses performance to 40% correct on the standard criteria, gave 91.7% correct for the five-choice between-categories test and 87.5% correct for the eight-choice within-category measure for the 60% of words that produced below-threshold output. A 0.4 diconnection of *csconns*, which reduced explicit correct response to 24.5%, gave a performance on a between-categories test of 73.8% and a performance on a within-category test of 73.4% for the 75.5% of below-threshold trials.[12] Lesions to the primary pathways also showed the above-chance preservation of forced-choice responding in below-threshold situations. Thus, *disconnect*(*isconns*, 0.15), with correct explicit responding of 36.5%, gave a below-threshold performance of 62.6% and 64.1% on the two forced-choice measures. The effect was weaker for *disconnect*(*giconns*, 0.3), which gave a roughly equivalent correct explicit response performance (35.3%); the scores on the two forced-choice measures for below-threshold trials were 48.3% and 49.0%.

One might argue that a lower setting of either of the two threshold criteria would result in the elimination of the above-chance performance on below-threshold trials on the two forced-choice measures. However, for *disconnect*(*giconns*, 0.3), changing these criteria did not eliminate the effect. For example, taking the highest below-threshold value of proximity to be 0.76 (the median proximity between a stored-meaning

[11] Strictly, one needs to consider whether all types of error have ranges similar to the correct responses for each lesion site. The lesion types considered in most detail for errors were *disconnect*(*giconns*, 0.3) and *disconnect*(*isconns*, 0.2), for which 10 additional trials were run for each word. For the former, extended coverage included a semantic error with a proximity of 0.97 and a gap of 0.11. For the latter, it included a visual error of proximity 0.92 and a gap of 0.29 to add to a semantic error with proximity 0.90 and gap 0.15 in the original 10 trials.

[12] In all cases, correct responses were also correct on the between-categories measure.

Table 2
Percentage Correct Performance With Standard Criteria

Connection	Disconnection severity									Addition of noise to connections									
	0.05	0.1	0.15	0.2	0.25	0.3	0.4	0.5	0.7	1.0	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	6.0
Network A																			
giconns	—	82.0	65.8	54.5	41.3	35.3	18.0	—	—	—	—	86.0	70.8	53.5	38.8	33.8	26.5	19.5	—
isconns	81.0	49.3	36.5	21.5	9.3	—	—	—	—	—	73.5	52.8	30.3	19.3	10.3	—	—	—	—
scconns	—	99.5	99.0	99.3	96.3	95.8	92.5	84.5	70.8	40.0	—	—	90.3	—	63.3	—	50.0	—	39.5
csconns	—	92.0	78.0	59.8	46.3	43.4	24.5	13.0	—	—	81.5	60.5	26.8	14.0	5.8	—	—	—	—
Network B																			
giconns	—	—	—	60.5	45.3	38.3	20.3	—	—	—	—	—	79.8	68.5	50.3	—	34.8	—	—
isconns	—	54.0	44.3	24.5	—	—	—	—	—	—	—	65.8	38.5	—	—	—	—	—	—
scconns	—	—	—	—	—	—	—	—	76.0	50.0	—	—	—	—	68.5	—	62.0	—	41.5
csconns	—	—	—	81.5	—	62.8	51.5	26.0	—	—	—	67.8	47.3	—	—	—	—	—	—

Note. Connections are coded to match the first letter of each type of unit: *g* = grapheme; *i* = intermediate; *s* = sememe; *c* = cleanup. Dashes indicate conditions that were not tried.

vector and its closest neighbor) and setting the gap to 0 still gave below-threshold performance of 41.0% and 48.5% on the two forced-choice measures. Both these measures are significantly above chance (20% and 12.5%, respectively, $\chi^2(1, N = 161) = 44.3$ and $\chi^2(1, N = 112) = 274.6$, respectively, $p < .001$ in both cases. All these effects also hold for Network B.

The most surprising aspect of these results is that in all cases, the eight-alternative forced-choice within-category performance is as high as the apparently much easier five-alternative between-categories performance. Except for *scconns*, this cannot be explained by the correlation between the two measures of performance across trials. Thus, for *disconnect*(*giconns*, 0.3), there is a positive contingency (C) between the measures, but it is not that high ($C = 0.33$).

Category Effects

It is possible to investigate how much the processing of individual words is affected when lesions are made to different networks and to different sites within the same network. The effects obtained are complex and in general not of apparent theoretical interest.[13] However, we obtained a very unexpected finding about the way in which individual categories of words were affected by certain lesions. After *csconns* lesions, one category, foods, was selectively preserved in Network A (see Figure 7), and this category difference was significant when words were treated as a random variable, Kruskal-Wallis $H(4) = 13.38$, $p < 0.1$, for *disconnect*(*csconns*) lesions. This effect was highly specific. It did not, for instance, occur for *disconnect*(*giconns*) lesions for Network A, nor for *disconnect*(*csconns*) lesions for Network B.

Discussion

A major attraction of connectionist models is that they perform computations by means of processing units that behave on broadly similar principles to those used in the brain. Their appeal would be strengthened if it could be shown that analogous operations on the elements of the two systems lead to similar consequences for the behavior of the overall systems. An obvious candidate operation is the effect of lesions. In this article, we take a relatively simple connectionist model and show that the lesioned model has properties that are similar to those already described in certain neurological patients.

The domain we chose—reading-to-meaning—has been much studied in single-case studies over the last 20 years. The network maps letter-level information into semantic information. Three principal qualitative phonomena are shown by the model:

1. It produces both more semantic and more visual errors than would be expected by chance. It also produces more mixed errors than would be expected if semantic and visual errors were independent. In more abstract terms, the errors reflect the similarity metric of both the input and the output.

[13] For further information on this issue, on the effect of lesion density on error type, on the consistency of performance across trials, and on an attempt to simulate lexical decision, see Hinton and Shallice (1989).

Table 3
Percentage Correct Performance With Standard Criteria for Ablating Units

Units ablated	% lesioned	Network A	Network B
Intermediate			
3	7.5	56.5	71.8
4	10.0	49.3	59.0
5	12.5	32.0	36.8
7	17.5	23.3	—
10	25.0	16.8	—
Cleanup			
10	16.7	71.0	81.8
20	33.3	38.3	56.3
30	50.0	18.8	—
40	66.7	4.8	—

2. The same combination of error types occurs whatever set of connections or hidden units are lesioned and whichever of the two types of lesioning procedure is used (with the exception of one set of connections, involving the cleanup units).
3. If sufficiently severe lesions are made so that an explicit response can no longer be produced, the network still performs above chance in forced-choice situations.

Two key questions can be posed. First, what aspects of the simulation are critical for these effects to occur? Second, do they correspond to the behavior shown in patients?

Design of the Network

To simulate the effects of damage, it is necessary to use a particular network with particular groups of units, particular patterns of connections, and particular input and output encodings. These requirements immediately raise the issue of whether the architecture and the input and output encodings are too simple or too complex to be plausible and whether the qualitative results are sensitive to fine details of the architecture. It is impossible to try even a small fraction of all possible architectures, and given the slowness of the learning, we could not experiment with alternatives. We therefore chose a particular architecture that was relatively simple but contained the major ingredients that we would expect to be present in the brain. Our design decisions and our reasons for them are as follows:

1. The input representation is simple. Even though the brain may not contain neurons that represent the conjunction of a grapheme and its position within a word, it probably uses a representation in which similar strings of graphemes are encoded by similar vectors, and this is probably the only property of the input encoding that is crucial for the qualitative results.
2. The output representation is also simple, and the same justification applies as for the input representation.
3. There must be at least one layer of hidden units between the input and output. Given our input encoding (or any other encoding of comparable size), we can encode any four-letter word as a pattern of activity over 4×26 input units. So if there are more than 104 four-letter words, it is impossible for the input vectors to be linearly independent, and it is therefore impossible to associate arbitrary output vectors with them by a linear associator. With nonrandom words, the problem can be even more severe. With our input encoding, no linear associator could associate a semantic feature with the words GOAT and COAL and not associate that feature with the words GOAL and COAT.
4. There must be recurrent connections between the output units or between them and other groups of units to create the attractors that are a central postulate of our theory. It would be possible to use only pairwise interactions among the output units, but this strongly limits the kinds of attractors that can be created. It would be possible to avoid this limitation by using recurrent connections involving the hidden units, in which case there would be no architectural distinction between cleanup units and intermediate units. We felt that it was more plausible to separate units used for mapping from the graphemes to the sememes from units used for building complex attractors. Given that the cleanup units are modeling the rapid part of the effects of the whole of the rest of the brain on the sememe units,

Table 4
Error Rates in the Different Conditions

	Overall error rate		Conditional probabilities			
Condition	*n*	Rate	Semantic	Visual & semantic	Visual	Other
Disconnect (*giconns*)	4	4.8	13.2	44.7	34.2	7.9
Noise (*giconns*)	4	3.9	20.6	27.0	46.0	6.3
Ablate (*intermediate*)	3	3.1	24.3	45.9	24.3	5.4
Disconnect (*isconns*)	2	3.4	55.6	29.6	11.1	3.7
Noise (*isconns*)	3	2.4	20.7	48.3	24.1	6.9
Disconnect (*scconns*)	2	0.23	—	100	—	—
Noise (*scconns*)	4	1.8	20.7	72.4	6.9	—
Ablate (*cleanup*)	2	3.4	25.9	63.0	7.4	3.7
Disconnect (*sconns*)	3	3.4	34.1	31.7	34.1	—
Noise (*csconns*)	2	2.3	33.3	38.9	27.8	—
Chance			11.9	6.1	29.6	52.5

Note. The results are amalgamated over runs that used values of the parameters that yielded an overall correct performance of between 25% and 75%. *n* = the number of sets of values from which the results are derived. *giconns* = connections between *grapheme* and *intermediate* units; *isconns* = connections between *intermediate* and *sememe* units; *scconns* = connections between *sememe* and *cleanup* units; *csconns* = connections between *cleanup* and *sememe* units. Dashes indicate conditions that were not tried.

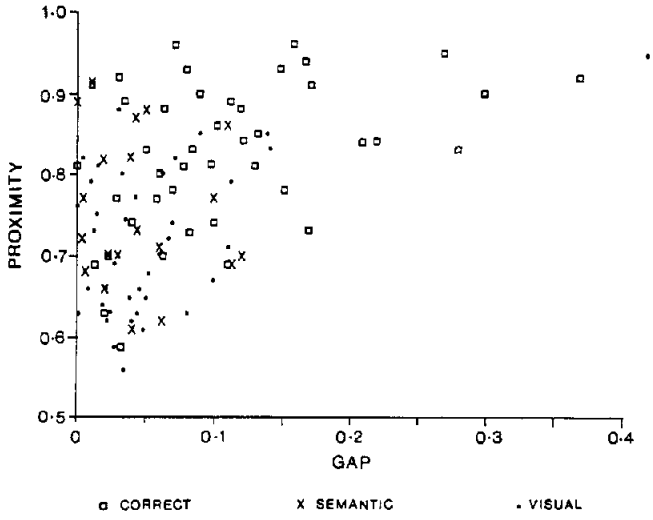

Figure 6. The values of proximity and gap from 400 trials with *disconnect*(*giconns*, 0.3), where 30% of the connections between grapheme and intermediate units were set to 0. (All responses where either a semantic error or a visual error was the closest stored-meaning vector are shown together with a randomly selected 20% of trials where the correct word was the closest stored-meaning vector.)

we view them as a gross simplification of much more complex interactions rather than as an unjustified elaboration.

The general design of the network was not guided by a desire to produce the phenomena of dyslexia. The network from which the present one was derived was actually designed as an illustration of how a neural net could efficiently represent an arbitrary mapping from graphemes to sememes without using a local unit for each word (Hinton, McClelland, & Rumelhart, 1986). Only after the network was designed and analyzed did it occur to us that it would give one of the phenomena of dyslexia (semantic errors). Even after we had run the present simulations, we were surprised that the network also exhibited other phenomena, such as the strong tendency to give mixed errors and the tendency for damage near the semantic representation to lead to visual errors.

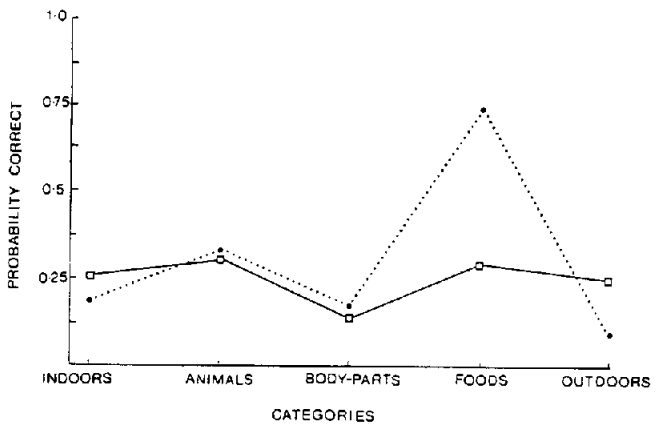

Figure 7. The effects of two different types of lesion on the performance of Network A on the five categories of word. [The solid line is for *disconnect*(*giconns*, 0.4), and the dotted line is for *disconnect*(*csconns*, 0.4), where 40% of the connections between the *grapheme* and *intermediate* units and the *cleanup* and *sememe* units, respectively, were set to 0.]

The network is not intended to provide a precise simulation of a particular class of dyslexic disorder. For technical reasons, four types of simplification were made, meaning that an attempt to produce a quantitative fit to empirical results would not be appropriate:

1. The words used are not a representative subset of the English language; important variables, such as word length, word frequency, syntactic class, and word imageability known to affect performance in acquired dyslexia, have not been manipulated, and the treatment of the semantic level—in terms related to probabilistic feature models—can at best provide only a crude approximation to the complex semantic representations of concrete nouns.
2. The selection of semantic features and of the quantitative values of numbers of units and of connections was not based on any empirical evidence. There is, however, some evidence that the qualitative results do not depend on many of the quantitative details, because we obtained similar results for an earlier network that was topologically broadly similar but that differed substantially in quantitative terms.
3. The modeling of performance on specific tasks is based on generally unsupported additional assumptions; in the most critical cases, we have, however, tried to provide secondary support for our conclusions independent of our additional assumptions.
4. In the disconnection or ablation mode, our procedure was to sum over 10 random samples with different specific connections or units of the same general type lesioned in each sample. This can only provide an approximation to the results of an individual patient for which a much larger set of words is subject to the effect of a single specific set of connections or units lesioned.

Fortunately, these simplifications are not critical, because our aim is to demonstrate qualitatively similar effects to those obtained with acquired-dyslexic patients. Given the assumptions present in connectionist modeling, there are natural ways to represent the effects of lesions within the framework. The plausibility of the approach is increased if interesting neuropsychological phenomena are exhibited in analogue by the model and is diminished if they are not found.

One might argue that the type of modeling we have carried out of the effects of neurological damage is unconstrained and that the model can be delicately manipulated to produce any phenomenon. In fact, however, the present investigation was only the second network of its type we investigated. Its parameters were selected on a priori computational grounds, not by a curve-fitting exercise, and the first network showed generally similar behaviors.

Modeling has a complementary function for neuropsychology. By lesioning a complex network that is plausible on independent grounds, one obtains a richer account of the relation between the functional lesion and observed behavior than is available when neuropsychological theorizing is based on the functional architecture alone.[14] More particularly, it shows how

[14] We are not claiming that connectionism provides the only procedure for obtaining such an account. Any fully specified mechanism would be equivalent in this respect. However, theories of neuropsychological performance based on examining damage to fully specified mechanisms are very few.

the same qualitative pattern of errors can arise from functionally different lesions and how phenomena can arise in a damaged system that are far from transparent (Gregory, 1961). Consider, for instance, the category effect. On a specific level, the investigation provides a potential explanation for two types of counterintuitive phenomena that have been observed in the study of the acquired dyslexias.

Neuropsychological Correspondences

The first set of phenomena for which the network suggests an explanation are those associated with the symptom-complex deep dyslexia. The simplest is the semantic error itself. The network was designed with semantic features for output so that the explanation it provides for these errors is a special case of the accounts of Howard (1985) and Hillis, Rapp, Romani, and Caramazza (1990) of inaccuracy in accessing or using such features. Their approach was in turn derived from explanations based on Katz and Fodor's (1963) semantic theory, such as those of Marshall and Newcombe (1966) and Coltheart (1980b). The additional element that our explanation offers is the presence of a mechanism—in essence, the existence of attractors—resulting in the accessing of clusters of features that correspond closely to the meanings of words related to the target word and the demonstration that such inaccurate access can follow from both partial ablation and partial disconnection.

More critically, the network suggests an explanation for the co-occurrence of different types of error in a wide range of patients who read by the semantic route, in particular the co-occurrence of both semantic errors and visual errors. Coltheart et al. (1987) in a recent review posed the question of why acquired-dyslexic patients who produce semantic errors have always been found to produce visual errors, but the authors were unable to provide a compelling answer. Although more recently, 3 patients have been described for whom the generalization is not correct, more than 30 patients for whom it is valid have been described. A third effect is that where the relative incidence of mixed errors (e.g., *can* → "pan") has been examined, a higher incidence than would be expected from the combined rates of visual errors and of semantic errors has indeed been found (e.g., Shallice & McGill, 1978).[15]

Our investigation offers a simple answer to the mix of different error types so frequently observed. In the network, the representation of words on levels between a letter level and the semantic level has both distributed and cascade properties; that both of these characteristics are appropriate is supported by evidence from normal subjects (e.g., Johnston & McClelland, 1980; Rumelhart & McClelland, 1982). In the present concrete realization of such a model, lesions anywhere in the system—except *disconnect*(*scconns*)—give rise to qualitatively the same error pattern. Thus, a straightforward explanation exists for why this error pattern is so widespread even though other aspects of the syndrome vary across patients. It reflects a breakdown characteristic of a network containing attractors when lesioned in various places. Such a mixture of error types may be as much a sign of the operation of a layered connectionist system with attractors as dissociations are of modular systems.[16]

Two objections may be made to this analysis. Certain patients who make semantic errors in reading aloud but who make hardly any visual errors have been described recently (Caramazza & Hillis, 1990; Hillis et al., 1990). Caramazza and Hillis argued that for 2 of the patients, the locus of the impairment is the phonological output lexicon. This would place it outside the domain of this model, which deals only with the process of accessing semantic representations. However, 1 patient, K.E, produced qualitatively similar semantic errors regardless of the form in which the information was presented—from written words to tactually presented objects—and regardless of the verbal output procedure used (writing or speaking). He produced virtually no visual errors. Hillis and her colleagues attributed the patient's semantic errors to the loss of certain semantic features required by whatever transformation was being carried out. Now, in our simulation, the effect of lesioning the sememes themselves could not be examined, because the difference between their activation and the desired level provided the error measure. Such a lesion would, however, have two indirect effects. First, it would remove input to other sememes through the direct sememe–sememe connections. As a consequence of the generally winner-take-all character of these interactions, this would tend to release inhibition from competing sememes. This would be a powerful effect, and the consequence would tend to be semantic errors. The second effect would be to reduce the input to the *cleanup* units. The closest tested analog was the disconnection of *scconns*. This disconnection, though, leads to an order of magnitude less errors than the other types of disconnection (see Table 4). So, if the two effects combine in a reasonably linear fashion, then the effect of a lesion to the semantic units would be to produce semantic errors with few, if any, visual ones. This, however, remains a speculation until simulated in a network, which would have to be more extensive then the present one.

A second argument that needs to be considered is that put forward by Gordon et al. (1987). The position they adopted for their deep-dyslexic patient FM was that visual errors arise from damage to an orthographic lexicon and semantic errors arise from damage at the semantic stage (see also Patterson, 1978). They did not discuss mixed errors. They tested their account in

---

[15] Absolute numbers of mixed errors comparable with those of the visual errors and the semantic errors are not found in patients. However, the relative proportions will depend on the proportions of response types that are counted as visual errors and semantic errors (see the discussion relating to Equation 9). For visual errors, this chance value is much higher in the case of the model than in the way responses are normally scored for patients.

[16] The ratio of semantic to visual errors varies considerably across patients. For example, Patient GR (Marshall & Newcombe, 1966) produced more than twice as many semantic as visual errors, but Patient K.F (Shallice & Warrington, 1975) produced 15 times as many visual errors as semantic errors. In the model, disconnections to the *giconns* produce a much higher ratio of visual to semantic errors than do more centrally located lesions. The factor of 13 for this ratio (see Table 4) is not qualitatively much different from the changes that occur among patients. Patients have been described who make only visual errors. However, in certain cases in the literature, as for neglect dyslexia (Costello & Warrington, 1987; Ellis, Flude, & Young, 1987), it is plausible that the impairment lies at or before the stage at which letters are categorized, and in others, as for phonological alexia (Derouesné & Beauvois, 1985), the phonological route or routes may be partially operative so that any putative semantic error will tend to be edited out (see Newcombe & Marshall, 1980b).

two main ways. One concerned how well words that were produced as visual error responses and as semantic error responses were read. The empirical effects they obtained were small and not completely consistent with their theoretical predictions. Much stronger effects were obtained to support a second prediction they made from their theory that words that give rise to a particular class of errors do so consistently; if another error is made later on the same word, it will tend to be of the same type. If different types of error occur from independent loci, however, this does not explain the phenomenon described in the introduction of visual errors occurring more frequently on words in semantic classes the patient has difficulty in processing. As Gordon et al. pointed out, to explain this phenomenon it seems necessary to assume the existence of a cascade type of process. Yet, as Appendix C explains, on our model, which has cascade-type characteristics, a single lesion to *isconns* can lead to a consistent pattern of errors having both visual and semantic aspects. Thus, the existence of a consistent pattern of errors containing both semantic and visual errors does not require one to assume the existence of two lesions in a patient that produces the pattern. Thus, an explanation based on the assumption that the co-occurrence of the two types can arise from a single lesion cannot be excluded by the observation.

Our model exhibits another type of phenomenon for which patients with other types of dyslexia are more relevant. These effects were originally described in a semantic-access dyslexic patient, AR, who was unable to read a word aloud or even identify it (e.g., by picture–word matching) but could make either between-categories or within-category judgments at well above chance level (Warrington & Shallice, 1979). When words could be read aloud or identified at only about 40% correct, five-choice between-categories performance for unidentified words was 69% correct. For an equivalent level of correct responding, *disconnect(giconns)* and *disconnect(isconns)* gave five-choice between-categories performances of 64% and 48% on trials where the word was not identified, values that are clearly in the same range.

The idea that the preserved aspects of reading in semantic-access dyslexia might arise from activation of a subset of the features of a word's semantic representation has been suggested by Howard (1985) and Rapp and Caramazza (1989). The model shows that a computational explanation of this general type can predict performance in the appropriate quantitative range. More critically, it provides a surprising finding that, on trials where no explicit response could be made, the eight-choice between-categories judgments tended to be performed at roughly the same level as the five-choice within-category ones, and both were well above chance. The model therefore makes the counterintuitive prediction that forced-choice within-category judgements are at least as good as selection between superordinate labels. Patient AR was not tested on this latter form of test directly. However, Rapp and Caramazza (1989) showed that in a patient showing the same general phenomenon, the semantic distance between the alternatives was not a significant factor in performance on two-alternative word–picture matching.

Rather similar effects observed in the pure alexic patients of Shallice and Saffran (1986) and Coslett and Saffran (1989) are also relevant. When a word is presented to these patients for too brief a time for the letter-by-letter procedure to be used, explicit word identification can only rarely be achieved, but between-categories discrimination can be carried out at a level that is well above chance. Again, on the few occasions when explicit responses are made, both semantic errors and visual errors occur. One appears to have an extreme form of semantic-access dyslexia; it would fit with how the model behaves with a severe lesion to, say, *giconns* or to the *intermediate* units. Most strikingly, using a word-to-picture matching test, Coslett and Saffran have found performance as good in a within-category test as in a between-categories one—strikingly similar, although not identical to the phenomena generated by the model with, say, a severe lesion to *giconns*.

There was an additional noteworthy result obtained from the simulation. One surprising failure to replicate an aspect of the lesioned system's behavior occurred with a seemingly noncritical change in parameters. The network was trained again with a new set of initial weights to check that the effects obtained were independent of the small, random initial values chosen for the weights.[17] Qualitatively, the same overall pattern was obtained, with one exception.

Network A exhibited category specificity in that one particular category, foods, was much better preserved when one particular lesion, *disconnect(csconns)*, was made. Category-specific effects have now been found in many neurological patients and for a wide set of categories, most of the findings coming from individual case studies (see Shallice, 1988, for review).[18] Several different types of explanation have been put forward (e.g., see Humphreys et al., 1988; Warrington & McCarthy, 1983, 1987; Warrington & Shallice, 1984). To our surprise, our modeling of the category-specificity effect obtained in Network A was completely absent in Network B, which basically differed from Network A only in the initial starting weights. This finding implies that the overall structure of the trained network may depend initially not only on the environment but also on the small initial weights used. This has the further consequence that any characterization of the effect in terms of the pattern of stimulus–response contingencies alone cannot be correct; thus, the explanation for the present example of category specificity must inevitably be more complex than explanations offered concerning the neurological examples of the phenomenon. Because the attempt to provide simple accounts of such phenomena in patients has encountered internal difficulties (e.g., see Hart, Berndt, & Caramazza, 1985; Warrington & McCarthy, 1987), the presence of a far-from-transparent example of category specificity in our simula-

[17] These random values are necessary in most networks to break symmetries among the hidden units. Because backpropagation is a deterministic algorithm, identically connected hidden units with identical initial weights can never become different.

[18] An example of category specificity relevant to the categories used in the simulation has been obtained with certain patients with herpes simplex encephalitis who have a much greater difficulty identifying, from verbal or visual presentation, living things and foods rather than artifacts (Sartori & Job, 1988; Silveri & Gainotti, 1988; Warrington & Shallice, 1984; see also McCarthy & Warrington, 1988). By contrast, certain global aphasic patients have exactly the opposite pattern of performance in picture–word matching (Warrington & McCarthy, 1983, 1987). Relatively limited information only is available on the reading of these patients.

tion suggests that neurological examples of the phenomena should be interpreted with caution.

Conclusions

In the present simulation, the lesioning of a connectionist model that maps orthographic inputs onto semantic features produces several counterintuitive behaviors that are also shown by acquired-dyslexic patients. The two main ones concerned the range of error types exhibited and the below-threshold forced-choice performance. Because the effects occurred at rather different densities of lesion in the model, we have in general treated the effects as relevant for different patients, although the semantic-access patients studied have shown both effects.

The simulations were intended to relate to only certain aspects of the reading of the dyslexic patients we considered.[19] Whether extrapolations from the basic approach advocated here can provide an account of other aspects of the syndrome must await more elaborate and specific simulations. However, there seem to be certain obvious lines of development. Thus, the effect of the sheer density of semantic features that a word possesses is a potentially important variable for predicting how well a word can still be processed after lesions are made; it would seem likely that different types of word differ markedly with respect to this variable (e.g., see Jones, 1985). The difficulty with abstract words and parts of speech other than nouns found in many dyslexic patients who cannot read by spelling-to-sound translation might well be explicable in related terms if they were modeled in an appropriate fashion.

The relatively preserved lexical decision performance found in some of the patients (Coslett & Saffran, 1989) may present more of a problem. Our initial attempt to simulate it was not successful. One possibility that needs to be considered is that the intermediate units may form attractors, so that the familiarity of a letter string can have an effect that does not just depend on the familiarity of the meaning.

Overall, a similarity exists, in the present domain, between the effects of lesions in a connectionist model and in certain types of neurological patient. Because the relevant phenomena are counterintuitive, this similarity strengthens the plausibility that the connectionist approach is capturing a key aspect of human cognitive processing. A central aspect of the model that enables it to produce the error phenomena is that it builds attractors to represent the meanings of words.

[19] There is no necessary conflict between an explanation of deep dyslexia in terms of lesions at varying points in a connectionist network and one in terms of right-hemisphere reading. The two explanations are orthogonal. If some part of a set of units—at one or more levels—were located in the right hemisphere, then a right-hemisphere reading system would correspond to the network quantitatively reduced in the appropriate fashion.

References

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

Anderson, J. A., Silverstein, J. W., Ritz, S. A., & Jones, R. S. (1977). Distinctive features, categorical perception, and probability learning: Some applications of a neural model. *Psychological Review, 84,* 413–451.

Barry, C. (1985). Consistency and types of semantic errors in a deep dyslexic patient. In R. N. Malatesha & H. A. Whitaker (Eds.), *Dyslexia: A global issue* (pp. 311–337). The Hague, The Netherlands, Martinus Nijhoff.

Barry, C., & Richardson, J. T. E. (1990). Accounts of oral reading in deep dyslexia. In H. A. Whitaker (Ed.), *Phonological processing and brain mechanisms* (pp. 118–171). New York: Springer.

Bub, D., Cancelliere, A., & Kertesz, A. (1985). Whole-word and analytic translation of spelling-to-sound in a non-semantic reader. In K. E. Patterson, M. Coltheart, & J. C. Marshall (Eds.), *Surface dyslexia* (pp. 15–34). London: Erlbaum.

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

Caramazza, A. (1986). On drawing inferences about the structure of normal cognitive systems from the analysis of patterns of impaired performance: The case for single-patient studies. *Brain and Cognition, 5,* 41–66.

Caramazza, A., & Hillis, P E. (1990). Where do semantic errors come from? *Cortex, 26,* 95–122.

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review, 82,* 407–428.

Coltheart, M. (1980a). Deep dyslexia: A review of the syndrome. In M. Coltheart, K. E. Patterson, & J. C. Marshall (Eds.), *Deep dyslexia* (pp. 22–47). London: Routledge & Kegan Paul.

Coltheart, M. (1980b). A right-hemisphere hypothesis. In M. Coltheart, K. E. Patterson, & J. C. Marshall (Eds.), *Deep dyslexia* (pp. 326–380). London: Routledge & Kegan Paul.

Coltheart, M. (1980c). The semantic error: Types and theories. In M. Coltheart, K. E. Patterson, & J. C. Marshall (Eds.), *Deep dyslexia* (pp. 146–159). London: Routledge & Kegan Paul.

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

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

Coltheart, M., Patterson, K. E., & Marshall, J. C. (1987). Deep dyslexia since 1980. In M. Coltheart, K. E. Patterson, & J. C. Marshall (Eds.), *Deep dyslexia* (2nd ed., pp. 407–451). London: Routledge & Kegan Paul.

Coslett, H. B., Rothi, L. G., & Heilman, K. M. (1985). Reading: Dissociation of lexical and phonological mechanisms. *Brain and Language, 24,* 20–35.

Coslett, H. B., & Saffran, E. M. (1989). Evidence for preserved reading in 'pure alexia.' *Brain, 112,* 327–359.

Costello, A., & Warrington, E. K. (1987). Dissociation of visuo-spatial neglect and neglect dyslexia. *Journal of Neurology, Neurosurgery and Psychiatry, 50,* 1110–1116.

Dejerine, J. (1892). Contribution àl'études anatomoclinique et clinique des differentes variétié de cécité verbale [A contribution to the anatomical and clinical study of the different varieties of word blindness]. *Memoires de la Société de Biologie, 4,* 61–90.

Derouesné, J., & Beauvois, M. F. (1985). The phonemic stage in the non-lexical reading process: Evidence from a case of phonological alexia. In K. E. Patterson, M. Coltheart, & J. C. Marshall (Eds.), *Surface dyslexia* (pp. 399–457). London: Erlbaum.

Derthick, M. (1987). Counterfactual reasoning with direct models. In *Proceedings of the Sixth National Conference on Artificial Intelligence* (pp. 346–351). Los Altos, CA: Morgan Kauffman.

Ellis, A. W., Flude, B. M., & Young, A. W. (1987). 'Neglect dyslexia' and

the early visual processing of letters in words and nonwords. *Cognitive Neuropsychology, 4,* 439–464.

Friedman, R. B., & Perlman, M. B. (1982). On the underlying causes of semantic paralexias in a patient with deep dyslexia. *Neuropsychologia, 20,* 559–568.

Geschwind, N. (1985). Mechanisms of change after brain lesions. *Annals of the New York Academy of Sciences, 457,* 1–11.

Gordon, B., Goodman-Schulman, R., & Caramazza, A. (1987). *Separating the stages of reading errors* (Tech. Rep. No. 28). Baltimore: Johns Hopkins University, Cognitive Neuropsychology Laboratory.

Gregory, R. L. (1961). The brain as an engineering problem. In W. H. Thorpe & O. L. Zangwill (Eds.), *Current problems in animal behaviour.* Cambridge, England: Cambridge University Press.

Hart, J., Berndt, R. S., & Caramazza, A. (1985). Category specific naming deficit following cerebral infarction. *Nature, 316,* 439–440.

Hillis, A. E., Rapp, B., Romani, C., & Caramazza, A. (1990). Selective impairments of semantics in lexical processing. *Cognitive Neuropsychology, 7,* 191–243.

Hinton, G. E. (1981). Implementing semantic networks in parallel hardware. In *Parallel models of associative memory* (pp. 161–187). Hillsdale, NJ: Erlbaum.

Hinton, G. E. (1986). Learning distributed representations of concepts. In *Proceedings of the Eighth Annual Conference of the Cognitive Science Society* (pp. 1–12). Hillsdale, NJ: Erlbaum.

Hinton, G. E. (1989). Connectionist learning procedures. *Artificial Intelligence, 40,* 185–234.

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

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

Hinton, G. E., & Shallice, T. (1989). *Lesioning a connectionist network: Investigations of acquired dyslexia* (Tech. Rep. No. CRG-TR-89-3). University of Toronto, Ontario, Canada.

Howard, D. (1985). *The semantic organisation of the lexicon: Evidence from aphasia.* Unpublished doctoral dissertation, University of London.

Humphreys, G. W., Riddoch, M. J., & Quinlan, P. T. (1988). Cascade processes in picture identification. *Cognitive Neuropsychology, 5,* 67–103.

Johnston, J. C., & McClelland, J. L. (1980). Experimental tests of a hierarchical model of word identification. *Journal of Verbal Learning and Verbal Behavior, 19,* 503–524.

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

Katz, J. L., & Fodor, J. A. (1963). The structure of a semantic theory. *Language, 39,* 170–210.

Kohonen, T. (1977). *Associative memory: A system-theoretical approach.* Berlin, Federal Republic of Germany: Springer.

Kremin, H. (1982). Alexia: Theory and research. In R. N. Malatesha & P. G. Aaron (Eds.), *Reading disorders: Varieties and treatments* (pp. 341–367). New York: Academic Press.

Landis, T., Regard, M., & Serrat, A. (1980). Iconic reading in a case of alexia without agraphia caused by a brain tumour: A tachistoscopic study. *Brain and Language, 11,* 43–53.

Marshall, J. C., & Newcombe, F. (1966). Syntactic and semantic errors in paralexia. *Journal of Psycholinguistic Research, 4,* 169–176.

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

McCarthy, R., & Warrington, E. K. (1986). Phonological reading: Phenomena and paradoxes. *Cortex, 22,* 359–380.

McCarthy, R., & Warrington, E. K. (1988). Evidence for modality-specific meaning systems in the brain. *Nature, 334,* 638–640.

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,* 375–407.

Minsky, M. L. (1975). A framework for representing knowledge. In P. H. Winston (Ed.), *The psychology of computer vision* (pp. 211–277). New York: McGraw-Hill.

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

Mozer, M. (1983). Letter migration in word perception. *Journal of Experimental Psychology: Human Perception and Performance, 9,* 531–546.

Newcombe, F., & Marshall, J. C. (1980a). Response monitoring and response blocking in deep dyslexia. In M. Coltheart, K. E. Patterson, & J. C. Marshall (Eds.), *Deep dyslexia* (pp. 160–175). London: Routledge & Kegan Paul.

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

Nolan, K. A., & Caramazza, A. (1982). Modality-independent impairments in word processing in a deep dyslexia patient. *Brain and Language, 16,* 237–264.

Patterson, K. E. (1978). Phonemic dyslexia: Errors of meaning and meaning of errors. *Quarterly Journal of Experimental Psychology, 30,* 587–601.

Patterson, K. E. (1979). What is right with 'deep' dyslexic patients? *Brain and Language, 8,* 111–129.

Patterson, K. E. (1981). Neuropsychological approaches to the study of reading. *British Journal of Psychology, 72,* 151–174.

Patterson, K. E., & Marcel, A. J. (1977). Aphasia, dyslexia and the phonological coding of written words. *Quarterly Journal of Experimental Psychology, 29,* 307–318.

Patterson, K. E., Seidenberg, M. S., & McClelland, J. L. (1989). Connections and disconnections: Acquired dyslexia in a computational model of reading. In R. G. M. Morris (Ed.), *Parallel distributed processing: Implications for psychology and neurobiology* (pp. 523–568). Oxford, England: Oxford University Press.

Rapp, B. C., & Caramazza, A. (1989). General to specific access to word meaning: A claim re-examined. *Cognitive Neuropsychology, 6,* 251–272.

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

Rumelhart, D. E., & McClelland, J. L. (1982). An interactive activation model of context effects in letter perception: Part 2: The contextual enhancement effect and some tests and extensions of the model. *Psychological Review, 89,* 60–94.

Sartori, G., Bruno, S., Serena, M., & Bardin, P. (1984). Deep dyslexia in a patient with crossed aphasia. *European Neurology, 23,* 95–99.

Sartori, G., & Job, R. (1988). The oyster with four legs: A neuropsychological study on the interaction of visual and semantic information. *Cognitive Neuropsychology, 5,* 105–132.

Schwartz, M. F. (1984). What the classical aphesia categories don't do for us and why. *Brain and Language, 21,* 3–8.

Schwartz, M. F., Saffran, E. M., & Marin, O. S. M. (1980). Fractionating the reading process in dementia. In M. Coltheart, K. E. Patterson, & J. C. Marshall (Eds.), *Deep dyslexia* (pp. 259–269). London: Routledge & Kegan Paul.

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. (1986). *NETtalk: A parallel network that learns to read aloud* (Tech. Rep. No. 86-01). Baltimore: Johns Hopkins University, Department of Electrical Engineering and Computer Science.

Shallice, T. (1988). *From neuropsychology to mental structure.* Cambridge, England: Cambridge University Press.

Shallice, T., & Coughlan, A. K. (1980). Modality specific word comprehension deficits in deep dyslexia. *Journal of Neurology, Neurosurgery and Psychiatry, 43,* 866–872.

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

Shallice, T., & Saffran, E. M. (1986). Lexical processing in the absence of explicit word identification: Evidence from a letter-by-letter reader. *Cognitive Neuropsychology, 3,* 429–458.

Shallice, T., & Warrington, E. K. (1975). Word recognition in a phonemic dyslexic patient. *Quarterly Journal of Experimental Psychology, 27,* 187–199.

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* (pp. 119–145). London: Routledge & Kegan Paul.

Shallice, T., Warrington, E. K., & McCarthy, R. (1983). Reading without semantics. *Quarterly Journal of Experimental Psychology, 35A,* 111–138.

Silveri, M. C., & Gainotti, G. (1988). Interaction between vision and language in category specific impairment for living things. *Cognitive Neuropsychology, 5,* 677–709.

Smith, E. E., & Medin, D. L. (1981). *Categories and concepts.* Cambridge, MA: Harvard University Press.

Van Orden, G. C. (1987). A ROWS is a ROSE: Spelling, sound and reading. *Memory & Cognition, 15,* 181–198.

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

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

Warrington, E. K., & Shallice, T. (1979). Semantic access dyslexia. *Brain, 102,* 43–63.

Warrington, E. K., & Shallice, T. (1984). Category specific semantic impairments. *Brain, 107,* 829–853.

Willshaw, D. J., Buneman, O. P., & Longuet-Higgins, H. C. (1969). Non-holographic associative memory. *Nature, 222,* 960–962.

Appendix A

Consider two randomly chosen points A and C in a space of dimension N. If B is the midpoint between A and C, then for large N there is a high probability of B being closer to A than to any member of a set of M points randomly chosen from the same distribution as A and C. If this is the case, then a hyperspherical attractor centered on target A will generally be able to come close to a hyperspherical attractor centered on target C without intersecting similar-sized attractors centered on the other targets in the set.

We are concerned with the case when the chosen points are the vertices of a unit hypercube, where there is a fixed probability p of choosing a 1 on each dimension; this approximates the semantic vectors used in our study.

The features (dimensions) can be divided into three subsets, the sizes of which are denoted by r, s, and t. For the r features where A and C disagree, B will have a value of 0.5 and will therefore contribute exactly the same amount to the squared distance from any vertex. For the s features where A and C are both 1, B will have a value of 1 and will therefore contribute nothing to the squared distance from A or C, but for randomly chosen other vertices, B will produce a contribution to the squared distance distributed according to a binomial with parameters s and $1 - p$. Similarly, for the t features where A and C are both 0, the contribution for a random vertex will be binomially distributed with parameters t and p.

For large N, almost all the variance in the squared distance between B and a randomly chosen vertex is contributed by these binomials, and almost none of this variance comes from the variance in the sizes of s and t because of the choices of A and C. So we can use the expected values for s and t. If we then approximate the binomials by gaussians, we get the following mean and standard deviation for the squared distance between B and a randomly chosen vertex:

$$\mu = 0.5^2 r + s(1 - p) + tp, \qquad \sigma = \sqrt{sp(1 - p) + tp(1 - p)}.$$

For $N = 68$ and $p = 0.22$, the expected values of s and t are 3.3 and 41.4, respectively. With these values for s and t, the expected squared distance of a random vertex from B is 17.5 with a standard deviation of 2.8. For A or C, the expected squared distance is 5.82.

The assumptions we made allow us to get a quantitative insight into the relatively large distances from B of other, random vertices. In fact, no vertex can actually be closer to B than either A or C.

(Appendixes continue on next page)

Appendix B

Table B1
Semantic Features

No.	Feature	No.	Feature
1	max-size-less-foot	34	partof-limb
2	max-size-foot-to-two-yards	35	surfaceof-body
3	max-size-greater-two-yards	36	interiorof-body
4	main-shape-2D	37	above-waist
5	main-shape-3D	38	mammal
6	cross-section-rectangular	39	wild
7	cross-section-circular	40	fierce
8	has-legs	41	does-fly
9	white	42	does-swim
10	brown	43	does-run
11	green	44	living
12	color-other-strong	45	carnivore
13	varied-colors	46	madeof-metal
14	transparent	47	madeof-wood
15	dark	48	madeof-liquid
16	hard	49	madeof-other-nonliving
17	soft	50	gotfrom-plants
18	sweet	51	gotfrom-animals
19	tastes-strong	52	pleasant
20	moves	53	unpleasant
21	indoors	54	man-made
22	in-kitchen	55	container
23	in-bedroom	56	for-cooking
24	in-living-room	57	for-eating-drinking
25	on-ground	58	for-other
26	on-surface	59	used-alone
27	otherwise-supported	60	for-breakfast
28	in-country	61	for-lunch-dinner
29	found-woods	62	for-snack
30	found-near-sea	63	for-drink
31	found-near-streams	64	particularly-assoc-child
32	found-mountains	65	particularly-assoc-adult
33	found-on-farms	66	used-for-recreation
		67	human
		68	component

Note. Directly interconnected sememes occur within the same section.

Table B2
The Words and Their Positive Semantic Features

Word	Features from Table B1																			
Bed	2	4	6	8	13	17	21	23	25	47	50	52	54	58						
Can	1	5	7	13	16	21	22	26	46	54	55	56	57	59	61	62	63			
Cot	2	5	6	13	17	21	23	25	47	50	52	54	58	64						
Cup	1	5	7	9	16	21	22	24	27	49	54	55	56	57	60	62	63			
Gem	1	5	13	14	16	21	23	24	27	49	52	54	58	59	65					
Mat	2	4	6	10	15	17	21	22	24	25	26	49	54	57	58					
Mug	1	5	7	13	16	21	22	24	26	49	54	55	57	59	60	62	63			
Pan	2	5	7	9	16	21	22	26	46	54	55	56	61	65						
Bug	1	5	8	13	15	20	25	28	29	31	33	39	41	43	44	53				
Cat	2	5	7	8	13	15	17	20	21	22	24	25	28	33	38	40	43	44	45	52	66
Cow	3	5	7	8	13	15	20	25	28	31	33	38	43	44	52	57	65				
Dog	2	5	7	8	10	20	21	22	24	25	28	33	38	40	42	43	44	45	52	58	66
Hawk	2	5	8	10	15	20	27	28	29	32	39	40	41	44	45	66					
Pig	2	5	7	8	12	17	20	25	28	33	38	43	44	57	65						
Ram	2	5	7	8	9	20	25	28	31	32	33	38	40	43	44	65					
Rat	1	5	7	8	10	15	20	25	28	33	38	39	40	42	43	44	45	53			
Back	2	4	6	12	16	20	21	27	35	37	44	67	68								
Bone	1	5	7	9	16	21	27	34	36	37	44	67	68								
Gut	3	4	7	9	17	21	27	36	37	44	53	67	68								
Hip	1	4	12	16	20	21	27	36	44	67	68										
Leg	2	5	7	12	16	20	21	25	34	35	36	44	67	68							
Lip	1	4	6	12	17	20	21	27	35	37	44	67	68								
Pore	1	4	7	12	17	21	27	34	35	37	44	67	68								
Rib	1	4	16	21	22	23	24	27	36	37	44	67	68								
Bun	1	5	7	10	17	18	19	21	22	26	50	52	54	62	64						
Ham	1	5	12	15	17	19	21	22	26	33	51	52	54	56	57	61	62				
Hock	2	5	7	14	19	21	22	26	48	50	52	54	57	59	61	63	65				
Lime	1	5	7	11	17	18	19	21	22	26	44	50	52	57	59	61	62				
Nut	1	5	7	10	15	16	19	21	22	24	26	28	29	44	50	52	57	59	62		
Pop	2	5	7	14	18	19	20	21	22	24	26	48	52	54	57	59	62	63	64		
Pork	1	5	9	17	19	21	22	26	33	51	52	56	57	61							
Rum	2	5	7	14	19	21	24	26	48	50	52	54	57	59	63	65					
Bog	3	4	11	17	20	25	28	31	32	39	48	49	53								
Dew	1	4	14	17	25	28	29	31	32	33	48	49	52								
Dune	3	5	10	17	25	28	30	39	49	52	66										
Log	2	5	7	10	15	16	25	28	29	33	47	50	58								
Mud	3	4	10	15	17	20	25	28	29	30	31	32	33	49	53	58	64				
Park	3	4	11	25	28	31	39	49	50	52	54	58	64	66							
Rock	3	5	10	15	16	25	28	30	31	32	39	49	58	66							
Tor	3	5	10	15	16	25	28	32	39	49	52	66									

Appendix C

The consistency of error type can be examined for a particular lesion site that Barry (1985) and Gordon, Goodman-Schulman, and Caramazza (1987) have examined in deep dyslexia. Gordon et al. have argued that it is a theoretically important variable for determining the locus of lesion sites responsible for the visual errors and for the semantic errors. In the model, some lesion sites produce inconsistent error types, but others produce consistent ones. Twenty trials per word were available for two lesions that produce a reasonable number of the three main types of error: *disconnect*(*isconns*, 0.2) and *disconnect*(*giconns*, 0.3). The nature of second occurrence of an error on a particular word was compared with the first occurence of an error on that word. For the former lesion type, 10 of 13 words produced at least two or more errors that were consistent in their error type (3 semantic, 5 mixed, and 2 visual; $C = 0.71$, significantly different from 0, $p < .01$). However, for the latter lesion type, only 5 of 14 were consistent, which is very close to the chance value (4.1 of 14). Repetition in the disconnection procedure in fact gives rise to different lesions, albeit qualitatively and quantitatively equivalent ones. However, this does not affect the general point made in the Discussion section concerning inferences that one can make about location of lesion from consistency of error pattern. If different lesions at the same locus can give consistent performance across words as far as error type is concerned, then presumably so will the identical lesion.

Received April 21, 1989
Revision received May 6, 1990
Accepted June 6, 1990 ■