Chapter 2

Connectionist modeling in neuropsychology

This thesis is concerned with the application of connectionist models in cognitive neuropsychology, with the intent of better understanding both patients and networks. This chapter presents background material that serves as the context and motivation for research presented in subsequent chapters. It begins with an overview of cognitive neuropsychology, emphasizing studies of selective deficits in reading, known as "acquired dyslexias." Following this, the role of computational modeling in cognitive neuropsychology is discussed. The advantages of adopting a connectionist approach are described and illustrated by specific applications in a number of domains. We focus on efforts to reproduce the behavior of different types of dyslexic patients within connectionist models of reading. The reading behavior of "deep" dyslexics in particular is then described in more detail. A previous preliminary attempt at modeling these patients is presented and evaluated. The specific strengths and weaknesses of the design of this model serve to motivate much of the research presented in the thesis.

2.1 Cognitive neuropsychology

Many cognitive abilities can be selectively impaired as a result of brain damage. Among these are object recognition, selective attention, reading, writing, language understanding, speech production, learning, memory, planning, and reasoning. The field of cognitive neuropsychology studies the patterns of impaired and preserved abilities of brain-injured patients, and attempts to relate them to models of normal cognitive functioning. The aim is both to explain the behavior of the patients in terms of the effects damage in the model, and to inform the model based on the observed behavior of patients (Coltheart, 1985; Ellis & Young, 1987).

At first it may seem that little can be learned about the normal operation of an information-processing device as complicated as the brain on the basis of how it behaves under damage. To put it bluntly, "What can you possibly learn about the way a car works (or a vacuum cleaner, or a computer) by pounding it with a sledgehammer" (Marin et al., 1976, p. 868; also see Gregory, 1961). A detailed consideration of the theoretical difficulties in relating studies of normal and impaired cognitive processes is beyond the scope of this overview—readers are referred to Shallice (1988) for a thorough treatment. For our purposes, we will consider it an open issue as to what and how studies of brain-injured patients can contribute to our understanding of normal cognition. In a sense, the degree to which this endeavor succeeds may be taken as evidence that brains are not organized like cars, or (conventional) computers for that matter.

2.1.1 Modularity and dissociations

At its most basic level, the viability of cognitive neuropsychology rests on what is known as the "modularity hypothesis" (Fodor, 1983; Marr, 1976; Simon, 1969). On this hypothesis, the "functional architecture" of the mind is composed of relatively independent subsystems, or modules, that specialize in carrying out a particular function or in processing a particular type of information. The modules are also claimed to be neuroanatomically separate, and so can be independently impaired by brain damage. The most basic argument in favor of a modular organization is that it simplifies the design and improvement of the system (Marr, 1976; Simon, 1969).

> Any large computation should be split up and implemented as a collection of small sub-parts that are as nearly independent of one another as the overall task allows. If a process is not designed in this way, a small change in one place will have consequences in many other places. This means that the process as a whole becomes extremely difficult to debug or to improve, whether by a human designer or in the course of natural evolution, because a small change to improve one part has to be accompanied by many simultaneous compensating changes elsewhere. [Marr, 1976, p. 485]

Empirical support for modularity comes from existence of highly specialized cortical areas (Van Essen, 1985), and the success of the "additive factors" method in psychology (Sternberg, 1969). Perhaps the most direct evidence for isolable cognitive subsystems comes from the occurrence of selective cognitive deficits in some neurological patients.

An important methodology in cognitive neuropsychology for isolating cognitive modules involves demonstrating "dissociations" between the performance of a patient on different tasks. Two tasks are dissociated when the patient performs significantly worse on one than the other. Ideally, performance on the poor task is much worse than that of normals, while performance on the good task is within the normal range.

A single dissociation provides only limited information because it is possible that the poorly performed task is simply much more difficult, but still well within the abilities of normals. Of particular importance is the demonstration of a "double" dissociation, in which a second patient is found who also exhibits a dissociation on the same two tasks, but in the *opposite* direction. Under the assumption that the organization of cognitive processes in the two patients is essentially the same, a double dissociation rules out an explanation based simply on the relative difficulties of the tasks. In this case, the fact that each task can be selectively impaired while leaving the other unaffected provides strong evidence that the two tasks are carried out by different cognitive mechanisms.[1]

Rather more problematic, but still informative, are "associations" among symptoms exhibited by patients. If a patient is impaired on both of two tasks, it suggests that they are subserved by the same mechanism, which is now impaired. However, it may also be that they are subserved by separate mechanisms that just happen to be affected by the same brain lesion because they are neuroanatomically close, or perhaps the patient has suffered multiple lesions. The argument for a common mechanism is strengthened somewhat if all or most of a large number of similar patients exhibit the same association, but neuroanatomical proximity remains a possible, if perhaps unpalatable, explanation. In fact, a major empirical puzzle addressed by simulations in this thesis is the remarkable consistency in the association of large, diverse set of symptoms exhibited by virtually all patients with a particular type of acquired reading disorder.

2.1.2 Acquired dyslexia

One of the most intensely studied areas in cognitive neuropsychology in recent years has been the types of selective deficits in reading that can follow brain damage, known as "acquired dyslexias." Prior to the early 1970's, the major distinction among such patients was simply whether the reading deficit was accompanied by a deficit in writing—"alexia with agraphia"—or whether it occurred in isolation—"alexia without agraphia," or "pure alexia" (Dejerine, 1892).[2] Little attempt was made to distinguish among different types of reading deficits until Marshall & Newcombe (1973) identified three separate types of acquired dyslexia based on the typical patterns of errors that patients made in reading aloud. "Surface" dyslexia involved phonological confusions in the procedure by which words are sounded-out based on the typical pronunciations of sets of letters (e.g. INSECT ⇒ "insist", hard to soft *c*). "Deep" dyslexia involved semantic confusions, in which words were often misread as semantically related words (e.g. DINNER ⇒ "food"). The third type that Marshall & Newcombe identified, involving visual confusions between words, has not been as widely recognized as either surface or deep dyslexia.

A particularly notable aspect of Marshall & Newcombe's work is that they explained the existence of these distinct types of dyslexia in terms of damage to a "dual-route" model of normal reading (see Figure 2.1). In the model, word pronunciations can be generated through either of two mechanisms. The first is a phonological system that translates from spelling to sound by the use of grapheme-phoneme correspondences (GPCs). This system enables people to read word-

[1] The logic of making inferences from single and double dissociations is, in general, far more complicated than as described here, particularly when a patient's performance on the "unimpaired" task is not within normal limits.

[2] The terms "alexia" and "dyslexia," referring to deficits in reading, are interchangeable for our purposes. "Acquired" dyslexia refers to reading disorders in a previously literate individual, usually as a result of brain injury. "Developmental" dyslexia refer to reading disorders in individuals who never learned to read normally.

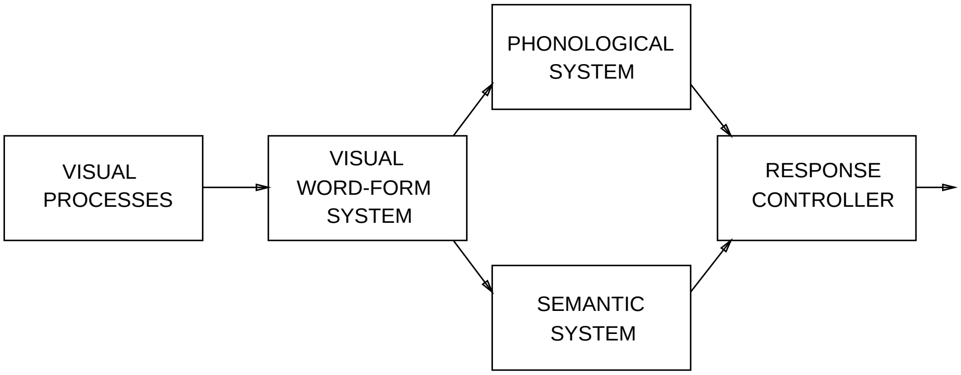

Figure 2.1: Marshall & Newcombe's (1973) dual-route model of reading (from Shallice, 1988, p. 71).

like nonsense letter strings (e.g. MAVE) as well as words with "regular" pronunciations.[3] The second mechanism for pronouncing words is a semantic system that recognizes and assigns them meaning. The specific pronunciation of a word can then be directly accessed from its meaning. This semantic route enables people to read "exception" words that violate the standard GPC rules. Surface dyslexics were held to have damage to the semantic route, and thus read only by the phonological route. Conversely, deep dyslexics were held to read only via the semantic route because the phonological route was damaged. The errors produced by these patients reflected the imperfect operation of the remaining route in isolation. The third type of dyslexia was thought to involve damage to visual processes prior to each of these routes—in more recent terminology it would be termed a "peripheral" dyslexia, in contrast to surface and deep dyslexia, which are "central" (Shallice & Warrington, 1980).

Further research has examined the characteristics of both surface and deep dyslexia in more detail. Some of this research has prompted a division of surface dyslexics into two separate types (Shallice & McCarthy, 1985). Type I patients (e.g. M.P., Bub et al., 1985; K.T., McCarthy & Warrington, 1986; H.T.R., Shallice et al., 1983) exhibit fluent and correct reading of words with regular spelling-to-sound correspondences (e.g. MUST), even though word comprehension is severely impaired. These patients can also appropriately pronounce non-words (e.g. NUST). However, they often misread exception words, usually by giving a more regular pronunciation (e.g. PINT mispronounced to rhyme with MINT). This type of "regularization" error is most common for exception words that occur infrequently—common exception words are read almost as well as regular words. Thus there is an interaction between the effects of frequency and of regularity. Type II surface dyslexics (e.g. J.C., S.T., Marshall & Newcombe, 1973; R.O.G., Shallice & Warrington, 1980) are also better at reading regular words than exception words, and show a frequency ×

[3] "Regular" words, such as MUST, contain a spelling pattern (–UST) that occur in many other words, always with the same pronunciation. "Exception" words, such as HAVE, also contain a common spelling pattern, but it is usually pronounced *differently* than it is in this word (cf. GAVE, SAVE).

regularity interaction. However, compared with Type I patients, they read more slowly with many corrections, make more errors on regular words and non-words, produce a lower proportion of regularizations of exception words, and have better comprehension. In fact, Shallice (1988; Shallice & McCarthy, 1985) argues that Type II surface dyslexia is caused by a more peripheral visual disturbance, and that only Type I surface dyslexia, renamed "semantic" dyslexia, truly reflects the isolated operation of the phonological route(s). Readers are referred to Patterson et al. (1985) for a more comprehensive presentation of surface/semantic dyslexia.

In contrast to surface dyslexics, deep dyslexics cannot read non-words, and their ability to read a word is unaffected by the nature of its spelling-to-sound correspondences (Coltheart et al., 1980). In addition to semantic errors, they also make visual errors (e.g. SCANDAL ⇒ "sandals"), derivational or morphological errors (e.g. HIRE ⇒ "hired"), and function word substitutions (e.g. AND ⇒ "of"). They also find content words much easier to read than function words, and concrete words much easier to read than abstract words. The behavior of deep dyslexics is described in more detail below.

In addition to surface and deep dyslexia, a third type of central dyslexia, involving a more selective impairment of the phonological route, was identified by Beauvois & Derouesné (1979; also see Funnell, 1983; Patterson, 1982; Shallice & Warrington, 1980). Similar to deep dyslexics, patients with "phonological" alexia cannot read non-words, and yet have no problem pronouncing them. However, unlike deep dyslexics, phonological alexics have almost completely intact word reading, and they show little if any effect of abstraction or part-of-speech, with the possible exception of some difficulty on function words. Some theories of reading (e.g. Coltheart, 1985; Morton & Patterson, 1980; Sartori et al., 1987; Schwartz et al., 1980) interpret the behavior of these patients as providing evidence for a third, lexical non-semantic route, separate from both the phonological and semantic routes. Other theories (e.g. Seidenberg & McClelland, 1989; Shallice & McCarthy, 1985; Van Orden et al., 1990) propose a single, "broad" phonological route that maps between orthography and phonology at different levels of structure (individual graphemes-phonemes to whole words).

Although these additional empirical findings show that Marshall & Newcombe's early model of selective impairments in reading is far too simplistic, the form and underlying assumptions of their explanation have survived and flourished in cognitive neuropsychology (Caramazza, 1984; 1986; Coltheart, 1985; Coltheart et al., 1987b; Patterson et al., 1985). The most basic among these assumptions is that cognitive processes like reading should be described in terms of simple interactions between functionally isolable subsystems. The resulting "functional architecture" is typically cast as an information-processing flow diagram, otherwise known as a "box-and-arrow" model. It is assumed that brain damage can selectively impair or eliminate particular components in the model, while the remaining components continue to operate normally. Giving an account of the deficits of a particular patient involves specifying a functional architecture, together with

Figure 2.2: An explanation for deep dyslexia in terms of lesions (enclosed by dotted lines) to a functional architecture for understanding and pronouncing written and spoken words (from Morton & Patterson, 1980, p. 115). Written words are presented as stimuli to the upper-right "visual analysis" component, and its pronunciation is read out of the "response buffer" at the bottom. The "grapheme-phoneme conversion" component corresponds to phonological system in Figure 2.1, while the pathway through the "cognitive system" corresponds to the semantic system. The third (lexical non-semantic) route goes from "visual input logogens" directly to "output logogens." Notice that Morton & Patterson are claiming that deep dyslexics have multiple partial impairments along the semantic route in addition to complete impairment of the non-semantic routes.

a set of "lesions" to the architecture, such that the resulting system "exhibits" the same pattern of impaired and preserved behavior as the patient. Figure 2.2 illustrates this approach for deep dyslexia (Morton & Patterson, 1980).

The fact that "exhibits" must be scare-quoted in the preceding paragraph reveals a limitation in the way this methodology has typically been applied. Specifically, most effort has focused on identifying and fractionating the components of the functional architecture, whereas relatively little effort has gone into specifying the representations and computations within each component—exactly how the boxes work. In this way, conventional theorizing in cognitive neuropsychology implicitly subscribes to the philosophy that it is possible to characterize a computation independently of the precise details of how that computation is implemented (Marr, 1982). However, this approach makes it is difficult to derive predictions of the behavior of the model without more specific claims about the nature of the representations and computations that actually produce the behavior (Seidenberg, 1988).

The most explicit formulations of the internal operations of components in neuropsychological models of reading are based on Morton's (1969) "logogen" model. At the core of the approach is a set of evidence-accumulating threshold devices, called "logogens," that mediate the recognition and/or response to particular words. A distinguishing feature of the model is that the logogens themselves contain no semantic information, but can interact directly with what is typically referred to as the "cognitive system." The exact nature of the cognitive system is unspecified, but is typically assumed to contain all of the higher-order processing not explicitly accounted for by other parts of the model (Morton & Patterson, 1980). Unfortunately, models of this form are rarely explicitly implemented and damaged—more often the predictions of the behavior of the damaged model are based on little more than informed intuition.

2.2 Computational modeling in cognitive neuropsychology

Typically, predictions of the behavior of a model, both in normal operation and under damage, have consisted of descriptions based on fairly general notions about how the various modules would operate and interact. While these types of predictions may suffice for capturing the more general characteristics of normal and impaired cognitive functioning, they become increasingly unreliable as the model is elaborated to account for more detailed phenomena. Computational modeling makes it possible to demonstrate the *sufficiency* of the underlying theory in accounting for the phenomena by making the behavior of a detailed cognitive model explicit. A working simulation guarantees that the underlying theory is neither vague nor internally inconsistent, and the behavior of the simulation can be used to generate specific predictions of the theory. However, building a working simulation involves making design decisions that introduce theoretically irrelevant details. It is often difficult to identify what aspects of a model are responsible for its success, and the degree to which these aspects are theoretically motivated. The most comprehensive approach involves an interplay between computational and empirical work, in which simulations focus experimentation on particular issues, and empirical results constrain the development of the computational model.

2.2.1 Conventional implementations

There are many alternative computational formalisms within which to develop computational models of neuropsychological disorders. Perhaps the most straightforward modeling approach denies the importance of the choice of formalism. In this way it retains the perspective of most box-and-arrow theorizing, that the identity and function of each component in the model can be abstracted from the details of how that function is implemented. The clearest example of this style of research is found in Kosslyn et al.'s (1990) work in modeling object recognition and visuospatial processing in high-level vision.

Although a detailed description of the operation of the model falls outside our current concerns, a consideration of the general form and approach of the model is instructive. The model is cast within a conventional box-and-arrow framework (see Figure 2.3). There are components for recognizing objects based on features derived from information in a "visual buffer" that falls within an "attentional window." Other components associate object identity with the identities and positions of parts of the object, and carry out the spatial transformations that relate the predicted parts with what is found in the visual buffer. The model aims to reproduce, and thereby account for, a wide range of normal and impaired behavior in high-level vision.

A guiding principle in the design of the model is the "hierarchical decomposition constraint," which requires that elaboration of the model must only involve subdividing existing components rather than introducing components that cut across existing boundaries. The assumption is that consideration of implementation details is only germane *after* the overall organization of the system is settled. In fact, Kosslyn et al. (1990) provide very few details of the implementation of the model, stating "we are interested in the ways subsystems interact, not in how they actually process input" (p. 243). Nonetheless, some aspects of the implementation can be inferred from descriptions in the text and from personal communication with S. Kosslyn.

The model is implemented as a conventional program, in which each component is a subroutine that manipulates data structures such as lists and arrays. Components can operate in parallel; however, each component must complete its computation before it generates output to other components (cf. McClelland, 1988). The control structure that determines when components operate is specified externally based on the task that is being performed. The effects of "damage" to each component, or pathway between components, are explicitly defined based on the nature of the computation performed by the component.

The implemented model performs a range of recognition and discrimination tasks on a small set of schematic, hand-segmented inputs. When the definitions of damage are applied to the model, it fails to perform some or all of the tasks depending on the nature of the damage. Some of these impairments are similar to the basic characteristics of some neuropsychological patients. Although many types of patients are discussed in the context of the model, no attempt is made to actually replicate their behavior in any detail.

In evaluating this type of work, it is important to determine how the development of a computational model contributes to our understanding beyond the unimplemented description of the theory. The role of computational modeling is less interesting if it merely verifies the coherence of the theory. Ideally, modeling should contribute to the development of the theory *per se*. In fact, Kosslyn and his colleagues identify a number of ways in which the development of the model from earlier work (Kosslyn, 1987) was shaped by the process of developing a working imple-

Figure 2.3: Kosslyn et al.'s (1990) model of high-level vision. The top figure (p. 214) presents the major subsystems of the model; the bottom figure (p. 237) presents the decomposition of these subsystems into specific functional components.

mentation (footnote p. 241). Unfortunately, the possible opportunities for this type of interaction between modeling and theorizing are significantly limited by the lack of attention paid to implementation details. In essence, the computational formalism of conventional programming is too unconstrained—it provides little bias towards carrying out computations in one way vs. another. Computational modeling can have a much more profound impact on theorizing when the nature of the formalism strongly influences the organization of the representations and computations in the model.

2.2.2 Connectionist approaches

Computational modeling is most interesting when the formalism significantly contributes to a natural explanation for empirical phenomena that are counterintuitive when viewed within other formalisms. For this reason, connectionist networks are becoming increasingly influential in a number of areas of psychology as a methodology for developing computational models of cognitive processes. These networks compute via the massively parallel cooperative and competitive interactions of a large number of simple neuron-like processing units. Typically, the specific unit interactions are not directly under the control of the experimenter, but are determined by a general learning procedure. In this way, the formalism places strong constraints on the nature of the computation that is available in modeling, often with interesting consequences. Networks of this form have been applied to a wide range of problems in perception, language, and reasoning (McClelland et al., 1986).

In addition to their usefulness in modeling normal cognitive functioning, a number of general characteristics of connectionist networks suggest that they may be particularly well-suited for modeling neuropsychological phenomena (Allport, 1985).

- “Modular” theories of cognitive processes can be expressed naturally by dedicating separate groups of units to represent different types of information. In this way the approach can be viewed as an elaboration of, rather than alternative to, more traditional “box-and-arrow” theorizing within cognitive neuropsychology (cf. Seidenberg, 1988).
- Partial lesions of neurological areas and pathways can be modeled in a straightforward way by removing a proportion of units in a group and/or connections between groups. In contrast, simulations of neuropsychological findings within more traditional computational formalisms (e.g. Kosslyn et al., 1990) must typically make more specific assumptions about how damage affects particular components of the system.[4]

[4] Kosslyn et al. (1990) suggest that each component of their model could be implemented by a separate connectionist network, but they mention no implications of doing so. In fact, the logic of their “hierarchical decomposition” approach implies that they believe a connectionist implementation would have no interesting consequences for the theoretical design of their system.

- Since knowledge and processing in a connectionist network is distributed across a large number of units and connections, performance degrades gracefully under partial damage (Hinton & Sejnowski, 1986). This means that a range of intermediate states between perfect performance and total impairment can occur. Together with the richness of the computational formalism, this allows behavior more detailed than the simple presence or absence of abilities to be investigated (Patterson, 1990).

A number of authors have attempted to explain patient behavior based on intuitions about how connectionist networks or other "cascaded" systems (McClelland, 1979) would behave under damage, without actually carrying out the simulations (e.g. Miller & Ellis, 1987; Riddoch & Humphreys, 1987; Shallice & McGill, 1978; Stemberger, 1985). However, the highly distributed and dynamical nature of these networks makes such unverified predictions somewhat suspect. More recently, a few researchers have begun to explore the correspondence of the behavior of damaged connectionist networks and patient behavior in a wide range of domains. McClelland & Rumelhart (1986) modeled the patterns of retrograde and anterograde memory deficits in amnesics by degrading the consolidation of weight changes during learning. Cohen & Servan-Schreiber (1989) reproduced the characteristics of the abnormal use of context in schizophrenics by adjusting the sensitivity of units to their inputs in a way that corresponded to the influence on individual neurons of the abnormal levels of excitatory neurotransmitters in these patients. Levine (1986; Bapi & Levine, 1990; Levine & Prueitt, 1989) modeled the tendency of patients with frontal lobe damage to repeat previous responses, and be overly distracted by novel stimuli, by disrupting the interactions between sensory and reinforcement representations in an ART network (Carpenter & Grossberg, 1987). Farah & McClelland (1991) reproduced a selective deficit in recognizing and recalling functional information about living vs. non-living things (Warrington & Shallice, 1984) by introducing damage to visual semantics, under the hypothesis that the representations of living things rely more heavily on visual vs. functional information. More recently, Cohen et al. (1992) mimicked the apparent difficulty that some patients with parietal damage have in "disengaging" attention from an ipsilesional location to attend to a contralesional stimuli (Posner et al., 1984) by unilaterally damaging a competitive mechanism for allocating attention. Each of these studies shows how particular computational properties of connectionist networks can contribute to our understanding of complex, often counterintuitive, neuropsychological phenomena.

2.3 Connectionist modeling of acquired dyslexia

Perhaps the most detailed attempts at relating the behavior of damaged connectionist networks to that of brain-injured patients has been in the domain of acquired dyslexia (Hinton & Shallice, 1991; Mozer & Behrmann, 1990; Patterson et al., 1990). This is in part because investigations of reading in both cognitive psychology and neuropsychology (Coltheart, 1987) have produced a rich, and often counterintuitive, set of empirical findings. In addition, reading is appealing as a domain for computational modeling because the surface forms (i.e. strings of letters and phonemes) are fairly simple. Furthermore, reading is a particularly appropriate domain in which to use relatively unstructured connectionist models (cf. Feldman et al., 1988; Sejnowski et al., 1989) to study the neural implementation of cognitive processes as it is unlikely that the brain relies on specialized neural circuitry to accomplish such an evolutionarily recent skill. Nonetheless, separate neuroanatomical *areas* may become specialized for particular reading processes as a result of experience during reading acquisition, and thus be independently susceptible to brain damage (cf. Farah, 1990). Because much of the research presented in this thesis concerns connectionist modeling of particular deficits in reading, we will review this work in more detail.

2.3.1 Neglect and attentional dyslexias

Neglect dyslexia is a peripheral reading disorder, typically following right parietal damage, in which patients often ignore the leftmost portion of written material, even when it falls entirely within the intact portions of their visual fields (Caplan, 1987; Kinsbourne & Warrington, 1962; Sieroff et al., 1988; also see the special issue of *Cognitive Neuropsychology*, 7(5–6), 1991, on “Neglect and the Peripheral Dyslexias”). The accuracy of reading a string letters is better when the stimulus is presented further to the right (Behrmann et al., 1990; Ellis et al., 1987), or when it forms a word (Behrmann et al., 1990; Brunn & Farah, 1991; Sieroff et al., 1988). Incorrect responses typically consist of words in which the leftmost letters of the stimulus are omitted (e.g. CHAIR ⇒ “hair”), replaced (e.g. HOUSE ⇒ “mouse”), or augmented (e.g. LOVE ⇒ “glove”). When two letter strings are presented (e.g. SUN FLY), the left one is often ignored (Sieroff et al., 1988), but this occurs less frequently when they combine to form a compound word (e.g. COW BOY; Behrmann et al., 1990). Thus, the severity of the deficit is influenced both by peripheral (sensory) and central (lexical) manipulations. Neglect dyslexia often accompanies more generalized hemispatial neglect (Bisiach & Vallar, 1988; Friedland & Weinstein, 1977) but has been dissociated from it in a least some patients (Costello & Warrington, 1987). It is traditionally interpreted as a deficit in allocating spatial attention to contralesional stimuli (Posner et al., 1984). In fact, explicit instructions or cueing manipulations that bias attention towards the left can often alleviate the deficit (Karnath, 1988; Riddoch & Humphreys, 1983; Riddoch et al., 1991), and neglect-like attentional manipulations in normals can elicit analogous lexical effects (Behrmann, 1991; Behrmann et al., 1991).

Mozer & Behrmann (1990) reproduced these characteristics of neglect dyslexia in a pre-existing connectionist model of word recognition, known as MORSEL (for Multiple Object Recognition and SELective attention; Mozer, 1988; 1990; see Figure 2.4). Retinotopic letter features are combined into letters, and then into “letter clusters,”[5] by the operation of a hierarchically organized

[5] Letter clusters are context-dependent triples of letters that explicitly represent spaces (designated by “*”) and can span across an intermediate position (designated by “_”). For example, the letter clusters for the isolated word CAT

Figure 2.4: The main components of MORSEL (from Mozer & Behrmann, 1990, p. 98).

subnetwork called BLIRNET (for Builds Location Invariant Representations). The input of letter features to BLIRNET is partially gated by an Attentional Mechanism (AM) that attempts to form a spatially contiguous "spotlight" of activity on the basis of where letter features occur. The letter cluster activity produced by BLIRNET is cleaned-up into the pattern for a particular word under the top-down influence of lexical/semantic units within a Pull Out network.

Mozer & Behrmann (1990) model the attentional impairment in neglect dyslexia by introducing a monotonic gradient of damage to the connections from the letter features to the AM, with damage most severe on the left and least on the right. This damage biases the AM towards forming an inaccurate spotlight that includes only the rightmost letters of a single input string, or the rightmost of two input strings. Letter features that fall outside the spotlight are transmitted to BLIRNET much less effectively, so that the resulting letter cluster activity is inaccurate, particularly in representing the left-hand side of the input. The clean-up of the Pull Out network can often reconstruct the correct pattern of activity for the letter string from this corrupted activity, particularly when the entire input forms a word (corresponding to some of the lexical/semantic units). However, when clean-up within the Pull Out network fails, the result is often the pattern for another word that differs from the presented word only on the left. Reading accuracy is better if the letter string is presented further to the right because the damage from these positions to the AM is less severe. Similarly, accuracy is improved by cueing in the model through the unimpaired top-down input to the AM from so-called "Higher Levels of Cognition." In this way, the damaged model reproduces the main characteristics of neglect dyslexia.

In addition to neglect dyslexia, Mozer & Behrmann (in press) describe how damage in MORSEL might also account for another type of peripheral dyslexia, known as "attentional" dyslexia (Shallice & Warrington, 1977). Attentional dyslexics can correctly read single words or letters when presented in isolation, but have difficulty when multiple items are presented together. Thus, these patients often cannot identify the individual letters within a word that they can read. Identifying a letter is somewhat improved when it is flanked by digits rather than other letters (e.g. 83B40 vs. LHBMC). In addition, when multiple words are presented, letters can migrate between the words in the response (e.g. WIN FED read as FIN FED). Similar letter migration errors occur in normals under brief masked exposure (Mozer, 1983; Shallice & McGill, 1978).

Mozer & Behrmann propose that attentional dyslexia arises when the AM is unable to focus its spotlight on only one of multiple items. Many types of damage in the model would cause the spotlight to capture everything in the visual field. One possibility is that reduced connection weights in the AM slow the formation of the spotlight. In fact, Mozer (1988) produced letter migration errors in MORSEL when the AM was given insufficient time to settle into focusing on just one of two words. The Pull Out network occasionally recombines the letter clusters from

are **C, **_A, *CA, *_AT, CAT, C_T*, AT*, A_**, and T**. Letter clusters provide a unique representation for most words (but see Pinker & Prince, 1988, for a discussion of their limitations). The scheme is loosely modeled after that of Wickelgren (1969).

Figure 2.5: Seidenberg & McClelland's general framework for lexical processing (left), and the structure of the portion of this framework that was implemented (right) (from Seidenberg & McClelland, 1989, pp. 526–527).

both words inappropriately. Individual letters in words cannot be identified because the AM must focus its spotlight on a single letter in order for BLIRNET to generate the letter cluster activity that corresponds to that letter as the response. Reading a letter in the context of digits is easier than in the context of other letters because the network does not form clusters between letters and digits, so in the former case the letter cluster activity more closely approximates that for isolated letter presentation.

Thus one type of attentional manipulation in MORSEL leads to neglect dyslexia, while another leads to attentional dyslexia. MORSEL was originally developed to account for aspects of word reading in normals—the fact that it exhibits both neglect and attentional dyslexia under damage provides independent support for the model.

2.3.2 Surface dyslexia

Recall that surface dyslexics can correctly pronounce regular words and non-words, but often make regularization errors on low-frequency words (e.g. YACHT ⇒ "yatched"). Thus they would appear to be reading entirely by a phonological route that sounds-out words based on spelling-to-sound correspondences. Patterson et al. (1990) attempted to reproduce similar effects by damaging a model of word pronunciation that had been previously demonstrated to account for a wide range of effects in normal reading (Seidenberg & McClelland, 1989, see Figure 2.5). The model takes the form of a connectionist network that maps orthographic representations of the written forms of words onto phonological representations of their pronunciations. In this way it is closely related to Sejnowski & Rosenberg's (1987) NETtalk model. Orthographic representations in the Seidenberg & McClelland model are roughly similar to the letter clusters in MORSEL, except that each of the 400 orthographic units is involved in representing a number of related letter clusters instead of just one. Phonological representations, composed of triples of phonemic features, are distributed similarly over 460 phonological units (for details, see Rumelhart & McClelland, 1986). In one version of the network, there are 200 hidden units that receive connections from each orthographic unit, and that send connections back to these units as well as to each phonological unit.[6] The network was trained with back-propagation (Rumelhart et al., 1986b) to generate both the orthographic and phonological representation of each of 2897 monosyllabic English words when presented with its orthographic representation as input. The frequency with which each word was presented to the network during learning was proportional to the logarithm of its frequency of occurrence in written English (Kucera & Francis, 1967). After 250 sweeps through the training corpus (about 150,000 word presentations), the network's pronunciation of 97.3% of the words matched the correct pronunciation better than any alternative pronunciation that differed by a single phoneme. A number of the incorrect responses were regularizations of low-frequency words (e.g. SOOT ⇒ "suit").

The model succeeds at simulating a broad range of empirical phenomena in reading. The most important results for our purposes involve the time required to pronounce different types of words, where naming latency in the model is defined to be directly proportion to the distance between the generated and correct pronunciation. Specifically, in normal subjects and in the model, (a) there is a main effect of frequency, with high-frequency words read more quickly that low-frequency words (Forster & Chambers, 1973; Frederiksen & Kroll, 1976), and (b) there is a frequency × regularity interaction, with the faster naming times for regular vs. exception words being much larger for those words with low frequency compared with those with high frequency (Seidenberg, 1985; Seidenberg et al., 1984; Taraban & McClelland, 1987; Waters & Seidenberg, 1985), and (c) words are read better than non-words (McCann & Besner, 1987). These effects occur in the model because both frequency and regularity combine to strengthen common associations between subpatterns of the orthographic and phonological representations.

Perhaps the main contribution of the Seidenberg & McClelland model is that it demonstrates that a single mechanism can pronounce regular words, exception words, and non-words (although see Besner et al., 1990, for criticism of the model's word and non-word reading performance, and Seidenberg & McClelland, 1990, for a rejoinder). In contrast, most models of reading claim that exception words must be pronounced by a separate "lexical" mechanism because their spelling-to-sound correspondences violate the rules that apply to regular words and non-words (see Patterson

[6] This connectivity would suggest that the implemented network is recurrent and would have to repeatedly update unit states in processing a given input. However, unit states are computed only once in the Seidenberg & McClelland simulations. Thus, it is more accurate to think of the implemented model as a feed-forward network that sends connections to a *separate* group of orthographic (output) units rather than back to the input units.

& Coltheart, 1987, for a review). Strong evidence in favor of this separation comes from the characteristics of two classes of acquired dyslexics: "surface" dyslexics are much better at reading regular words than exception words, and "phonological" alexics are much better at reading words than non-words. Accounting for these two patterns of deficits in terms of damage to a single naming mechanism presents a difficult challenge for the Seidenberg & McClelland model.

Unfortunately, Patterson et al. (1990) decline to consider how the model might account for phonological alexia. However, the model does seems ideally suited for reproducing surface dyslexia because neither it nor the patients (at least Type I) can rely on semantic mediation in reading. Accordingly, Patterson and her colleagues "lesioned" the model, by removing different proportions of units or connections, and compared its performance on different classes of words. The pronunciation generated by the damage network to a given word was compared with the correct pronunciation as well as a plausible alternative—for exception words, the alternative consisted of the "regularized" pronunciation. After damage, regular and exception words are read equally well, and there is no effect of frequency in reading exception words. Exception words are much more likely than regular words to produce the alternative (regularized) pronunciation, but a comparison of the phonemic features in errors revealed that the network shows no greater tendency to produce regularizations than other errors that differ from the correct pronunciation by the same number of features. This pattern of results is unlike that of surface dyslexics, who read regular words much better than exception words, and are worse at reading exception words with low frequency, and are particularly prone to regularization errors (also see Behrmann & Bub, in press).

Using a more detailed procedure for analyzing responses, Patterson (1990) found that removing 20% of the hidden units produced better performance on regular vs. exception words and a slight (non-significant) trend towards a frequency × regularity interaction. Thus, both high- and low-frequency regular words are read well (93% for both). High-frequency exception words are also read reasonably (86%) but low-frequency exception words are more impaired (78%). In addition, half of the errors to exception words were regularizations. These effects are similar to those seen in some surface dyslexics, but not as dramatic.

Taken together, the attempts to model surface dyslexia by damaging the Seidenberg & McClelland model have been less successful than its ability to account for normal reading behavior.[7] However, Patterson et al. (1990) and Seidenberg & McClelland (1990) point out a number of possible elaborations to the model that might improve its ability to behave like surface dyslexics when it is damaged.

[7] Interestingly, Olsen & Caramazza (1988) were also unsuccessful at reproducing the characteristics of "lexical agraphia" (Beauvois & Derouesné, 1981), a writing disorder analogous to surface dyslexia, by damaging a connectionist model of spelling analogous to NETtalk (Sejnowski & Rosenberg, 1987). However, more recent work by Loosemore et al. (1991) appears promising.

2.3.3 Deep dyslexia

In the conclusion of their review article, “Deep Dyslexia since 1980,” Coltheart, Patterson & Marshall (1987a) argue that deep dyslexia presents cognitive neuropsychology with a major challenge. They raise two main issues specific to the domain of reading. First, they argue that standard “box-and-arrow” information-processing accounts of deep dyslexia (e.g. Morton & Patterson, 1980, see Figure 2.2) provide no explanation for the observed combinations of symptoms. If a patient makes semantic errors in reading aloud, why are many other types of behavior virtually always observed? Second, they point out that the standard explanations for semantic errors and for effects of abstractness involve *different* impairments along the semantic route.

> The loss of semantic information for abstract words that explained visual errors in oral reading cannot readily explain semantic errors in oral reading, since semantic errors typically occur on moderately concrete words.... The deficit in the semantic routine that gives a pretty account of semantic errors is, rather, an abnormal sloppiness in the procedure of addressing a phonological output code from a set of semantic features. .... Must we now postulate several different semantic-routine impairments in deep dyslexia, and if so, why do we not observe patients who have one but not the other: in particular, patients who make semantic errors but do not have difficulty with abstract words? [Coltheart et al., 1987a, pp. 421–422]

Recently, Hinton & Shallice (1991) have put forward a connectionist approach to deep dyslexia that addresses the first of the above points. They reproduced the co-occurrence of semantic, visual, and mixed visual-and-semantic errors by lesioning a connectionist network that develops “attractors” for word meanings. While the success of their simulations is quite encouraging, there is little understanding of what underlying principles are responsible for them. A major focus of this thesis is to evaluate and, where possible, improve on the most important design decisions made by Hinton & Shallice.[8] First, we improve on the rather arbitrary way that the model realized an explicit response by extending it to generate phonological output from semantics. Next, we demonstrate the robustness of the account by examining network architectures different from the original model. Thirdly, we evaluate the significance of the particular learning procedure used to train the original model by re-implementing it in a more plausible connectionist formalism. Finally, we investigate whether the remaining characteristics of deep dyslexia—in particular, Coltheart, Patterson & Marshall’s third issue relating to effects of abstractness—can be explained by essentially the same account proposed for the co-occurrence of error types.

The remainder of this chapter presents a brief overview of the reading behavior of deep dyslexics, motivations for a connectionist account, a summary of the Hinton & Shallice results, and a general evaluation of these results that serves to motivate much of the research presented in the thesis.

[8] This research was done in collaboration with Tim Shallice—a more condensed description can be found in Plaut & Shallice (1991a).

2.4 Deep dyslexia

Despite its familiarity as a concept in cognitive neuropsychology, deep dyslexia remains controversial. It was first suggested as a symptom-complex by Marshall & Newcombe (1973), who described two patients (G.R. and K.U.). Both made semantic errors in attempting to read aloud and also made visual and derivational errors. Coltheart (1980a) was able to add another 15 cases. Kremin (1982) added another eight and over ten more are referred to in Coltheart et al. (1987a).

Beginning with the semantic error, Coltheart (1980a) also extended the list of common properties to 12, namely (examples of errors are from D.E., Patterson & Marcel, 1977)

1. Semantic errors (e.g. BLOWING ⇒ “wind”, VIEW ⇒ “scene”, NIGHT ⇒ “sleep”, GONE ⇒ “lost”);
2. Visual errors (e.g. WHILE ⇒ “white”, SCANDAL ⇒ “sandals”, POLITE ⇒ “politics”, BADGE ⇒ “bandage”);
3. Function-word substitutions (e.g. WAS ⇒ “and”, ME ⇒ “my”, OFF ⇒ “from”, THEY ⇒ “the”);
4. Derivational errors (e.g. CLASSIFY ⇒ “class”, FACT ⇒ “facts”, MARRIAGE ⇒ “married”, BUY ⇒ “bought”);
5. Non-lexical derivation of phonology from print is impossible (e.g. pronouncing non-words, judging if two non-words rhyme);
6. Lexical derivation of phonology from print is impaired (e.g. judging if two words rhyme);
7. Words with low imageability/concreteness (e.g. JUSTICE) are harder to read than words with high imageability/concreteness (e.g. TABLE);
8. Verbs are harder than adjectives which are harder than nouns in reading aloud;
9. Functions words are more difficult than content words in reading aloud;
10. Writing is impaired (spontaneous or to dictation);
11. Auditory-verbal short-term memory is impaired;
12. Whether a word can be read at all depends on its sentence context (e.g. FLY as a noun is easier than FLY as a verb).

Given the uniformity of the patients’ symptoms, Coltheart characterized the symptom-complex as a syndrome.

In fact, not all these properties are always observed when an acquired dyslexic patient makes semantic errors in reading. Thus patient A.R. (Warrington & Shallice, 1979) did not show the concreteness and content word effects (7 and 9), and had relatively intact writing and auditory short-term memory (10 and 11). Three other patients have been described who make semantic errors in reading aloud (and do so also when any other speech responses are required) and yet make few if any visual errors (Caramazza & Hillis, 1990; Hillis et al., 1990).[9] The lack of complete consistency across patients therefore led to criticisms of the attempt to characterize the symptom-complex as directly reflecting an impairment to some specific processing component. Some of these arguments were specific to deep dyslexia. Thus Shallice & Warrington (1980) held that deep dyslexia was not a "pure syndrome." Others, though, made more general critiques. Morton & Patterson (1980) and Caramazza (1984; 1986) denied the theoretical utility of generalizing over patients for extrapolation to normal function, and Shallice (1988) more specifically claimed that error patterns did not provide an appropriate basis for this purpose.

Despite these objections to the theoretical utility of the deep dyslexia symptom-complex, Coltheart et al. (1987a) stress that work since 1980 reinforces the virtually complete uniformity of symptom pattern found across a large number of patients. This means that to dismiss deep dyslexia as theoretically irrelevant would be at least as dangerous as to accept it uncritically as the manifestation of some specific impairment. For the present we will leave consideration of these methodological criticisms of deep dyslexia until the General Discussion. We will provisionally assume that it is a valid theoretical concept.

Many other properties of the reading of individual deep dyslexic patients have been recorded. In this thesis we will be particularly concerned with four.

1. *Additional types of reading errors*. Mixed visual-and-semantic (e.g. SHIRT ⇒ "skirt") were recorded in all of the patients reviewed by (Coltheart, 1980a) on whom there is adequate data; in K.F. (Shallice & McGill, 1978) and P.S. (Shallice & Coughlan, 1980) they were also shown to occur at above the rate which one would expect if they were all arising as visual errors or as semantic errors independently. Another error type which was observed even earlier by Marshall & Newcombe (1966) is that of visual-then-semantic errors (e.g. SYMPATHY ⇒ "orchestra", presumably via *symphony*), described in eight of the patients reviewed by (Coltheart, 1980a).

2. *Influences of semantic variables on visual errors*. In general, the abstract/concrete dimension does not just relate to the issue of how successfully different types of words are read. The stimuli on which visual errors occur tend to be more abstract than the responses produced and also more abstract than the stimuli for which other types of responses occur (see e.g. Shallice & Warrington, 1980).

3. *Confidence in errors*. The confidence with which errors are produced has been studied in

[9] One could argue that two of these patients at least are hardly "acquired dyslexics" since their problem is held to be at the phonological output lexicon. This though, presupposes that one can make a clear distinction between reading impairments and other difficulties. Yet, while it remains generally accepted that non-semantic phonological reading procedures are grossly impaired in deep dyslexic patients (see e.g. Marshall & Newcombe, 1973), it has been argued that there are additional deficits in the semantic reading route *and* that these can differ in their location, with some patients even being "output" deep dyslexics (Friedman & Perlman, 1982; Shallice & Warrington, 1980). Thus, the "clear distinction" between reading and non-reading difficulties is absent from the literature.

three patients. P.W. and D.E. (Patterson, 1978) were much more likely to be sure that they were correct for visual errors than for semantic errors, but G.R. gave as high confidence ratings both for visual errors and for semantic errors as for correct responses (Barry & Richardson, 1988).

4. *Lexical decision*. Deep dyslexics can often distinguish words from orthographically regular non-words, even when they are quite poor at explicitly reading the words (Patterson, 1979). Lexical decision was "surprisingly good" for nine of the 11 cases listed by Coltheart (1980a) for which there was data.

Turning to theoretical accounts of the symptom-complex, we will follow Marshall & Newcombe (1973) and many others by presuming that phonological reading procedures are grossly impaired in these patients and that this can account for characteristics (5), (6), and presumably (11) (see discussions in Coltheart, 1980a; Coltheart et al., 1987a). However, if it is held that the complete cluster of properties have a common functional origin, what can it be? The most prosaic possibility is that the syndrome arises from a set of functional deficits which co-occur for anatomical reasons (e.g. Morton & Patterson, 1980; Shallice, 1988; Shallice & Warrington, 1980). If, however, the impairments are only specified in terms of damage to hypothetical subcomponents or transmission routes, many questions remain to be answered. Why do visual and derivational errors so often co-occur with semantic ones? Why do mixed visual-and-semantic and visual-then-semantic errors occur? If the general advantage for concrete words results from impaired access to abstract semantics *per se*, why has only one patient (C.A.V., Warrington, 1981) been observed with superior performance on *abstract* words? How does one account for the effects of concreteness on visual errors? *Ad hoc* explanations have been given for some of these points (see Morton & Patterson, 1980; Shallice & Warrington, 1980) but nothing resembling a well-developed theory along these lines exists.

An interesting version of the "anatomical coincidence" explanation is the claim that deep dyslexic reading reflects reading by the right hemisphere (Coltheart, 1980b; 1983; Saffran et al., 1980). The attraction of this hypothesis is the similarities that have been demonstrated between reading in deep dyslexia and in patients reading with an isolated right hemisphere (e.g. Patterson et al., 1989; Zaidel & Peters, 1981). However, these analogies have been criticized (see e.g. Patterson & Besner, 1984a; Shallice, 1988) and at least one patient has been described with many deep dyslexic characteristics whose reading was abolished after a second *left* hemisphere stroke (Roeltgen, 1987). Overall, while the theory is based on empirical analogues for certain deep dyslexic characteristics (e.g. semantics by which the right hemisphere might produce the symptom-complex), it is principally an attempt to localize rather than to provide a mechanistic account. Since no mechanistic account exists for any other neuropsychological syndrome except for neglect dyslexia (Mozer & Behrmann, 1990), this is hardly a strong criticism of the theory from present-day perspectives. However, an explanation oriented towards this more complex goal remains a major target for understanding deep dyslexia.

2.5 Motivation of a connectionist account

Much of the initial motivation for pursuing a connectionist account of deep dyslexia comes out of preliminary work by Hinton & Sejnowski (1986) on the effects of damage in networks. They were not primarily concerned with modeling deep dyslexia, but rather with investigating how distributed representations can mediate in mapping between arbitrarily related domains (Hinton et al., 1986). The task they chose was a highly simplified version of the mapping orthography to semantics: each of 20 three-letter words was to be associated with an arbitrary semantics consisting of a random subset of 30 semantic features. The network used to accomplish the mapping had three layers of units. Thirty "grapheme" units, in three groups of 10, represented the three letters of each word. These units were fully connected to 20 "intermediate" units, which in turn were fully connected to 30 "sememe" units, one for each semantic feature. In addition, the sememe units were fully interconnected. The units produced stochastic binary output and all connections were symmetric. The network was trained with the Boltzmann Machine learning procedure (Ackley et al., 1985) to settle into the correct pattern of activity over the sememe units for each word when the grapheme units for the letters of the word were clamped on.

The undamaged network performed the task almost perfectly, but when single intermediate units were removed, 1.4% of the responses of the network were incorrect. Interestingly, 59% of these incorrect responses were the exact semantics of an alternative word, and these "word" errors were more semantically and visually similar to the correct word than would be expected by chance. Assuming that the pattern of semantic activity would be the basis for an overt naming response, a single locus of damage in the network produced semantic, visual, and mixed reading errors.

Hinton & Sejnowski interpret this behavior in the following way. Interactions among the sememe units enable them to "clean-up" an initially noisy or incomplete pattern of semantic activity into the pattern corresponding to the exact semantics of the input word. Under normal operation this initial pattern is always closer to the semantics of the correct word than to that of any other, and so the clean-up interactions produce a correct response. However, the damaged network occasionally produces an initial pattern of semantic activity that is closer to the meaning of another word, usually one that shares letters and/or semantic features with the correct word. When semantic clean-up is applied in this case, the network produces the exact semantics of the incorrect word, resulting in a "word" error that tends to be semantically and/or visually related to the correct word. While Hinton & Sejnowski's demonstration is highly oversimplified, it is quite suggestive that damage to networks that map from orthography to semantics can produce a pattern of errors qualitatively similar to that of deep dyslexics.

2.6 A preliminary connectionist model of deep dyslexia

Based on this promising initial work, Hinton & Shallice (1991, hereafter H&S) undertook to model deep dyslexia more thoroughly. Developing the model involved making four sets of design decisions that apply to the development of any connectionist simulation:

- *The task:* What input/output pairs is the network trained on and how are they represented as patterns of activity over groups of input and output units?
- *The network architecture:* What type of unit is used, how are the units organized into groups, and in what manner are the groups connected?
- *The training procedure:* How are examples presented to the network, what procedure is used to adjust the weights to accomplish the task, and what is the criterion for halting training?
- *The testing procedure:* How is the performance of the network evaluated—specifically, how are lesions carried out and how is the behavior of the damaged network interpreted in terms of overt responses that can be compared with those of patients?

The following four subsections describe the characteristics of the model in terms of each of these issues. The adequacy and limitations of these decisions are then discussed and serve to motivate the simulations presented in this thesis.

2.6.1 The task

H&S defined a version of the task of mapping orthography to semantics that is somewhat more sophisticated (although still far from realistic) than that used by Hinton & Sejnowski. Orthography was represented in a similar way, in terms of groups of position-specific letter units (McClelland & Rumelhart, 1981). In order to keep the task simple, 40 three- or four-letter words were chosen with restrictions on what letters could occur in each position, resulting in a total of 28 possible graphemes (see Table 2.1).

Rather than assign to each word a completely arbitrary semantics, H&S designed a set of 68 semantic features intended to capture intuitive semantic distinctions (see Table 2.2). On average, about 15 of the 68 features were present in the semantic representation of a word. The words were chosen to fall within five concrete semantic categories: indoor objects, animals, body parts, foods, and outdoor objects. The assignment of semantic features to words ensured that, in general, objects in the same category tended to be more similar (i.e. shared more features) than objects in different categories (see Figure 2.6). However, H&S did not directly demonstrate that their semantic categories faithfully reflect the actual semantic similarity among words. Figure 2.6 conveys some sense of the similarity within and between categories, but a more direct impression can be obtained from a full display of the similarity (i.e. proximity in semantic space) of each

Letters allowed in each position			
B C D G H L M N P R T	A E I O U	B C D G K M P R T W	E K

Words in each category				
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

Table 2.1: The words used by H&S, organized into categories.

Semantic features					
1	max-size-less-foot	21	indoors	46	made-of-metal
2	max-size-foot-to-two-yards	22	in-kitchen	47	made-of-wood
3	max-size-greater-two-yards	23	in-bedroom	48	made-of-liquid
4	main-shape-1D	24	in-livingroom	49	made-of-other-nonliving
5	main-shape-2D	25	on-ground	50	got-from-plants
6	cross-section-rectangular	26	on-surface	51	got-from-animals
7	cross-section-circular	27	otherwise-supported	52	pleasant
8	has-legs	28	in-country	53	unpleasant
9	white	29	found-woods	54	man-made
10	brown	30	found-near-sea	55	container
11	green	31	found-near-streams	56	for-cooking
12	color-other-strong	32	found-mountains	57	for-eating-drinking
13	varied-colors	33	found-on-farms	58	for-other
14	tranparent	34	part-of-limb	59	used-alone
15	dark	35	surface-of-body	60	for-breakfast
16	hard	36	interior-of-body	61	for-lunch-dinner
17	soft	37	above-waist	62	for-snack
18	sweet	38	mammal	63	for-drink
19	tastes-strong	39	wild	64	particularly-assoc-child
20	moves	40	fierce	65	particularly-assoc-adult
		41	does-fly	66	used-for-recreation
		42	does-swim	67	human
		43	does-run	68	component
		44	living		
		45	carnivore		

Table 2.2: Semantic features used by H&S. Features within a block were considered "closely related" for the purposes of determining the network architecture.

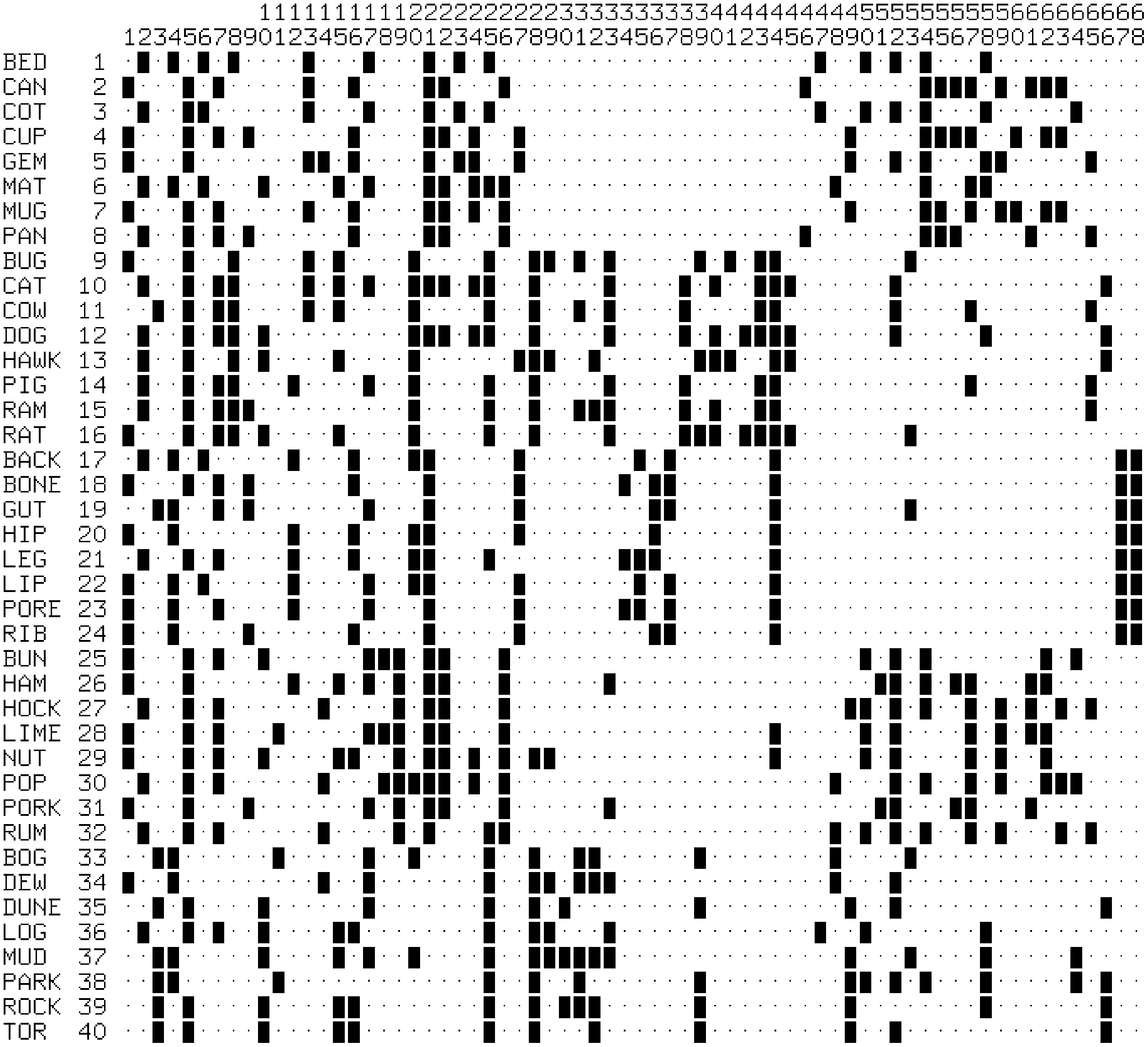

Figure 2.6: The assignment of semantic features to words used by H&S. A black rectangle indicates that the semantic representation of the word listed on the left contains the feature whose number (from Table 2.2) is listed at the top.

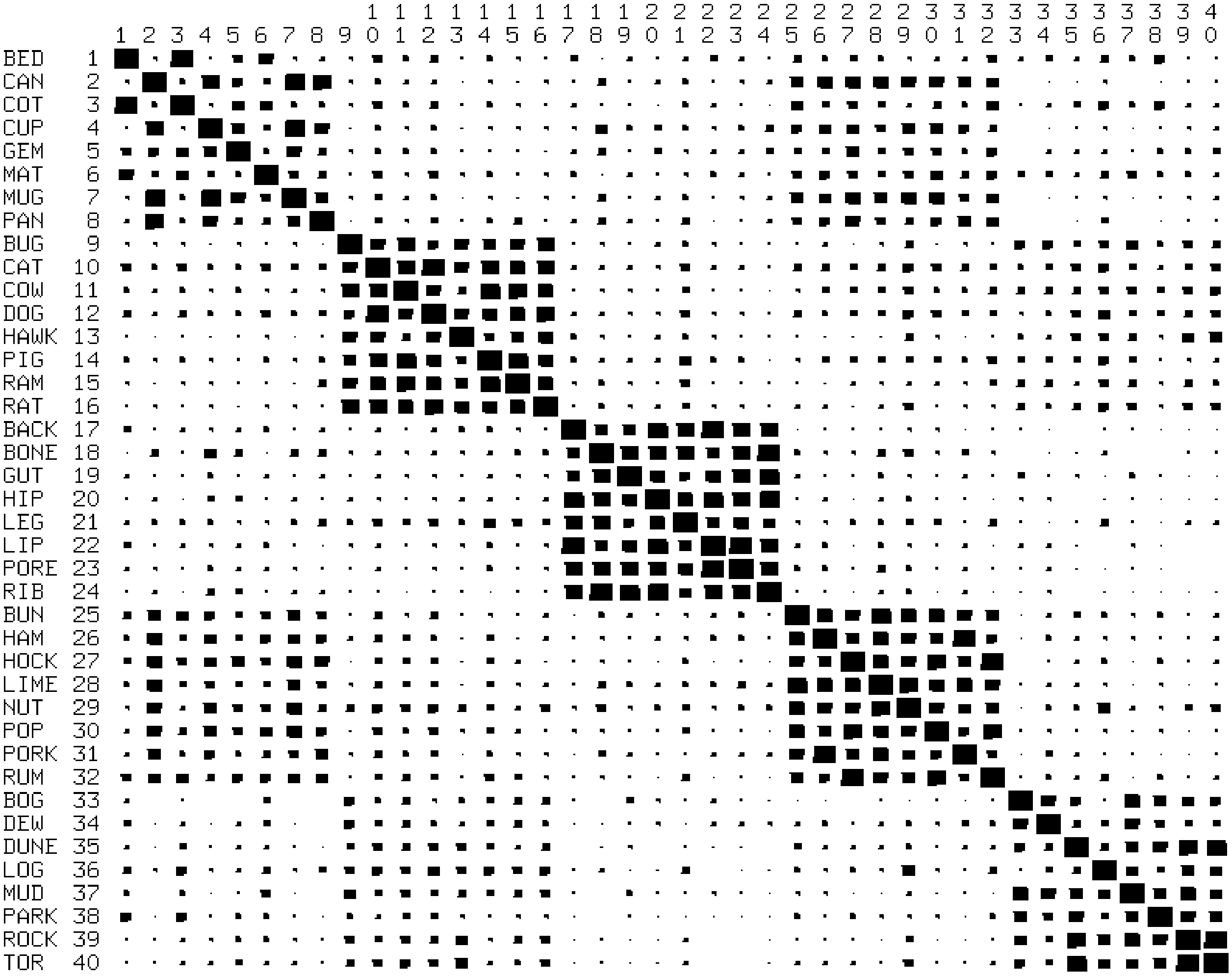

Figure 2.7: The similarity matrix for the semantic representations of words. The size of each square represents the proximity of the representations of a pair of words, where the largest squares (along the diagonal) represent the closest possible proximity (1.0) and a blank square represents the farthest possible proximity (0.0).

pair of words, shown in Figure 2.7. Because the words are ordered by category, the extent and uniformity of the similarity within each category is reflected by an 8-by-8 block along the diagonal of the matrix, while between-category similarity is reflected in off-diagonal blocks. A number of interesting characteristics are apparent from the similarity matrix. Words for "body parts" are quite similar to each other and quite different from words in other categories. In contrast, "indoor objects" are not uniformly similar to each other, and many are quite similar to "foods," particularly those that are used with food (i.e. CUP, CAN, MUG, PAN). "Outdoor objects" also vary considerably in their similarities with each other, and are often also similar to "animals" (which are also found outdoors). However, the overall strength of the five on-diagonal blocks supports the use of category membership as a general measure of semantic similarity.

A further requirement of a satisfactory approximation of the task of mapping orthography to semantics that H&S did not verify for their representations is that the relationship between the visual and semantic representations of a word is arbitrary in general. In other words, the visual similarity of two words (as defined below) provides no information about their semantic similarity, and *vice versa*. While this holds for the full set of English morphemes, it is possible that visual similarity is somewhat predictive of semantic similarity for the representations of the particular 40 words used by H&S. If this were the case, it would compromise their arguments about the expected "chance" rates of mixed visual-and-semantic errors assuming visual and semantic influences operate *independently*. One way to test the independence of visual and semantic similarity is that the probability of a randomly selected word pair being both visually and semantically similar, m, should be approximately equal to the product of the independent probabilities of visual, v, and semantic, s, similarity. Among all possible non-identical word pairs in the H&S word set, $m = .062$, $v = .36$, and $s = .18$, so $vs = .065$ is roughly equal to m. Thus visual and semantic similarity are approximately independent in the H&S word set.

2.6.2 The network

Unlike the binary stochastic units and symmetric connections used by Hinton & Sejnowski, H&S used real-valued deterministic units and one-way connections. The 28 grapheme units were connected to a group of 40 intermediate units, which in turn were connected to the 68 sememe units. In order to reduce the number of connections, only a random fourth of the possible connections were included.

Following Hinton & Sejnowski's argument for the importance of allowing the sememe units to interact, H&S introduced connections at the semantic level in two ways. First, they added direct connections between sememe units. Rather than include all possible 4624 such connections, only sememe units that represent closely related features (defined in Table 2.2) were connected. While these direct connections help the network ensure that sememes are locally consistent, not all relationships among semantic features can be encoded by pairwise interactions alone. In order

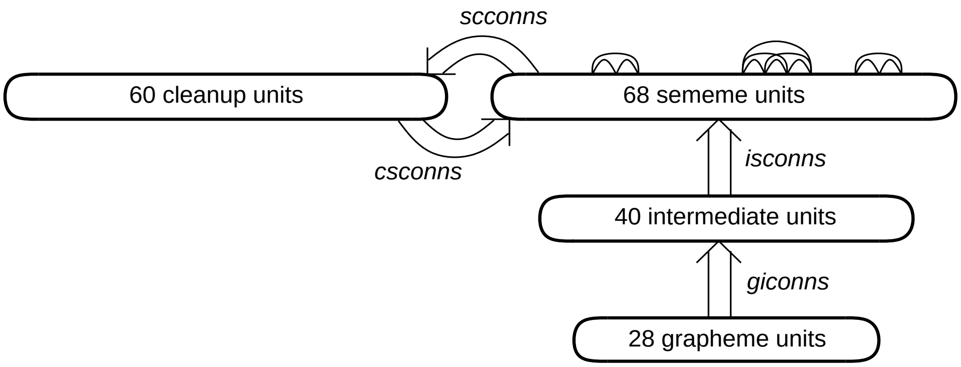

Figure 2.8: The network used by H&S. Notice that sets of connections are named with the initials of the names of the source and destination unit groups (e.g. *giconns* for grapheme-to-intermediate connections).

to allow *combinations* of sememes to directly influence each other, H&S also introduced a fourth group of 60 “clean-up” units that receive connections from, and send connections to, the sememe units. This pathway can enforce more global consistency among semantic features. As in the “direct” pathway from graphemes to sememes via the intermediate units, only a random fourth of the possible connections in this clean-up pathway were included. The resulting network, depicted in Figure 2.8, had about 3300 connections.

2.6.3 The training procedure

The network was trained in the following way. The grapheme units were set to the appropriate input pattern for a word, and all other units were set to 0.2. The network was then run for seven iterations in which each unit updated its state once per iteration, generating a pattern of activity over the sememe units. The network was initialized to have small random weights, so that at the beginning of training the pattern of semantic activity produced by the word was quite different from its correct semantics. An iterative version of the back-propagation learning procedure, known as “back-propagation through time” (Rumelhart et al., 1986b, see Appendix 10), was used to compute the way that each weight in the network should change so as to reduce this difference for the last three iterations. These weight changes were calculated for each word in turn, at which point the accumulated weight changes were carried out and the procedure was repeated. After about 1000 sweeps through the 40 words, when the network was presented with each word, the activity of each sememe unit was within 0.1 of its correct value for that word, at which point training was considered complete.

2.6.4 The testing procedure

After training, the intact network produced the correct semantics of each word when presented with its orthography. The network was then "lesioned" in three ways: (a) *ablation*: removing a random subset of the units in a layer, (b) *disconnection*: removing a subset of the connections between layers, and (c) *noise*: adding uniformly distributed random noise to the weights on connections between layers. Under damage, the semantics produced by a word typically differed somewhat from the exact correct semantics. Yet even though the corrupted semantics would fail the training criteria, it still might suffice for the purposes of naming. H&S defined two criteria that had to be satisfied in order for the damaged network to be considered to have made a response:

1. A *proximity* criterion ensured that the corrupted semantics was sufficiently close to the correct semantics of some word. Specifically, the cosine of the angle (i.e. normalized dot product) between the semantic vector produced by the network and the actual semantic vector of some word (in the 68-dimensional space of sememes) had to be greater than 0.8.[10]
2. A *gap* criterion ensured that no other word matched nearly as well. Specifically, the proximity to the generated semantics of the best matching word had to be at least 0.05 larger than that of any other word.

If either of these criteria failed, the output was interpreted as an omission; otherwise the best matching word was taken as the response, which either could be the correct word or an error.

In order to compare the behavior of the network under damage with that of deep dyslexics, H&S systematically lesioned sets of units or connections over a range of severity. For 10 instances of each lesion type, all 40 words were presented to the network and omission, correct, and error responses were accumulated. Figure 2.9 shows the overall correct response rate for lesions of each main set of connections over a range of severity.[11] As an approximation to the standard error classification used for patients (cf. Morton & Patterson, 1980), an error was defined to be visually similar to the input word if the two words overlapped in at least one letter, and semantically similar if the two words belonged to the same category. Errors were then classified into four types:

- *visual* (V): responses that are visually (but not semantically) similar to the stimulus (e.g. CAT ⇒ "cot").

[10] A number of distance metrics defined for vectors in the unit hypercube are reasonable candidates for comparing the network's output with known responses. Possibilities other than angle cosine include correlation, hamming distance, euclidean distance, and norms for powers higher than 2. Many of these metrics behave similarly. H&S chose angle cosine over the more familiar metric of euclidean distance because not all directions of difference between two vectors would be equally disruptive to an output system. In particular, they argued that differences in direction are more significant than differences in magnitude (which maintain the *relative* levels of unit activity).

[11] To make it easier to interpret this and subsequent graphs we will adopt the convention of using "closed" markers (i.e. dot and asterisk) for sets of connections in the direct pathway, "open" markers (i.e. square and diamond) for sets of connections in the clean-up pathway, and "line" markers (i.e. plus and cross) for any other sets of connections.

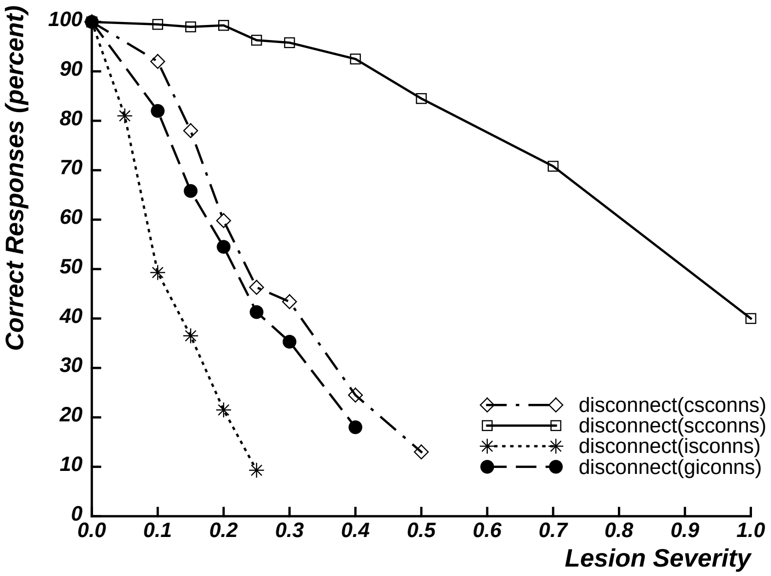

Figure 2.9: Overall correct response rate of the H&S model for "disconnections" of a range of severity of each main set of connections. See Figure 2.8 for the meaning of "csconns," etc.

- *semantic* (S): responses that are semantically (but not visually) similar to the stimulus (e.g. CAT ⇒ "dog").
- *mixed visual-and-semantic* (V+S): responses that are both visually and semantically similar to the stimulus (e.g. CAT ⇒ "rat").
- *other* (O): responses that are unrelated to the stimulus (e.g. CAT ⇒ "mug").

Table 2.3 shows the distribution of error types for all types of lesions, summed over instances which resulted in between 25–75% correct responses. The most important result is that all lesions produced semantic, visual-and-semantic, and visual errors at rates higher than would be expected by chance (with the sole exception of the lesion type most resistant to damage). "Chance" is determined by comparing the ratio of each error rate to that of "other" errors with the predicted ratio under the assumption that error responses are generated randomly from the word set. Also, for all but one lesion type—disconnect(*isconns*)—the number of mixed visual-and-semantic errors was greater than would be expected if visual and semantic similarity were caused independently. Furthermore, the network showed a greater tendency to produce visual errors with early damage (closer to the graphemes) and semantic errors with later damage (closer to the sememes) although even damage completely within the semantic clean-up system produced an above-chance rate of visual errors. It is clear that these errors were not produced randomly because then there would have been a high rate of "other" errors (based on the distribution of possible error types), whereas all errors produced by clean-up damage were either visual, visual-and-semantic, or semantic.

H&S also demonstrated that, even when the semantics produced by the system were insufficient to plausibly drive a response system, enough information was often available to make between- and

	Overall		Conditional probabilities			
	Error Rates			Vis&		
Lesion	n	Rate	Vis	Sem	Sem	Other
disconnect(*giconns*)	4	4.8	34.2	44.7	13.2	7.9
noise(*giconns*)	4	3.9	46.0	27.0	20.6	6.3
ablate(*intermediate*)	3	3.1	24.3	45.9	24.3	5.4
disconnect(*isconns*)	2	3.4	11.1	29.6	55.6	3.7
noise(*isconns*)	3	2.4	24.1	48.3	20.7	6.9
disconnect(*scconns*)	2	0.2	—	100.0	—	—
noise(*scconns*)	4	1.8	6.9	72.4	20.7	—
ablate(*cleanup*)	2	3.4	7.4	63.0	25.9	—
disconnect(*csconns*)	3	3.4	34.1	31.7	34.1	—
noise(*csconns*)	2	2.3	27.8	38.9	33.3	—
Chance			29.9	6.2	11.8	52.2

Table 2.3: The distribution of error types produced by lesions of all types and locations that resulted in 25–75% correct performance in the H&S model. "*n*" refers to the number of lesion severities producing performance falling within the 25–75% range, and "Rate" is the average percentage of word presentations producing explicit error responses for these lesions. "Chance" refers to the distribution of error types if responses were chosen from the word set at random. Notice that there were few if any "Other" errors with many of the lesions even though more than 50% of the possible error response are of this type.

within-category discriminations. For instance, removing all of the connections from the sememe to clean-up units reduced explicit correct performance to 40%. However, of the 60% remaining trials producing an omission, 91.7% of these resulted in semantics that were closer to the centroid of the correct category than to that of any other category (chance is 20%), and 87.5% were closer to the semantics of correct word in that category than to that of any other word in the category (chance is 12.5%). The effect was weaker with earlier damage: removing 30% of the grapheme-to-intermediate connections produced 35.3% correct performance with 48.3% between-category and 49.0% within-category discrimination on omission trials.

Finally, a peculiar and interesting effect emerged when the connections from the clean-up to sememe units were lesioned. The network showed a significant selective preservation of words in the "foods" category (75% correct) relative to those in other categories (next best, 34% correct).[12] The effect was quite specific; it did not occur for other lesions in the network, nor for the same lesion in a second version of the network trained with different initial random weights.

2.6.5 Attractors

An important concept in understanding these results is that of an "attractor." The sememe units in the H&S network change their states over time in response to a particular orthographic input. The initial pattern of semantic activity generated by the direct pathway may be quite different from the exact semantics of the word. Interactions among sememe units, either directly via intra-sememe connections or indirectly via the clean-up units, serve to gradually modify and "clean-up" the initial pattern into the final, correct pattern. As described in the Introduction, this process can be conceptualized in terms of movement in the 68-dimensional space of possible semantic representations, in which the state of each sememe unit is represented along a separate dimension. Since unit states are bounded between 0 and 1, this space is actually a hypercube. At any instant in processing a word, the entire pattern of activity over the sememe units correspond to a particular point in semantic space. The exact meanings of familiar words correspond to other points in the space—to be precise, particular corners of the hypercube. The states of sememe units change over time in such a way that the point representing the current pattern of semantic activity "moves" to the point representing the nearest familiar meaning. In this way, familiar meanings are attractors in the space of semantic representations.

H&S offer an intuitive explanation for co-occurrence of visual and semantic influences on errors in terms of the effects of damage in a network that builds attractors in mapping between two arbitrarily related domains. Connectionist networks have difficulty learning to produce quite different outputs from very similar inputs, yet very often visually similar words (e.g. CAT and COT) have quite different meanings. One effective way a network can accomplish this mapping is to

[12] This effect was significant at the 0.01 level and not at the 0.1 level as incorrectly stated in Hinton & Shallice (1991).

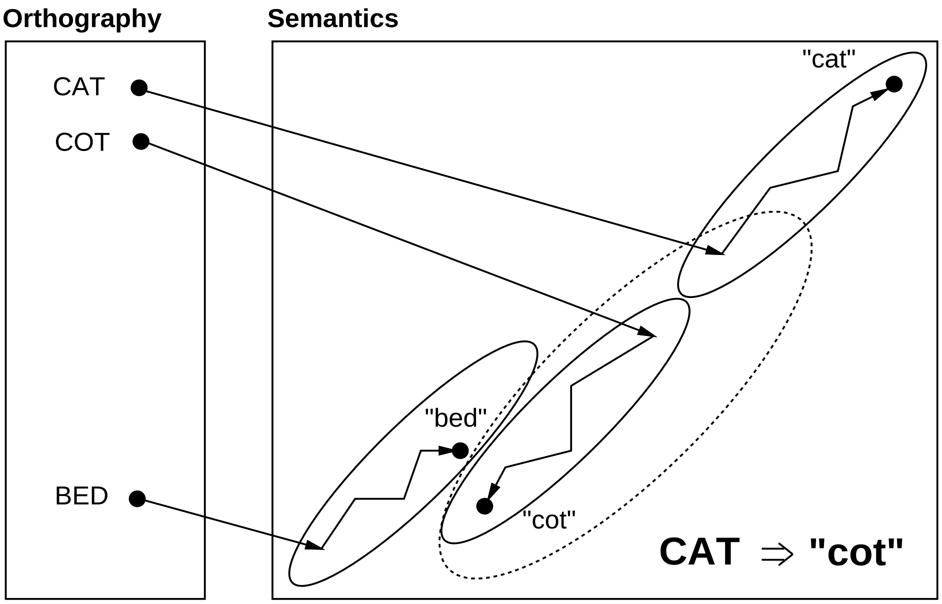

Figure 2.10: How damage to semantic attractors can cause visual errors. The solid ovals depict the normal basins of attraction; the dotted one depicts a basin after semantic damage.

construct large basins of attraction around each familiar meaning, such that any initial semantic pattern within the basin will move to that meaning (see Figure 2.10). Visually similar words are then free to generate fairly similar initial semantic patterns as long as they each manage to fall somewhere within the appropriate basin of attraction. In this way the network learns to shape and position the basins so as to “pull apart” visually similar words into their final distinct semantics. Damage to the semantic clean-up distorts these basins, occasionally causing the normal initial semantic pattern of a word to be “captured” within the basin of a visually similar word. Essentially, the layout of attractor basins must be sensitive to both visual and semantic similarity, and so these metrics are reflected in the types of errors that occur as a result of damage.

2.7 Evaluation of the model

The aim of H&S’s work was to provide a unified account of the nature and co-occurrence of semantic, visual and mixed reading errors in deep dyslexia. Most previous explanations of why virtually all patients who make semantic errors also make visual errors (e.g. Gordon et al., 1987; Morton & Patterson, 1980) have had to resort to proposing lesions at multiple locations along the semantic route. Shallice & Warrington (1980) speculated that an inability to adequately access part of the semantic system might give rise to the occurrence of errors. However, H&S actually demonstrated that all of these error types arise naturally from single lesions anywhere in a connectionist network that builds attractors in mapping orthography to semantics. Only the quantitative distribution of error types varied systematically with lesion location.

There are two main types of criticism leveled against the H&S model. The first has to do with the limited range of empirical phenomena it addresses. Of the aspects of deep dyslexia which pose problems for theory, only three were modeled—the very existence of semantic errors in reading aloud, the frequent co-occurrence of visual errors with semantic errors, and the relatively high rates of occurrence of mixed visual-and-semantic errors. However, an adequate theory of deep dyslexia would also need to account for a fair number of other aspects of the syndrome. Certain aspects—(5), (6) and (10) of Section 2.4—involve difficulties in mapping directly between print and sound and are covered by the assumption of the gross impairment in the operation of the non-semantic route(s). Two others—(3) function word substitutions and (4) derivational errors—can interpreted as special cases of semantic or mixed visual-and-semantic errors, and so can be explained in the way that these errors are (see Funnell, 1987). Another two—(11) auditory short-term memory impairments and (12) context effects—are dismissed by Coltheart et al. (1987a) as too vague. However, this still leaves (7) the effects of imageability on reading, (8) and (9) the effects of part-of-speech, and also a number of the additional effects—the interactions between the abstract/concrete dimension and visual errors, confidence ratings, lexical decision, and the visual-then-semantic errors. These phenomena will all be considered directly in this thesis. One final effect, the impaired writing, will be addressed in the General Discussion.

The second type of criticism of the H&S model relates to its generality. Most attempts to model acquired dyslexia by lesioning connectionist networks (Mozer & Behrmann, 1990; Patterson et al., 1990) have been based on pre-existing models of word reading in normals (Mozer, 1988; Seidenberg & McClelland, 1989). These studies have primarily aimed to provide independent validation of the properties of the normal models that enable them to reproduce phenomena they were not initially designed to address. The work of H&S is rather different in nature in that they were not concerned with supporting a particular model of normal word comprehension. Rather, H&S had the more general goal of investigating the effects of damage in a fairly general type of network in the domain of reading via meaning. To the extent that the behavior of the damaged network mimicked that of deep dyslexics, the principles that underly the network's behavior may provide insight into the cognitive mechanisms of reading in normals, and their breakdown in patients. In this way, the relevance of H&S's simulations to cognitive neuropsychology depends on identifying and evaluating those aspects of the model which are responsible for its ability to reproduce patient behavior.

H&S argue that the co-occurrence of different error types obtained in deep dyslexia is a natural consequence of lesioning a connectionist network that maps orthography to semantics using attractors. However, their conclusions were essentially based on a single type of network that inevitably had many specific features. It was only an assumption that these specific features did not significantly contribute to the overall behavior of the network under damage. Clearly it is impossible to evaluate every possible aspect of the model. H&S attempt to motivate and justify many of the decisions that went into developing their model. In considering the significance of these decisions, it is important to bear in mind that they each reflect a tradeoff between (at least) three types of constraint: (a) empirical data from cognitive psychology and neuropsychology, (b) principles of what connectionist networks find easy, difficult or impossible to do, and (c) limitations of the computational resources available for running simulations. Each of the following major design issues serves to motivate the investigations described in a subsequent chapter.

2.7.1 The task

The grapheme and sememe representations used by H&S clearly fail to reflect the full range of orthographic and semantic structure in word reading. The use of position-specific letter units, the selection of semantic features, and their assignment to words, was based more on computational than empirical grounds. However, these representations exhibit the characteristics that are essential for demonstrating the influences of both visual and semantic similarity on deep dyslexic reading: (a) visually similar words (with overlapping letters) have similar representations over the grapheme units, (b) semantically similar words (in the same category) have similar representations over the sememe units, and (c) there is no systematic relationship between the orthographic and semantic representations of a word.

One concern involves the adequacy of the definitions of visual and semantic similarity. These were chosen to be analogous to those used for patients, but they only approximate the actual similarity structure of the visual and semantic representations used for words. The impact of the adequacy of this approximation on the error pattern produced under damage was not evaluated.

A more severe limitation is that the model was trained on only 40 words, allowing only a very coarse approximation to the range of visual and semantic similarity among words in a patient's vocabulary. In particular, important variables known to affect patients' reading behavior, such as word length, frequency, syntactic class, and imageability/concreteness, were not manipulated. In addition, there is the potential problem that some of the observed effects may arise from operations of a small subset of the stimulus set with statistically unusual properties. The general impact of this limitation will be addressed in the General Discussion. More specifically, simulations presented in Chapter 6 attempt to extend the H&S approach to account for effects of concreteness in deep dyslexic reading performance.

2.7.2 The network

H&S provide only a general justification for the network architecture they chose. Hidden units are needed because the problem of mapping orthography to semantics is not linearly separable. Recurrent connections are required to allow the network to develop semantic attractors, whose existence constitutes the major theoretical claim of the work. The choices of numbers of intermediate and clean-up units, restrictions on intra-sememe connections, and connectivity density were an attempt to give the network sufficient flexibility to solve the task and build strong semantic attractors, while keeping the size of the network manageable. Some aspects of the design, particularly the selective use of intra-sememe connections, were rather inelegant and *ad hoc*. Chapter 4 elaborates on the implications of these distinctions and describes simulations involving a range of network architectures that attempt to directly evaluate their impact on the pattern of errors produced under damage.

2.7.3 The training procedure

H&S justify the use of an admittedly implausible learning procedure in two ways. The first is to emphasize that they were not directly concerned with simulating aspects of reading *acquisition*, but only its breakdown in mature, skilled readers. Thus the learning procedure can be viewed solely as a programming technique for determining a set of weights that is effective for performing the task. The second justification they use is to point out that back-propagation is only one of a number of ways of performing gradient descent learning in connectionist networks. Other more plausible gradient descent procedures, such as contrastive Hebbian learning in deterministic Boltzmann Machines (Hinton, 1989b; Peterson & Anderson, 1987), are more computationally intensive than back-propagation but typically develop similar representations. In Chapter 5 we present simulations that attempt to replicate and extend the H&S results using a deterministic Boltzmann Machine and a closely related stochastic GRAIN network (McClelland, 1990; 1991).

2.7.4 The testing procedure

Perhaps the most serious limitation of H&S's work involves the use of proximity and gap criteria in determining the response produced by the network under damage. These criteria were intended to approximate the requirements of a system that would actually generate responses based on semantic activity. H&S provided evidence that the main qualitative effects obtained do not depend on specific values for these criteria, but their adequacy as an approximation to an output system was left unverified.

Ideally, the response criteria would be replaced by extending the network to produce an actual phonological response. This response could then be compared directly with the oral responses of patients. Unfortunately, preliminary attempts to implement such an output system produced a high rate of phonological "blends" (literal paraphasias) under damage to the input network, which are almost never produced by deep dyslexics. Chapter 3 illustrates this problem and presents simulations that overcome it, allowing explicit phonological responses to replace the criteria H&S used to evaluate the effects of lesions.