

Double Dissociation Without Modularity: Evidence from Connectionist Neuropsychology

David C. Plaut

Department of Psychology
Carnegie Mellon University
Pittsburgh, PA 15213-3890
plaut@cmu.edu

Journal of Clinical and Experimental Neuropsychology, 1995, 17, 291-321.

Abstract

Many theorists assume that the cognitive system is composed of a collection of encapsulated processing components or modules, each dedicated to performing a particular cognitive function. On this view, selective impairments of cognitive tasks following brain damage, as evidenced by double dissociations, are naturally interpreted in terms of the loss of particular processing components. By contrast, the current investigation examines in detail a double dissociation between concrete and abstract word reading after damage to a connectionist network that pronounces words via meaning and yet has no separable components (Plaut & Shallice, 1993). The functional specialization in the network that gives rise to the double dissociation is not transparently related to the network's structure, as modular theories assume. Furthermore, a consideration of the distribution of effects across quantitatively equivalent individual lesions in the network raises specific concerns about the interpretation of single-case studies. The findings underscore the necessity of relating neuropsychological data to cognitive theories in the context of specific computational assumptions about how the cognitive system operates normally and after damage.

*This paper developed out of a presentation given as part of a Workshop on Modularity and the Brain sponsored by the Institute for Advanced Studies at Hebrew University in Jerusalem. I would like to express my appreciation to the Institute for the opportunity and financial support to participate in the workshop. In addition to the Institute, I'd like to thank the McDonnell-Pew Program in Cognitive Neuroscience (Grant T89-01245-016), the National Science Foundation (Grant ASC-9109215), and the National Institute of Mental Health (Grant MH47566) for providing financial support for this work. I'd also like to acknowledge Tim Shallice for his contribution to the larger collaborative project that forms the context for the work described in this paper. Finally, I'd like to thank Marlene Behrmann, Asher Koriat, and an anonymous reviewer for helpful comments on an earlier draft of the paper. Address correspondence to: David C. Plaut, Ph. D., Department of Psychology, Carnegie Mellon University, Pittsburgh, PA 15213-3890, U.S.A. Email: plaut@cmu.edu.

Cognitive neuropsychology attempts to uncover the structure of the human cognitive system by studying the patterns of impaired and preserved cognitive abilities of brain-damaged patients. In this endeavor, perhaps the most powerful weapon in the neuropsychologist's armamentarium is the *double dissociation* (Kinsbourne, 1971; Teuber, 1955). A single or one-way dissociation occurs when, as a result of brain damage, a patient's performance is significantly more impaired on one cognitive task than on another. For example, although neurologically intact individuals have no difficulty pronouncing written words regardless of whether they have a concrete meaning (e.g., TABLE) or an abstract meaning (e.g., TRUTH), after a severe left-hemisphere stroke, patient PW (Patterson & Marcel, 1977) correctly pronounced 67% of concrete words but only 13% of abstract words.

The finding that concrete and abstract words are differentially susceptible to brain damage suggests that they are represented separately in the brain. An alternative view, however, is that PW's brain damage has affected concrete and abstract representations equally, but that abstract words are more impaired because they are inherently more difficult to pronounce (although still well within the abilities of normals). On this latter account, one would not expect to see the opposite dissociation: better reading of abstract than concrete words following brain damage. However, this is exactly what Warrington (1981) observed in a patient CAV with a left-hemisphere tumor: he read correctly 36% of concrete words but 55% of abstract words.

Together, PW and CAV constitute a double dissociation of concrete and abstract word reading.¹ In a similar way, double dissociations among brain-damaged patients have been identified for many other pairs of tasks, across a wide range of specificity in the cognitive system. These include auditory verbal short-term memory versus long-term memory (see Vallar & Shallice, 1990), episodic versus semantic memory (see Tulving, 1983), language comprehension versus production and syntax versus semantics (see Caplan, 1992), sublexical versus lexical processing in reading and in writing (see Coltheart, Patterson, & Marshall, 1980; Patterson, Coltheart, & Marshall, 1985; Shallice, 1988), visual recognition of words versus faces versus other objects (see Farah, 1990), and naming pictures of natural kinds versus artifacts (see Warrington & Shallice, 1984). In each of these cases, a natural interpretation of the pattern of results is that the different tasks are subserved by separate neural mechanisms, such that these mechanisms can be selectively impaired by brain damage. For instance, Warrington (1981) states, "the only plausible interpretation of a double dissociation between abstract-word deficit and concrete-word deficit . . . is that the functional and structural organization of semantic representations of words is categorical" (p. 185). That is to say, the semantics of concrete words and those of abstract words must be represented separately in the brain. The logic of this interpretation dovetails well with the view that the cognitive system is composed of a collection of relatively independent processing components or modules, each dedicated to performing a particular cognitive function (Chomsky, 1980; Coltheart, 1985; Fodor, 1983; Marr, 1982; Morton, 1981). In fact, double dissociations and modularity fit together so naturally that this theoretical perspective has completely dominated the field of cognitive neuropsychology. As Ellis (1987) has put it, "There can be no argument with the fact of modularity, only about its nature and extent" (p. 402).

It is important to bear in mind, though, that different authors use the terms "modules" and "modularity" in the service of very different claims about the structure of the cognitive system. For example, as defined by Fodor (1983), modules are domain specific, innately specified, hard-wired, autonomous, non-assembled, and most critically, informationally encapsulated. By contrast, Coltheart (1985) states explicitly,

I am not using the term *module* in the sense adopted by Fodor (1983). . . . the model I describe . . . would be regarded as modular in character by information-processing theorists, but its components are not modules in Fodor's sense—for example, they are not innate, not informationally encapsulated, and not hard-wired. (p. 4)

Nonetheless, there is still a common thread that runs through the various usages of the term "module": that the cognitive system is composed of components, that the function of each component can be characterized independently of the functions of other components, and that these components can be selectively impaired by brain damage. This is the interpretation of modularity that is adopted in the current paper. Notice that not all forms of differentiation of function qualify as modular on this view. If, as Fodor (1983) suggests, "the condition for successful science . . . is that nature should have joints to carve it at" (p. 128), the modularity hypothesis says that, fortunately for cognitive science, the cognitive system (or at least most of it) has *very* clear joints.

¹For the purpose of making theoretical inferences from the data, it is important not only that the effects were in opposite directions for the two patients, but that PW read concrete words better than CAV, while the reverse is true for abstract words. If it were the case that one patient performed worse than the other on both tasks, the double dissociation might be due simply to one task being more sensitive to the available resources of a process common to both tasks (see Shallice, 1988, Chapter 10).

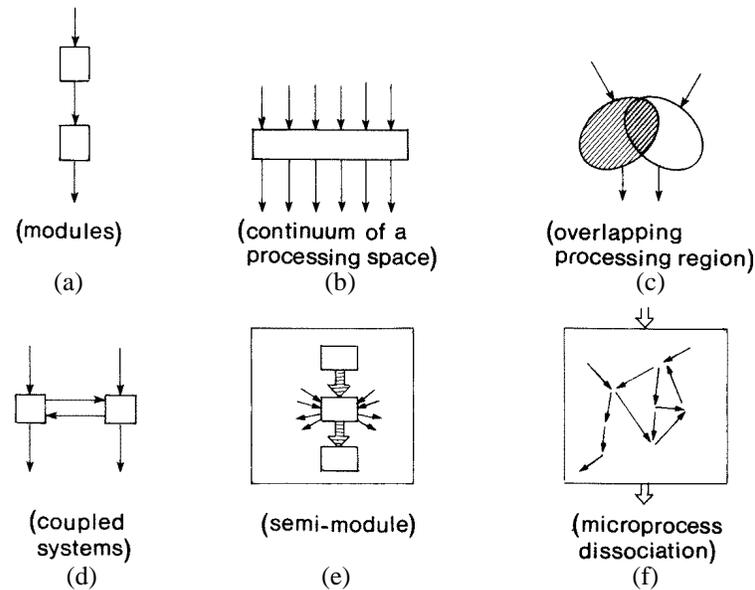


Figure 1: Depictions of six types of system capable of producing double dissociations when damaged (from Shallice, 1988, p. 250). (a) *Traditional modules*. (b) *Continuum of a processing space*. For example, damage to the retina at 9 vs. 15 degrees eccentricity would selectively impair visual processing at these locations without there being separate modules for each. (c) *Overlapping processing regions*. If the processing regions subserving two tasks partially overlap, selective damage to the nonoverlapping portions would doubly dissociate the tasks, but neither of the two regions nor any of their subregions constitute separate modules. (d) *Coupled systems*. If two subsystems (e.g., visual and auditory lexicons) were tightly coupled so that the two could not process contradictory information concurrently, but each could operate correctly in isolation, they would be *isolable* (Shallice, 1979) without being modules. (e) *Semi-modules*. Subsystems can be modules *to some degree*, depending on the extent to which its correct performance depends on variables within the system itself vs. influences from other subsystems. (f) *Multi-level systems*. Two tasks may depend on selectively impairable properties of the same physical system at different levels of description (e.g., learning may depend on some neurochemical whereas mature performance may depend on the structural integrity of the system).

Adopting a modular view of the cognitive system provides a natural way of describing many of the interesting and often counterintuitive double dissociations among cognitive tasks that have been identified among brain-damaged patients. Some researchers, however, seem to take the further step of interpreting the occurrence of double dissociations as *implying* the existence of separate components, each dedicated to performing one of the dissociated tasks (e.g., Warrington's, 1981, claim, quoted above, that separate concrete and abstract semantics is the "only plausible interpretation" of the relevant double dissociation). Shallice (1988) spells out the logic of the argument thus: "If modules exist, then . . . double dissociations are a relatively reliable way of uncovering them. Double dissociations do exist. Therefore modules exist" (p. 248). As Shallice points out, however, the inference is valid only if modular systems are the *only* sort of system that can produce double dissociations. As evidence against this, he describes a number of different types of processing systems that are not naturally described as modular, or are only partially modular, and yet could still give rise to double dissociations (see Figure 1). In view of these examples, the claim that a double dissociation implies the existence of separate modules dedicated to each task is clearly untenable.² Even the weaker notion of "functionally dissociable subsystems" (Shallice, 1979) does not capture all of the relevant distinctions (e.g., processing continuum, overlapping processing regions, multi-level systems). What, then, can be learned from the patterns of dissociations among brain-damaged patients?

Shallice (1988) proposes the notion of "functional specialization" as capturing the important distinctions in all of the relevant cases. In his formulation, the degree of specialization of part of a system is essentially defined by

² It should be pointed out that there are a number of motivations for adopting a modular perspective other than the existence of neuropsychological dissociations (see Fodor, 1983; Shallice, 1988, for discussion), but most if not all of these can be satisfied by nonmodular systems with some sort of functional specialization.

the specificity of the impairments that can arise following damage: the greater disparity in performance across the critical tasks, the more specialized the damaged region must be. Critically, the power of neuropsychological data for constraining cognitive theories is considerably weakened on this view. A double dissociation of two tasks cannot even serve to distinguish between the rather different types of system depicted in Figure 1. What is even more problematic is that we cannot know for sure whether the dimensions on which the behavioral dissociations are based (e.g., word concreteness) correspond in any discernible way to the underlying functional dimensions within the system that give rise to the behavioral differences. As Shallice (1988) puts it,

It remains logically possible that specialisation . . . could have no functional relevance. In biological systems, this seems implausible. What does remain conceivable is that in particular cases, the pattern of resource specialisation [i.e., the observed dissociations] might throw no useful light on what function is responsible for the specialisation. If and when examples of this are found will be the time to begin to consider this possibility. For the present, a reasonable conclusion is that determining the degree of specialisation within a system is a useful guide to system architecture and its functional organisation. (p. 258)

A shift away from a strictly modular perspective would have profound implications not only for the form of theoretical interpretation offered for neuropsychological phenomena, but also for the criteria for selecting brain-damaged patients for detailed study, and for the empirical methodologies that are or are not considered appropriate for studying such patients. Currently, researchers typically cast their explanations of the behavior of brain-damaged patients in terms of all-or-none functional “lesions” to one or more components of an information-processing diagram. Figure 2 illustrates this approach for the pattern of symptoms shown by PW—so-called *deep dyslexia*. In such a framework, the most productive way of determining what the components are and how they work is to study just those patients whose impairment is restricted to a single component—so-called *pure* cases (Lichtheim, 1885). Unfortunately, given the capricious nature of brain damage, such patients are exceedingly rare; brain damage will more typically impair a number of components to varying degrees. Such *mixed* cases—comprising the large majority of patients—are far more problematic to interpret and, thus, are deemed less theoretically relevant (although notice that, as reflected in Figure 2, PW is thought to be a mixed case). As Ellis (1987) has lamented, “the cognitive neuropsychologist will pass over 999 patients to find the one thousandth who comes close to being a pure case of ‘word meaning deafness’ or whatever” (p. 402). By contrast, a nonmodular framework may be more appropriate for capturing the effects of partial damage, thereby providing a basis for understanding the full distribution of effects that arise following brain damage.

Furthermore, given Caramazza’s (1986) *universality* assumption that the premorbid cognitive system is essentially equivalent across individuals, the only relevant distinction in a modular framework among patients is which component or components are impaired. Thus, patients with different underlying impairments are different sorts of patients, requiring different explanations, even if their impairments happen to give rise to the same overt pattern of behavior (Coltheart & Funnell, 1987; Shallice & Warrington, 1980). As there is no way to determine pre-theoretically which patients have the same underlying impairment, there is no legitimate basis for grouping patients together in neuropsychological investigations (Caramazza, 1984, 1986; Caramazza & McCloskey, 1988; McCloskey, 1993). Put another way, each component in a modular system operates according to its own separate principles, so there is no theoretical basis for explaining why damage to different modules should produce common effects. By contrast, various portions of a nonmodular system may operate according to common computational principles, so that different locations of damage produce the same symptoms *for the same reason*. In this case, it would be appropriate to group together patients with such damage for theoretical purposes.

In the context of examining varieties of nonmodular systems, Shallice (1988) specifically considers the relevance of “distributed-memory” systems, more commonly known as connectionist, neural, or parallel distributed processing (PDP) networks. In these networks, computation takes the form of cooperative and competitive interactions among simple, neuron-like processing units. At first glance, the distributed nature of such systems would appear at odds with the selective dissociations that can arise from brain damage. To the contrary, double dissociations of a sort have been demonstrated after damage to connectionist networks, but in these cases either the model had built-in structural distinctions corresponding directly to the observed dissociations (Farah & McClelland, 1991), or the lesions involved the removal of individual units or connections that had idiosyncratic effects (Bullinaria & Chater, 1993; Sartori, 1988; Wood, 1978). As Shallice (1988) points out,

A valid demonstration of a double dissociation arising from the effects of two different lesions to a distributed-memory system would need to satisfy two conditions. First, before the lesion is made, the

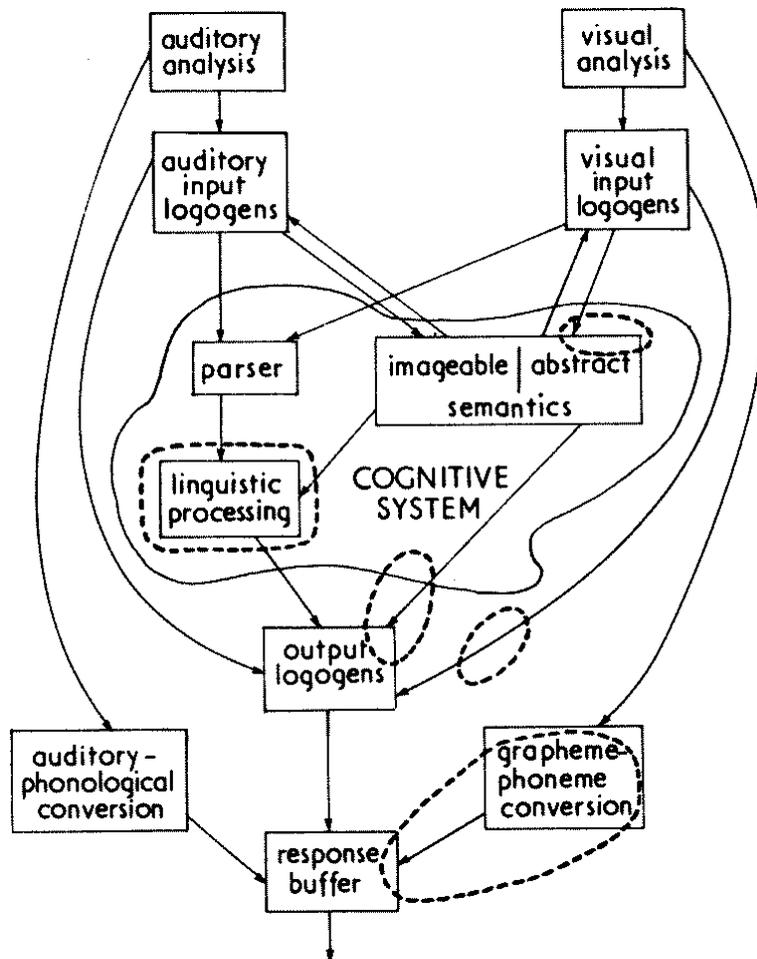


Figure 2: Morton and Patterson's (1980, p. 115) information-processing account of deep dyslexia. Note that the semantics of concrete (imageable) words and those of abstract words are assumed to be represented separately, and thus are independently susceptible to brain damage.

influence of any particular neuron on what output is produced should be small. Second, the neurons affected by the lesion should not be selected by some complex algorithm that is determined by the dissociation to be explained and that is not typical of those that arise naturally. It seems most unlikely that if these conditions are satisfied, a classical or a strong double dissociation could be demonstrated in a properly distributed memory system. (p. 256)

The current paper investigates in detail the effects of damage in a connectionist network that exhibits a double dissociation between concrete and abstract word reading (Plaut & Shallice, 1991, 1993), and yet does not suffer from the limitations of previous simulations mentioned above. The purpose of the investigation is to evaluate the theoretical status of a central pillar of cognitive neuropsychological methodology: demonstrating double dissociations among single-case studies. Specifically, a consideration of the distribution of effects across quantitatively equivalent lesions in the network (i.e., at the same location and severity) raises specific concerns about the interpretation of single-case patient studies. Furthermore, contrary to Caramazza's (1986) *transparency* assumption, the functional specialization in the network that gives rise to the double dissociation is not transparently related to the network's structure. Nonetheless, the occurrence of the double dissociation does "throw useful light" on the functional organization of the network *given specific assumptions about the computational principles governing its operation*. As a consequence, cognitive neuropsychologists cannot assume that the cognitive system must be modular in order to account for observed dissociations among brain-damaged patients. Rather, developing adequate accounts of neuropsychological deficits may require a close interplay between empirical and computational methodologies.

The next section presents in more detail the empirical data on the double dissociation between concrete and abstract word reading. Following this, the Plaut and Shallice connectionist simulation of reading via meaning is described. This network forms the basis for investigations aimed at identifying the conditions under which double dissociations occur within the network, and what can be inferred from the effects of individual lesions. The paper then concludes with a general discussion of the implications of the results for theorizing in cognitive neuropsychology.

Double Dissociation of Concrete and Abstract Word Reading

There is general agreement that normal readers have available to them (at least) two means by which to pronounce a written word: a *semantic* process that derives the pronunciation via the meaning the word, and a *phonological* process that bypasses semantics, deriving the pronunciation directly from orthography on the basis of common spelling-sound correspondences (for reviews and discussion, see Coltheart, 1987; Coltheart, Curtis, Atkins, & Haller, 1993; Henderson, 1982; Humphreys & Evett, 1985; Patterson et al., 1985). This second route also enables skilled readers to generate reasonable pronunciations for word-like nonsense letter strings (e.g., MAVE). In fact, most researchers believe that skilled readers rely primarily (perhaps even exclusively, see Van Orden, Pennington, & Stone, 1990) on the phonological route when reading words aloud. Only in cases where this route is rendered inoperative by brain damage, as in so-called deep and phonological dyslexic patients, are strong effects of semantic variables like concreteness observed (although see Strain, Patterson, & Seidenberg, 1995, for evidence of concreteness/imageability effects in normal word reading).

The hallmark symptom of deep dyslexia is the occurrence of semantic errors in oral reading (e.g., reading CAT as "dog"). In addition to semantic errors, these patients also exhibit a wide range of other symptoms in oral reading, including visual errors (e.g., CAT ⇒ "cot"), mixed visual-and-semantic errors (e.g., CAT ⇒ "rat"), morphological errors (e.g., GOES ⇒ "go"), a part-of-speech effect (nouns > verbs > adjectives > functors), a severe impairment in reading pronounceable nonwords, poor performance on other phonological tasks, and, as previously mentioned, better reading of concrete than abstract words (see Coltheart et al., 1980). In addition to severe impairment of the phonological route, it is commonly assumed that deep dyslexic patients also have partial impairment of the semantic route (e.g., Morton & Patterson, 1980; Nolan & Caramazza, 1982; Shallice & Warrington, 1980, but see Coltheart, 1980b; Newcombe & Marshall, 1980; Saffran, Bogyo, Schwartz, & Marin, 1980, for alternative accounts).

In his comprehensive 1980a review of cases of deep dyslexia in the literature at the time, Coltheart found that, among those patients for whom there was sufficient data, *all* were significantly better at reading aloud concrete than abstract words. According to an updated review (Coltheart, Patterson, & Marshall, 1987), this was still the case seven years later after numerous additional cases had been reported. As is the case for PW, cited in the Introduction, the effect can be quite large (e.g., 70% vs. 10% in DE, Patterson & Marcel, 1977; 73% vs. 14% in KF, Shallice & Warrington, 1975; and 50% vs. 10% in GR, Marshall & Newcombe, 1966). It should be noted, however, that there is considerable disagreement as to whether the relevant semantic variable is concreteness per se, or perhaps something

like imageability (Marcel & Patterson, 1978; Richardson, 1975; Shallice & Warrington, 1975) or ease-of-predication (Jones, 1985). In any case, these measures are highly intercorrelated and so may have a common underlying functional origin (see Barry & Richardson, 1988).

The concreteness of a word also appears to have a more subtle effect on the reading behavior of deep dyslexic patients, in terms of its likelihood to produce different types of errors. In particular, less concrete words are more likely to produce visual errors, and the responses in these cases tend to be more concrete than the stimuli (KF, Shallice & Warrington, 1975; BL, Nolan & Caramazza, 1982; GR, Barry & Richardson, 1988; PS, Shallice & Coughlan, 1980). Thus, a semantic variable—concreteness—clearly influences what would intuitively seem to be a more peripheral effect: visual similarity in error responses.

Phonological dyslexic patients may also exhibit effects of concreteness in single word reading. Although these patients, by definition, do not make semantic errors, they can be quite similar to deep dyslexic patients in many respects. In fact, Glosser and Friedman (1990; Friedman, 1996, also see Newcombe & Marshall, 1980) have argued that deep and phonological dyslexic patients fall on a continuum of severity. In particular, the relevant symptoms can be ordered such that each successive symptom is exhibited only by patients with an increasing severity of impairment. Thus, all phonological dyslexic patients have a selective deficit in nonword reading relative to word reading, and this may be the only symptom that the least impaired patients exhibit. Somewhat more impaired patients also make some visual and morphological errors in reading words (although word reading is still much superior to nonword reading). With still more impairment, a part-of-speech effect among words is also observed. Even more impaired patients also exhibit the concreteness effect. Finally, it is only the most severely impaired patients, arbitrarily relabeled “deep” dyslexics, who produce semantic errors (along with all the other symptoms). In corresponding fashion, Friedman (1996) has shown that the pattern of recovery in deep dyslexic patients follows the reverse ordering (also see Klein, Behrmann, & Doctor, 1994). The occurrence of semantic errors is the first symptom to resolve, followed by the concreteness effect, then the part-of-speech effect, then the visual and morphological errors, and only lastly, the impaired nonword reading.

Among those few documented phonological dyslexic patients whose impairment is severe enough to give rise to the concreteness effect but not so severe as to produce semantic errors, the magnitude of the concreteness effect tends to be smaller than is typical in deep dyslexia. For example, patient DV (Glosser & Friedman, 1990) was 100% correct on concrete words but only 87% correct on abstract words. This may, however, simply be a ceiling effect, as the overall word reading performance of such phonological dyslexic patients tends to be much better than that of deep dyslexic patients.

Thus, among deep and phonological dyslexic patients whose word reading is affected by concreteness, all find concrete words easier to read than abstract words. This concreteness effect mirrors the general advantage that concrete, highly imageable words have in a variety of cognitive and linguistic contexts (see, e.g., Kieras, 1978; Paivio, 1969, 1991; Schwanenflugel, 1991).

In striking contrast with the advantage for concrete words shown by deep and phonological dyslexic patients, patient CAV (Warrington, 1981) exhibited better performance in reading abstract words. On initial testing, CAV’s reading was very severely impaired: he failed at reading common words like MILK and TREE, but occasionally succeeded at reading abstract words like APPLAUSE, EVIDENCE, and INFERIOR. On more detailed testing, involving 387 words from the Brown and Ure (1969) list, CAV read correctly 36% of words that were more concrete than the mean, but 55% of less concrete ones. Most of his errors were visual in nature, and—also in contrast to the deep dyslexic pattern—his responses were more *abstract* than the stimuli in 67% of these, although the incidence of visual errors was approximately equal for concrete and abstract words. Although CAV made no more semantic errors than might be expected by chance (see Ellis & Marshall, 1978), he appeared to be relying at least in part on the semantic route because his performance improved when cued with the semantic category of the stimulus, and he showed a corresponding advantage for abstract words presented auditorily in a word/picture matching task. Also, as CAV could read some nonwords correctly (6/30), residual operation of the phonological route may have “edited out” some phonologically implausible semantic errors (Newcombe & Marshall, 1980).³

This double dissociation in reading aloud concrete and abstract words mirrors an analogous double dissociation in the comprehension of these word types (Warrington, 1975). When asked to define low-frequency words presented auditorily, patient EM (also see Coughlan & Warrington, 1981) could provide adequate definitions for 56% of concrete words but only 45% of abstract words. In contrast, patient AB could adequately define only 24% of concrete words but

³Visual errors would be less likely to be edited out by a partially operating phonological route as they would tend to be more phonologically plausible, given the high degree of systematicity between English orthography and phonology.

85% of abstract words.⁴ This second pattern was later replicated by Warrington and Shallice (1984) in patient SBY (50% concrete vs. 94% abstract). Warrington and Shallice also mention that SBY showed the same effect in auditory word/picture matching, although no data are given. More recently, Breedin, Saffran, and Coslett (1993, 1994) have documented an advantage for abstract words in a similar patient, DM, across a range of lexical tasks, including auditory word definition, auditory lexical decision, auditory word/picture matching, and written word synonym judgement.

The fact that performance on concrete and abstract words in oral reading and in comprehension tasks can be doubly dissociated by brain damage indicates that there are important differences in how these types of words are represented and processed in the brain. Exactly what these differences are, however, remains a matter of debate. Some researchers (e.g., Warrington, 1981) contend that, in order to account for the findings, the representations of concrete and abstract words must be neuroanatomically separate. Others, (e.g., Morton & Patterson, 1980), while not explicitly endorsing the necessity of this claim, nonetheless incorporate it into their accounts (see Figure 2). However, as pointed out in the introduction, a variety of general types of systems can give rise to double dissociations without such a strict separation of the representations and processes underlying the two tasks.

As a specific instantiation of this possibility, the next section describes the simulation of a connectionist network (Plaut & Shallice, 1993) that replicates the double dissociation of concrete and abstract word reading, but in which there is no structural separation between the semantic representations of these word types. The section begins with a discussion of a general framework for lexical processing, and situates Plaut and Shallice's general investigation of reading via meaning within this framework. The central concept of an "attractor" is introduced and used to explain the basic error pattern in deep dyslexia. The remainder of the section presents considerable detail on the architecture and representations used by the concrete/abstract network, as the simulation study presented in the subsequent chapter involves a detailed analysis of the effects of lesions to this network. The focus of the analysis is on the variability in the effects produced by quantitatively equivalent lesions.

A Simulation of Reading via Meaning

A full implementation of the word reading system would involve both semantic and phonological processes for generating the pronunciation of written words. An example of a general lexical framework embodying these processes is given by Seidenberg and McClelland (1989) and is shown in Figure 3. Within the framework, orthographic, phonological, and semantic information is represented in terms of distributed patterns of activity over separate groups of simple neuron-like processing units. Within each of these three domains, similar words are represented by similar patterns of activity. Lexical tasks involve transformations between these representations—for example, oral reading requires that the orthographic pattern for a word generate the appropriate phonological pattern. Such transformations are accomplished via the cooperative and competitive interactions among units, including additional "hidden" units that mediate between the orthographic, phonological, and semantic units. The interactions of units are governed by weighted connections between them, and it is these connection weights that collectively encode the system's knowledge about how the different types of information are related. The specific values of the weights are derived by a learning procedure on the basis of the system's exposure to written words, spoken words, and their meanings.

Seidenberg and McClelland (1989) implemented only the pathway from orthography to phonology (shown in bold in Figure 3), attempting to account for a wide range of aspects of normal skilled reading—particularly the nature of the interaction of word frequency and spelling-sound consistency in naming latency (Andrews, 1982; Seidenberg, 1985; Seidenberg, Waters, Barnes, & Tanenhaus, 1984; Taraban & McClelland, 1987; Waters & Seidenberg, 1985). Patterson, Seidenberg, and McClelland (1989, also see Patterson, 1990) attempted to replicate the symptoms of surface dyslexic patients (see Patterson et al., 1985)—in particular, the frequency-by-consistency interaction in naming accuracy and the tendency to "regularize" low-frequency exception words (e.g., DEAF ⇒ "deef")—by damaging the Seidenberg and McClelland model. More recently, Plaut and McClelland (1993) implemented a version of the Seidenberg and McClelland model using representations that better capture the relevant structure within and between orthography and phonology. The new model can read both words and pronounceable nonwords as well as skilled readers (see Plaut, McClelland, Seidenberg, & Patterson, 1994; Seidenberg, Petersen, MacDonald, & Plaut, 1996), and also forms the basis for a more adequate account of surface dyslexia (see Plaut, Behrmann, Patterson, & McClelland, 1993; Plaut et al., 1994).

⁴ Although the difference in EM's performance on concrete and abstract words was not statistically significant, his overall pattern of performance was shown to be significantly different from that of AB (Warrington, 1975).

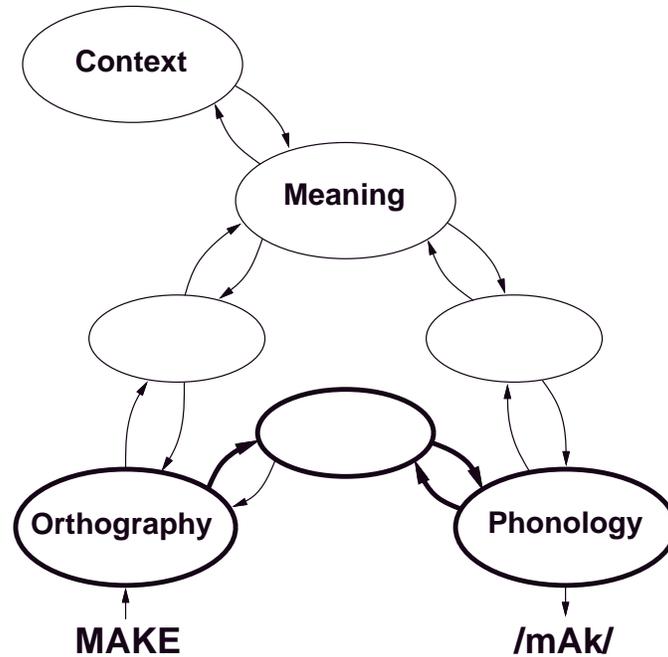


Figure 3: Seidenberg and McClelland's general framework for lexical processing. Each oval represents a group of units and each arrow represents a group of connections. The model they implemented is shown in bold. (Adapted from Seidenberg & McClelland, 1989, p. 526)

Plaut and Shallice (1993) investigated properties of various implementations of what can be thought of as the complementary portion of Seidenberg and McClelland's general framework: the pathway from orthography to phonology via meaning (semantics). Their primary motivation was to account for the specific pattern of impaired oral reading exhibited by deep dyslexic patients such as PW. Previous work by Hinton and Shallice (1991) had demonstrated that single lesions throughout a network that mapped from orthography to semantics would produce the co-occurrence of visual and semantic errors found in deep dyslexia. Plaut and Shallice systematically investigated each aspect of the design of the Hinton and Shallice simulation, attempting to identify which aspects are critical to producing the results and which are less central. The design issues included the definition of the task of reading via meaning, the network architecture (i.e., the numbers of units, their organization into layers, and how these groups are connected), the training procedure used for adjusting connection strengths, and the procedure for evaluating the behavior of the trained network in its normal state and after damage.

The major finding was that the occurrence of the qualitative error pattern was surprisingly insensitive to these detailed aspects of the simulation. Rather, what appeared critical was a more general property that all of the implementations shared: that units learned to interact in such a way that familiar patterns of activity over semantic features—corresponding to word meanings—formed stable *attractors* in the space of all possible semantic representations. Thus, after training, if the activity levels of units are set to correspond to a particular attractor pattern, unit interactions exactly balance and the network remains in that pattern. If the network is placed into a similar pattern—perhaps a distorted version of the meaning of a word—units will interact and gradually change their states so as to “clean-up” the distorted pattern into the exact meaning of the word.

It may be helpful to think about this process in the context of a high-dimensional *state space* in which the activity of each unit in the network is plotted along a separate dimension (see Figure 4). At any instant in processing an input, the pattern of activity over the entire network corresponds to a particular point within this space. This is easiest to imagine for a network with only two units; in this case, the pattern over the units corresponds to a two-dimensional point whose x and y coordinates are simply the activity levels of the two units, respectively. For a network with a hundred units, the point for a particular pattern of activity would have a hundred coordinates instead of just two. As units update their states, the global pattern of activity changes, so that the corresponding point *moves* in state space, eventually arriving at the point corresponding to the nearest word meaning. In fact, there is a region in state space

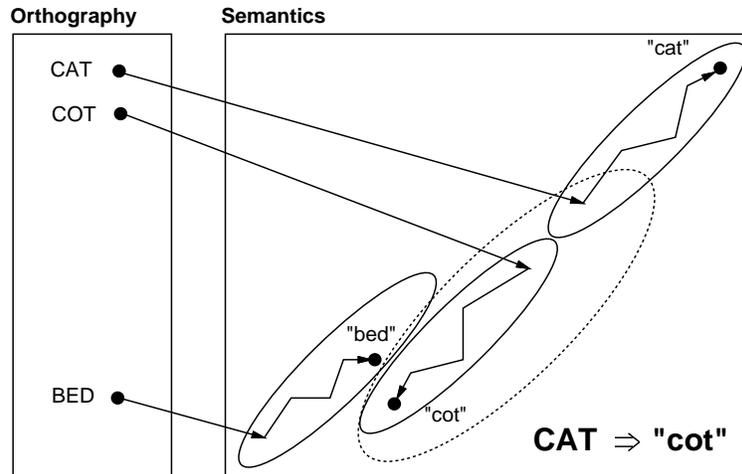


Figure 4: Mapping from orthography to semantics using attractors. For simplicity, only two dimensions of semantic space are depicted. Each point in semantics corresponds to the attractor for a word meaning; the solid oval around it corresponds to its “basin” of attraction (delimiting the other semantic patterns that will settle to it). The dotted oval depicts a basin after semantic damage, resulting in the *visual* error $CAT \Rightarrow \text{“cot”}$. (From Plaut & Shallice, 1993, p. 393)

around each familiar pattern—corresponding to a set of similar patterns—such that, if the network is set into a pattern falling anywhere within this region, the network will settle back to the familiar pattern. For this reason, each such familiar pattern is called an “attractor” in state space, and the region around it, its “basin of attraction.”

This shift in perspective, from thinking directly about the interactions of units to thinking about a changing pattern of activity as a moving point in state space, provides insight into the occurrence of errors after damage. Lesioning the network involves permanently removing some of its units and connections, affecting how the remaining units interact. In particular, it may change which initial patterns settle to which final patterns—that is, the shape and positions of attractor basins in state space. As a result, an input that, in the normal network, falls within the appropriate basin (thus leading to a correct response), may now fall within the basin of a neighboring attractor. The operation of the damaged network will thus “clean up” the pattern into the exact pattern for this neighboring attractor, producing an error response.

Notice that the use of attractors over distributed patterns of activity constitutes a controversial claim about the representational status of words (also see Seidenberg & McClelland, 1989). In most standard formulations, a word is explicitly represented in the *structure* of the reading system—for example, by the existence of a particular evidence-accumulating device called a “logogen” (Morton, 1969; Morton & Patterson, 1980). The same is true of so-called “local” connectionist models of lexical processing (e.g., Dell, 1986; McClelland & Rumelhart, 1981), in which each word is represented by a separate unit. By contrast, in a distributed attractor network, there is nothing in the structure of the system that corresponds to a word. Rather, the lexical status of a string of letters or phonemes depends solely on *functional* aspects of the system: how particular patterns of orthographic, phonological, and semantic activity interact to form stable patterns as a result of the system’s knowledge encoded in connection weights (also see van Gelder, 1990).

Plaut and Shallice’s (1993) second major objective was to extend the general approach of distributed attractor networks to account for the full range of oral reading behavior exhibited by deep dyslexic patients. Hinton and Shallice (1989) focused primarily on the co-occurrence of visual and semantic errors, and the relatively high rates of mixed visual-and-semantic errors, produced by these patients. However, many other aspect of deep dyslexia were not addressed, including other types of errors (e.g., visual-then-semantic errors such as SYMPATHY \Rightarrow “orchestra”), the part-of-speech effect, effects of concreteness and their interaction with visual errors, the existence of subvarieties of deep dyslexia, as well as less central aspects such as relatively good lexical decision and relative differences in confidence in error responses. Plaut and Shallice replicated many of these symptoms in the course of developing simulations to test the generality of the basic error pattern, but some—in particular, the concreteness effects—could not be investigated using the original Hinton and Shallice word set as it contains only concrete nouns.

Accordingly, Plaut and Shallice designed a version of the task of reading via meaning that would allow the effects of concreteness and visual similarity to be investigated directly. Twenty pairs of four-letter words were chosen such that one member of the pair was concrete, the other was abstract, and the two differed by only a single letter (e.g., ROPE and ROLE). The orthography of each word was encoded over 32 *orthographic* units, such that each of the 4 letters in the word was represented by a distributed pattern over a separate group of 8 units. The phonology of each word was represented in terms of 61 *phoneme* units, organized into position-specific groups of mutually exclusive phonemes (see Plaut & Shallice, 1993, for details). Although the resulting orthographic and phonological representations are not particularly realistic, and would not be sufficient for many other purposes, they do have the essential property that the similarities among words, either in their written forms or in their spoken forms, are reflected in the similarities of their representations.

The critical difference between concrete and abstract words relates to their semantic representations. Plaut and Shallice's (1993) approach to capturing this distinction was based in part on Jones' (1985) demonstration that words vary greatly in the ease with which predicates about them can be generated. For example, more predicates can be generated for basic-level words than for subordinate or superordinate words (Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976). Jones showed that there is a very high correlation (0.88) between ease-of-predication ratings and imageability (which also correlates highly with concreteness), and that the relative difficulty of parts-of-speech in deep dyslexia maps perfectly onto their ordered mean ease-of-predication scores. He argued that the effects of both imageability and part-of-speech in deep dyslexia can be accounted for by assuming that the semantic route is sensitive to ease-of-predication. Within the present framework, the natural way to realize this distinction is by assuming that concrete words have more semantic features (predicates) than do abstract words.

However, a literal interpretation of this manipulation would be misleading, as abstract words can certainly make rich and substantial contributions to meaning. Rather, a more appropriate interpretation relates to the degree of *variability* across contexts in the semantics generated by different types of words. As Saffran et al. (1980) point out,

A concrete word—a reference term like “rose”—has a core meaning little altered by context (a rose *is* a rose) The meanings of abstract words, on the other hand, tend to be more dependent on the contexts in which they are embedded. (p. 400)

A similar contrast appears to hold among different parts-of-speech—for example, between nouns and verbs (Gentner, 1981). Thus, using fixed semantic representations containing fewer features for abstract words should really be considered an approximation for a simulation in which abstract words have fewer semantic features that are consistently present across a variety of contexts. In fact, if a connectionist network were trained to generate pronunciations from such variable semantic representations, it would come to rely on just those few features that are consistently predictive of the correct response (see McClelland & Rumelhart, 1985, for illustrations of this property).

The particular list of 98 semantic features used by Plaut and Shallice (1993) to describe the meanings of concrete and abstract words is given in Table 1, and their assignment to words is given in Figure 5. The first 67 of the features (e.g., *found-on-farms*, *does-fly*) were based on those used by Hinton and Shallice (1991) and apply exclusively to concrete words. The remaining 31 features (e.g., *has-duration*, *relates-money*) apply primarily to abstract words but occasionally also to concrete words. Overall, concrete words contain an average of 18.2 semantic features while abstract words average only 4.7.

If the orthographic and phonological representations were suspect, these semantic representations must be approached with even more caution. However, as Plaut and Shallice (1993) are quick to point out,

We do not claim that this representation adequately captures the richness and subtlety of the true meanings of any of these words. Rather, we claim that it captures important qualitative distinctions about the relationships *between* word meanings—namely, that similar words (e.g., LACK and LOSS) have similar representations, and that there is a systematic difference between the semantics of concrete and abstract words that reflects their relative ease-of-predication. (p. 452)

In particular, the exact identity of the semantic features is of no importance. The operation of the network is insensitive to any external labels such as *made-from-other-nonliving* that may be attached to units for the benefit of external observers. What *is* important is that the semantic features induce the appropriate relative similarities among words. In this regard, the semantic features provide a coarse but adequate approximation of the relationships among these words.

The architecture of the network that Plaut and Shallice (1993) used to investigate effects of concreteness is shown in Figure 6. As can be seen by comparison with Figure 3, the network is broadly consistent with the semantic pathway

Table 1: Semantic Features for Concrete and Abstract Words

1	max-size-less-foot	35	found-in-transport	68	positive
2	max-size-foot-to-two-yards	36	found-in-factories	69	negative
3	max-size-greater-two-yards	37	surface-of-body	70	no-magnitude
4	main-shape-1D	38	above-waist	71	small
5	main-shape-2D	39	natural	72	large
6	main-shape-3D	40	mammal	73	measurement
7	cross-section-rectangular	41	bird	74	superordinate
8	cross-section-circular	42	wild	75	true
9	cross-section-other	43	does-fly	76	fiction
10	has-legs	44	does-swim	77	information
11	has-arms	45	does-run	78	action
12	has-neck-or-collar	46	living	79	state
13	white	47	carnivore	80	has-duration
14	brown	48	plant	81	unchanging
15	color-other-strong	49	made-of-metal	82	involves-change
16	varied-colors	50	made-of-liquid	83	temporary
17	dark	51	made-of-other-nonliving	84	time-before
18	hard	52	got-from-plants	85	future-potential
19	soft	53	got-from-animals	86	relates-event
20	sweet	54	pleasant	87	relates-location
21	moves	55	unpleasant	88	relates-money
22	indoors	56	dangerous	89	relates-possession
23	in-kitchen	57	man-made	90	relates-work
24	on-ground	58	container	91	relates-power
25	on-surface	59	for-eating-drinking	92	relates-reciprocation
26	otherwise-supported	60	for-wearing	93	relates-request
27	outdoors-in-city	61	for-other	94	relates-interpersonal
28	in-country	62	for-lunch-dinner	95	quality-difficulty
29	found-woods	63	particularly-assoc-child	96	quality-organized
30	found-near-sea	64	particularly-assoc-adult	97	quality-bravery
31	found-near-streams	65	used-for-games-or-recreation	98	quality-sensitivity
32	found-mountains	66	human		
33	found-on-farms	67	female		
34	found-in-public-buildings				

Note: From Plaut & Shallice, 1993, p. 450.

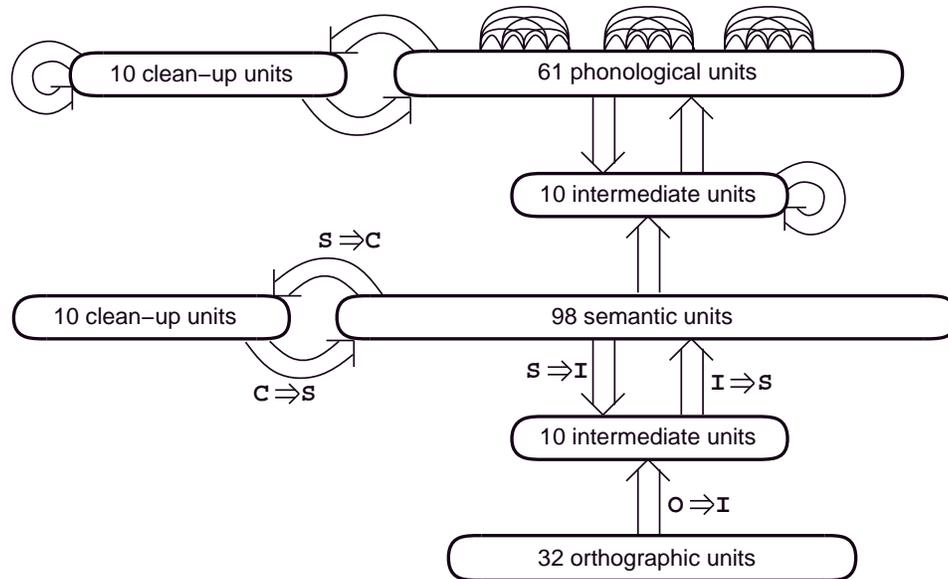


Figure 6: The architecture of the network used by Plaut and Shallice (1993) to investigate the effects of concreteness in deep dyslexia. Ovals represent groups of units, while large arrows represent complete connectivity within or between layers. Sets of connections are named in terms of the first letter of the names of the unit groups they connect (e.g., $O \Rightarrow I$ for orthographic-to-intermediate connections). (From Plaut & Shallice, 1993, p. 453)

in the Seidenberg and McClelland (1989) general framework for lexical processing.⁵ Orthographic input maps by way of 10 intermediate units to semantics, which in turn maps via 10 additional intermediate units to phonology. In addition, both semantics and phonology are each reciprocally connected with a separate set of 10 *clean-up* units. These clean-up units are critical because they implement the attractors for word meanings and pronunciations: during the processing of an input, the clean-up units interact with and influence the activities of the semantic units and the phoneme units, gradually pushing them towards the correct values. Each set of connections (indicated by an arrow in the Figure) represents complete connectivity from units in the sending group to units in the receiving group. In particular, any unit connected to a semantic unit is connected to *every* semantic unit. As units are insensitive to the ordering of their connections, the actual ordering of semantic features as listed in Table 1 and Figure 5—and in particular, the apparent separation of features applying to concrete words from features applying primarily to abstract words—is irrelevant to the simulation.

Input is presented to the network by clamping the states of the orthographic units to the representation of some word. Other units in the network have their states initialized to a low resting value. In processing the input, unclamped units repeatedly update their states based on the states of connected units and the weights on these connections. Initially, most effects are near the clamped input, but gradually the effects of this input are felt further into network. Initial semantic activity is progressively refined by interactions with clean-up units, while phonological units begin to become active based on partial semantic activity. Eventually, the semantic units settle into a pattern of activity representing the meaning of the input word, and the phonological units settle under pressure from their clean-up units into a pattern of activity representing the pronunciation of the word.

The network was trained with back-propagation through time (Rumelhart, Hinton, & Williams, 1986; Williams & Peng, 1990) to settle into the correct semantics and phonology of each of the 40 words when presented with its orthography (see Plaut & Shallice, 1993, for details). This procedure calculates how to change each connection weight so as to gradually reduce the total error on the task, where error is defined in terms of the discrepancy, for both semantics and phonology, between the patterns generated by the network and the correct patterns for each input.⁶

⁵Notice that, by implementing only the semantic pathway, Plaut and Shallice (1993) are implicitly assuming, along with most researchers (but not all; see, e.g., Buchanan, Hildebrandt, & MacKinnon, 1994) that the phonological pathway is completely inoperative in deep dyslexic patients. More subtly, they are assuming that the nature of the representations and processes in the semantic route do not depend critically on the fact that they develop in the context of a concurrently developing phonological route. In actuality, this claim is likely to be only approximately true (see Plaut et al., 1994, for simulations and discussion).

⁶Although back-propagation, in its literal form, is implausible as a neurobiological learning procedure (Crick, 1989), more plausible learning

The weights for all of the connections in the network were trained in the same way, including those involved in implementing the semantic and phonological attractors.

After the network had learned to pronounce all 40 words equally well, each set of connections within the “input” portion of the network (up to and including semantics) was subjected to lesions across a range of severities, in which some proportion of the connections in the set were randomly selected and removed. The network’s responses to the 40 words were accumulated after 50 specific random lesions at each combination of 5 locations and 9 severities. The following is a summary of the major findings (see Plaut & Shallice, 1993, for details):

1. Correct performance on concrete words was significantly better than on abstract words after lesions to the “direct” pathway from orthography to semantics (i.e., the $O \Rightarrow I$ and $I \Rightarrow S$ connections) at every level of severity.
2. Slight and moderate lesions to the “clean-up” pathway (i.e., $S \Rightarrow C$ and $C \Rightarrow S$) produced no relative difference in performance in concrete versus abstract words, but severe lesions to these sets of connections (i.e., 0.5 and 0.7) produced the *reverse* advantage: abstract words were read significantly better than concrete words.
3. For lesions of the direct pathway, visual errors (i.e., responses overlapping the stimulus in at least two letters) were more likely in response to abstract words than to concrete words, and the responses in these errors tended to be more concrete than the stimulus. For severe lesions of the clean-up pathway, the opposite effects obtained: visual errors were more likely on concrete words, and the responses tended to be more abstract than the stimulus.

Thus, lesions to the pathway from orthography to semantics replicated the effects of concreteness and their interaction with visual similarity found in deep dyslexia. Conversely, severe lesions to the semantic clean-up pathway replicated the effects observed in the concrete word dyslexic patient CAV (Warrington, 1981). In fact, the etiology of severe damage at the semantic level is consistent with other aspects of CAV’s behavior. His reading disorder was quite severe initially, and he also showed an advantage for abstract words in auditory word/picture matching, suggesting modality-independent damage at the semantic level.

Table 2 illustrates the double dissociation of concrete and abstract word reading produced after direct versus clean-up lesions to the network. The results listed are the averages from 50 instances of each type of lesion. Each of these 50 instances might be thought to correspond to a particular hypothetical patient who, presumably, would have a particular brain lesion. The standard objections to averaging the behavioral results of neuropsychological patients (Caramazza, 1984, 1986; Caramazza & McCloskey, 1988; McCloskey, 1993) do not apply, as the experimental manipulation guarantees that the various instances of the lesioned network have the same “functional” lesion. Nonetheless, there are interesting and important issues concerning the distribution of effects within a functionally equivalent patient population that cannot be addressed solely on the basis of data on average performance. In fact, almost all of the results reported by Plaut and Shallice (1993) are averaged, not only over quantitatively equivalent lesions, but also over a range of severities—those resulting in average correct performance between 15–85%.⁷ This averaging was done because Plaut and Shallice were concerned primarily with demonstrating that the general tendencies of their network under damage corresponded to the symptom-complex exhibited by deep dyslexic patients in general.

The current paper adopts a different emphasis. Specifically, it focuses on the variation in the effects of quantitative equivalent lesions (i.e., those at the same location and severity). Such an investigation is critical because much of neuropsychological theorizing is based in the behavior of individual patients. If we assume that a given patient has a particular location and severity of lesion within the cognitive system, then one can think of a population of such patients with different random instances of the same location and severity of lesion. In this way, any individual patient may be viewed as a sample from the (hypothetical) population of patients with equivalent lesions. In the simulation, it will turn out that different random instances of equivalent lesions can have qualitatively different effects—specifically, on whether concrete or abstract word reading is selectively impaired. The findings thus call into question the theoretical implications of reliance on single-case studies.

Simulation Study

The simulation study involves a detailed comparison of the effects of two types of lesions of the Plaut and Shallice (1993) concrete/abstract network shown in Figure 6. The two types studied are lesions of 20% of the orthographic-to-intermediate ($O \Rightarrow I$) connections, and lesions of 70% of the semantic-to-cleanup ($S \Rightarrow C$) connections. These two lesion

procedures—such as contrastive Hebbian learning (Peterson & Anderson, 1987)—produce qualitatively equivalent results (Plaut & Shallice, 1993).

⁷Plaut and Shallice (1993, pp. 426–428) do report data on the effects of lesion severity on error pattern, and on the effects on individual lesions, but only in the very restricted context of verifying the generality of the Hinton and Shallice (1991) results.

Table 2: Correct Performance on Concrete and Abstract Words after Lesions

Lesion		Percent Correct	
Location	Severity	Concrete Words	Abstract Words
$0 \Rightarrow I$	0.2	52.7	25.4
$S \Rightarrow C$	0.7	28.4	42.7

Note: “Location” refers to a particular set of connections as labeled in Figure 6. “Severity” refers to the proportion of these connections randomly selected and removed for each lesion. Results are averaged over 50 instances of such lesions. (From Plaut & Shallice, 1993, p. 457)

types were chosen because, among the locations and severities of lesions investigated by Plaut and Shallice, these two produce the clearest double dissociation between correct performance on concrete versus abstract words (as shown in Table 2). Notice that the two lesion types are not equated for the overall level of performance that they produce (average percent correct is 40.6% after $0 \Rightarrow I(0.2)$ lesions versus 34.1% after $S \Rightarrow C(0.7)$ lesions). This difference is not problematic for the current study as the goal is to explore the variability within each lesion condition in the effects caused by specific lesions.

Method

The concrete/abstract network was subjected to 1000 instances of each of the two lesion types: $0 \Rightarrow I(0.2)$ and $S \Rightarrow C(0.7)$. This large sampling of specific lesions enabled a better evaluation of the distribution of effects caused by lesions. Each of the 1000 lesions involved randomly selecting and removing the specified proportion of connections from the network. The damaged network was then presented with each of the 40 words for processing, as described in the previous section. Correct performance was measured and error responses were accumulated and categorized as described below.

As a result of the damage, the activity levels of the phonological units at the end of processing were not generally identical to the (correct) levels produced by the undamaged network. Nonetheless, they might still constitute a reasonable pronunciation. In order for the phonological activity generated by the network to be considered a well-formed pronunciation, exactly one phoneme unit (possibly corresponding to a “null” phoneme) had to be active at each position (see Plaut & Shallice, 1993, for the precise definition of “active”). If, in response to an input, the damaged network did not produce a well-formed pronunciation, that word presentation was interpreted as an omission. Otherwise, the concatenation of active phonemes was taken as the network’s response. This phoneme string could correspond to the pronunciation of the stimulus word (correct response) or it could be some other string of phonemes (error response). Thus, each word presentation produced either an omission, a correct response, or an error response.

Among error responses, if the string of phonemes produced by the network did not match the pronunciation of any of the 40 words in the training corpus, the error was classified as a nonword. The remaining (word) error responses were categorized in terms of their relation to the stimulus word. A response was considered visually similar to the stimulus if the two words overlap in at least two of their four letters. This corresponds to the standard (if somewhat vague) neuropsychological criterion (e.g., Morton & Patterson, 1980) of accepting a response as visually similar to a stimulus if the two share at least 50% of their letters in approximately the same order.

Defining semantic similarity is more problematic, both for patient responses and for the network. For the Hinton and Shallice (1991) word set, the semantic representations that were assigned to words ensured that, in general, words in the same category tended to have similar (overlapping) semantic representations. In the concrete/abstract word set used in the present simulation, words are not organized into semantic categories and so category membership cannot be used as the basis for deciding semantic relatedness. Instead, we use the degree of overlap of semantic representations as a direct measure of relatedness. Specifically, we consider two words to be semantically similar if their semantic representations overlapped in sufficiently many features. The exact number of features required differs between concrete and abstract words due to the systematic differences in their semantic representations (see Plaut & Shallice, 1993, for details). Fortunately, the exact values of these criteria are relatively unimportant as we will be

Table 3: Chance Error Proportions for Concrete and Abstract Words

Word Type	Error Type			
	Visual	Semantic	Vis&Sem	Other
Concrete	.121	.044	0.0	.835
Abstract	.115	.012	.072	.801

concerned only with comparing the resulting rates of semantic errors with the chance rates of such errors among randomly paired words.

Given the definitions of visual and semantic similarity, any word error response can be classified as either visual, semantic, visual-and-semantic, or “other” (unrelated to the stimulus). Table 3 lists the relative proportions of each of these error types for error responses chosen randomly from the training corpus. Notice that the large majority of possible error responses are unrelated to the stimulus. Also, concrete and abstract words have approximately equal chance rates of visual errors. Concrete words are somewhat more likely to produce semantic errors by chance although, according to the definitions of visual and semantic similarity, they *cannot* produce visual-and-semantic errors. This limitation is an artifact of the word set and similarity definitions, and is not problematic for the current work as we will be concerned only with comparing the relative rates on visual errors and on semantic errors.

Results and Discussion

Correct Performance

After 1000 lesions of 20% of the orthographic-to-intermediate ($0 \Rightarrow I$) connections, average correct performance on the 40 word training corpus is reduced from 100% to 40.1% ($SD = 12.1$). Replicating the findings of Plaut and Shallice (1993), correct performance on concrete words is reliably better than on abstract words (paired $t_{999} = 50.5$, $p < .001$). On average, 52.9% ($SD = 14.9$) of concrete words are read correctly compared with only 27.2% ($SD = 14.1$) of abstract words. Also replicating the previous findings, lesions of 70% of the semantic-to-cleanup ($S \Rightarrow C$) connections have the opposite effect. Correct performance is reduced to 36.1% ($SD = 7.0$) overall, but abstract words are read better (45.1%, $SD = 10.5$) than are concrete words (27.1%, $SD = 9.4$; paired $t_{999} = 40.0$, $p < .001$). Figure 7 presents these data in graphical form, illustrating the clear cross-over double dissociation.

Of more interest is the distributions of these effects across individual lesions. Figure 8 plots abstract word performances against concrete word performance separately for the two lesion types. In the plots, the radius of each circle is proportional to the number of lesions yielding the performance levels on concrete and abstract words indicated by the position of the circle. The diagonal lines in each plot correspond to equal levels of performance on the two sets of words. The overall double dissociation of concrete versus abstract word reading after $0 \Rightarrow I$ versus $S \Rightarrow C$ lesions is evident by comparing the two plots in the Figure. There is a strong tendency for $0 \Rightarrow I$ lesions to produce levels of performance below the main diagonal (i.e., concrete word performance $>$ abstract word performance) while the opposite is true of $S \Rightarrow C$ lesions (i.e., concrete word performance $<$ abstract word performance). However, while these dissociations are generally true after these lesions, they are not universally true. Specifically, 4.2% of lesions to the $0 \Rightarrow I$ connections produce better performance on abstract words, while 7.9% of lesions to the $S \Rightarrow C$ connections produce better performance on concrete words.

This overlap in the distributions of effects produced by the two types of lesions can be brought out further by plotting the proportion of lesions producing particular differences in correct performance on concrete word versus abstract words (i.e., percent correct on concrete words minus percent correct on abstract words) for the two lesion types (see Figure 9). For example, lesions producing 40% versus 10% and 75% versus 45% on concrete versus abstract word reading, respectively, would both contribute to the proportion of lesions producing a difference of 30%. More generally, these data can be interpreted as the results of projecting the data in each plot in Figure 8 onto the plane orthogonal to the main diagonal. The two lesion types clearly differ on this measure ($0 \Rightarrow I(0.2)$: mean 25.7, $SD = 16.1$; $S \Rightarrow C(0.7)$: mean -18.0 , $SD = 14.2$; $t_{1998} = 64.3$, $p < .001$). Nonetheless, there is substantial overlap between the two distributions. Furthermore, if we consider individual lesions, we can observe a double dissociation between concrete and abstract word reading after two instances of quantitatively equivalent lesions to the same set

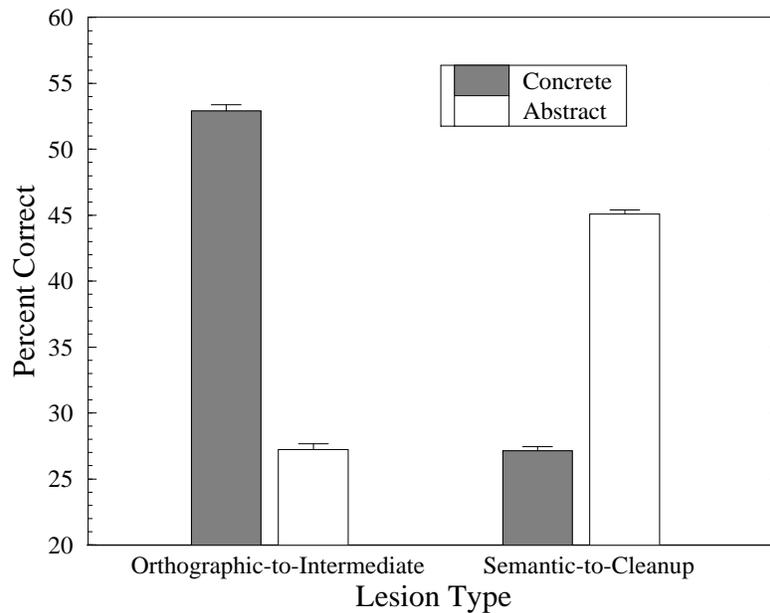


Figure 7: Average correct performance on concrete and abstract words as a function of lesion type. Results are averaged over 1000 lesions of each type. The “orthographic-to-semantic” lesions involve removal of 20% of those connections, while the “semantic-to-cleanup” lesions involve removal of 70% of those connections.

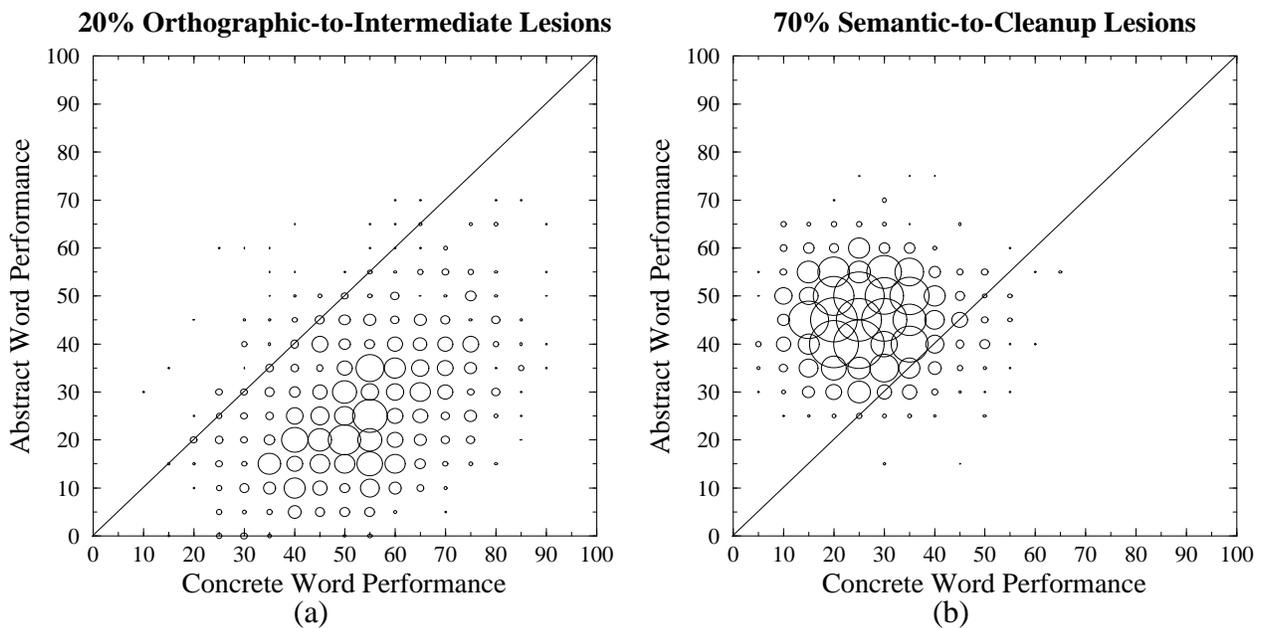


Figure 8: Percent correct performance on concrete versus abstract words after (a) lesions of 20% of orthographic-to-intermediate ($O \Rightarrow I$) connections and (b) lesions of 70% of semantic-to-cleanup ($S \Rightarrow C$) connections. The radius of each circle is proportional to the number of lesions yielding the performance levels indicated by the position of the circle. The diagonal lines correspond to equal levels of performance on concrete and abstract words.

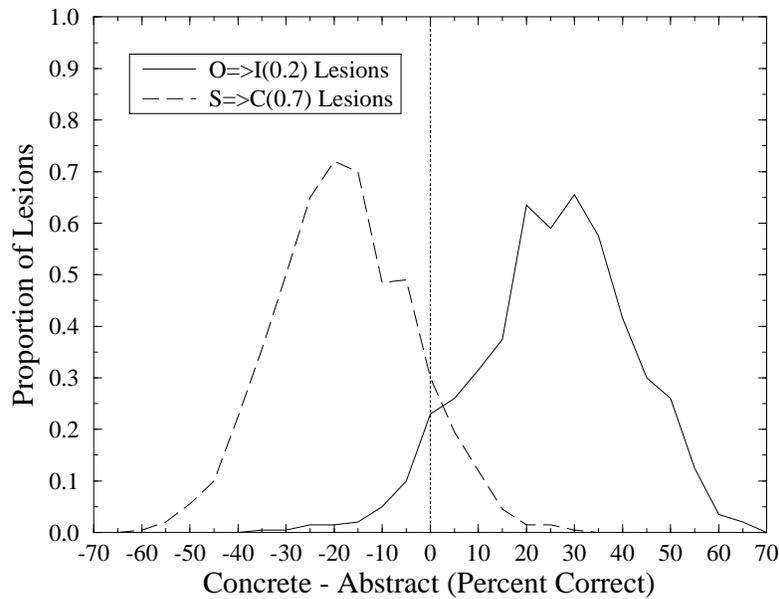


Figure 9: Distributions of the difference in percent correct on concrete and abstract words after lesions of 20% of orthographic-to-intermediate ($O \Rightarrow I$) connections and lesions of 70% of semantic-to-cleanup ($S \Rightarrow C$) connections.

Table 4: Correct Performance on Concrete and Abstract Words after Specific Instances of Lesions

Lesion	Percent Correct	
	Concrete Words	Abstract Words
$O \Rightarrow I(0.2)$		
Instance 636	80	15
Instance 88	25	60
$S \Rightarrow C(0.7)$		
Instance 831	45	15
Instance 143	10	65

of connections (see Table 4). Even though each cell in the Table reflects performance on only 20 words, all of the differences in performance, both between word types for a particular lesion and between lesion instances for a particular word type, are reliable at the .05 level by one-tailed Fisher exact tests. Thus, these differences constitute valid double dissociations by the criteria applied to the performance of neuropsychological patients (see Shallice, 1988, Chapter 10).

Error Pattern

Table 5 presents the distribution of outcomes of all word presentations to the damaged network, as either correct responses, omissions, or error responses. As can be seen from the Table, when damage prevents the network from responding correctly, it often fails to respond at all. The rates of explicit error responses are rather low, particularly after lesions to the semantic-to-cleanup connections. These connections are critical for implementing the attractors for word meanings; the fact that lesions to them yield very low error rates provides direct evidence that the network produces the error pattern in deep dyslexia only when intact attractors are operating to clean up distorted activity patterns. However, although deep dyslexic patients do fail to respond to some words, their explicit error rates are generally much higher than those of the network (see, e.g., Shallice & Warrington, 1980).

In evaluating this discrepancy between the model and the patients, it is important to bear in mind that many

Table 5: Percentages of Correct, Omission, and Error Responses after Lesions

	Lesion	
	$0 \Rightarrow I(0.2)$	$S \Rightarrow C(0.7)$
All Words		
Correct	40.1	36.1
Omissions	48.4	61.0
Errors	11.5	2.9
Concrete Words		
Correct	52.9	27.1
Omissions	39.2	68.3
Errors	7.9	4.6
Abstract Words		
Correct	27.2	45.1
Omissions	57.7	53.8
Errors	15.1	1.1

specific aspects of the model contribute to its quantitative behavior that are not central aspects of the proposed theory of reading via meaning (also see Plaut & Shallice, 1993, for discussion). In particular, the criterion used to decide when a pronunciation is “well-formed” affects the relative frequency of occurrence of omissions. If this criterion is relaxed, the network produced more explicit error responses relative to omissions. However, the rate of nonword responses—which are rare in deep dyslexia—also increases (see Plaut & Shallice, 1994, for details). Thus, the high relative rates of omissions would appear to reflect a real limitation in the ability of the network to generate phonological responses. The “output” portion of the network (i.e., from semantics to phonology) was developed primarily to avoid having to apply criteria directly to semantic activity—as Hinton and Shallice (1991) were forced to do—rather than as a realistic model of human speech production per se. The development of a more adequate distributed connectionist model of speech production must await further research (see Dell, Juliano, & Govindjee, 1993, for promising work along these lines). For the present, we must be content with comparisons of the *relative* rates of different error types after lesions. Fortunately, these comparisons are fairly insensitive to the particular procedure used to generate responses (Plaut & Shallice, 1993).

Table 6 presents the overall rates of each error type—visual, semantic, visual-and-semantic, nonword, and other—after $0 \Rightarrow I(0.2)$ and $S \Rightarrow C(0.7)$ lesions to the network. Considering the data for the entire word set, notice that, as in deep dyslexia, the rates of nonword responses to words are quite low after both types of lesion. Also notice that the rates of visual and semantic errors relative to “other” errors are greater than expected by chance similarity within the corpus. This is indicated by the numbers listed in parentheses next to each error rate. This number is the ratio of the rate for that error type with the rate of unrelated errors, divided by the corresponding ratio based on responses generated randomly from the corpus (see Table 3). If the network were responding randomly, the observed ratio of each error type with unrelated errors would be equal to the chance ratio, yielding values of 1.0. As can be observed in Table 6, the observed values are all significantly greater than one, although the rates of visual-and-semantic errors after $0 \Rightarrow I(0.2)$ lesions are only slightly greater than chance. In general, both locations of lesion produce above-chance rates of both visual and semantic errors (also see Hinton & Shallice, 1991).

Figure 10 provides data on the distributions of the co-occurrence of visual and semantic errors across individual lesions to the network. Considering $0 \Rightarrow I(0.2)$ lesions first, 41.1% of lesions produce both visual and semantic errors, 48.0% of lesions produce only visual errors, 4.7% produce only semantic errors, and 6.2% of lesions do not produce either type of error. The network shows a stronger tendency to produce visual errors than semantic errors, although this is partly due to the relative chance rates of these error types (see Table 3). Also, the occurrence of visual errors without semantic errors is common in most forms of acquired dyslexia (see Shallice, 1988, for review). What is peculiar to deep dyslexia among acquired dyslexias is the opposite relation: the co-occurrence of visual errors in patients who make semantic errors. In the network, 89.7% of lesions producing semantic errors also produce visual errors. Nonetheless, the occasional lesion will produce semantic errors with no visual errors. The occurrence of such

Table 6: Rates of Each Error Type and Ratios with “Other” Errors (Divided by Chance Ratio) after Lesions

Lesion	Word Type	Error Type				
		Visual	Semantic	Vis&Sem	Nonword	Other
$0 \Rightarrow I(0.2)$						
	All	5.77 (10.1)	1.43 (10.5)	0.30 (1.71)	0.01	3.99
	Concrete ^a	3.45 (11.0)	1.06 (7.7)	–	0.02	2.61
	Abstract ^a	8.20 (10.5)	2.79 (34.3)	0.61 (1.26)	0.00	5.42
$S \Rightarrow C(0.7)$						
	All	1.54 (14.4)	0.47 (18.6)	0.10 (3.07)	0.03	0.74
	Concrete ^a	2.69 (23.3)	0.53 (10.5)	–	0.05	0.96
	Abstract ^a	0.34 (4.6)	0.25 (32.7)	0.19 (4.15)	0.00	0.51

Note: Rates are percentages of all word presentations and are averaged over 1000 instances of each lesion type. Each number in parentheses is the ratio of that error rate with the rate of “other” (unrelated) errors in that condition, divided by the same ratio for “chance” error responses (chosen at random from the corpus). Numbers greater than 1.0 indicate that the network’s tendency to produce that type of error is greater than predicted by chance.

^aRates are normalized relative to the chance error rates for each word type and error type (see Table 3).

effects in the network, although rare, may provide an explanation for the existence of two documented cases of patients who make semantic errors but no visual errors (Caramazza & Hillis, 1990).⁸

The very low error rates produced by $S \Rightarrow C(0.7)$ make an analysis of the distribution across lesions difficult. Nonetheless, as can be seen in Figure 10, of the 18.2% of lesions that produce semantic errors, about half (47.8%) also produce visual errors.

Returning to Table 6, another important effect involves a comparison of the rates of visual errors produced by concrete versus abstract words. As mentioned in the review of the empirical data, a common observation (Barry & Richardson, 1988; Nolan & Caramazza, 1982; Shallice & Coughlan, 1980; Shallice & Warrington, 1975) is that deep dyslexia patients are particularly prone to produce visual errors in response to abstract words. The same is true of the network after $0 \Rightarrow I(0.2)$ lesions: the rate of visual errors is much higher for abstract words (8.20%) than for concrete words (3.45%). These rates have been normalized to take into account the chance rates of visual errors for each word type (as listed in Table 3). By contrast, $S \Rightarrow C(0.7)$ lesions produce the opposite effect: concrete words produce a higher rate of visual errors (2.69%) than do abstract words (0.34%). This latter finding is somewhat discrepant with Warrington’s (1981) observation that the visual error rates of concrete word dyslexic patient, CAV, was unaffected by the concreteness of the stimulus.

These findings are clarified further if we consider how the visual error rates on concrete words versus abstract words are distributed across lesion instances (see Figure 11). Abstract words produce more visual errors than concrete words after well over half (56.9%) of $0 \Rightarrow I(0.2)$ lesions. In the extreme, six lesions produced visual errors on 30% of abstract words and no visual errors on concrete words. However, a substantial proportion of such lesions (19.3%) give the opposite result, with concrete words producing more visual errors. For one lesion, this consisted of visual errors on 25% of concrete words and no visual errors on abstract words. By contrast, only 3.8% of $S \Rightarrow C(0.7)$ lesions produced more visual errors on abstract words than on concrete words. Forty-four percent of such lesions produced more visual errors on concrete words. Thus, while different locations of lesion produce distinctive patterns of visual errors across concrete and abstract words, there is also considerable overlap in the effects when individual lesions are considered (see Figure 12).

⁸A third patient, KE (Hillis, Rapp, Romani, & Caramazza, 1990), produced some visual errors on preliminary testing. Furthermore, in the main experiments reported, she was retested repeatedly on stimulus items from a fixed set of semantic categories (4 or 10). It seems likely that she could learn to restrict her responses to the relevant categories, thus eliminating many visual errors (which would be unlikely to fall in any of the semantic categories tested).

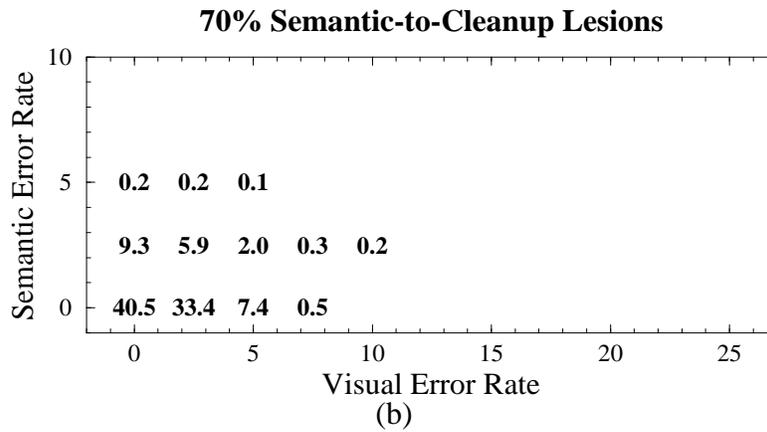
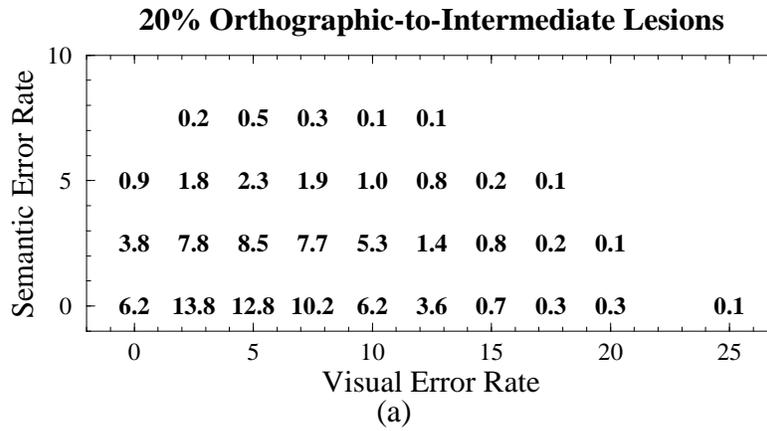


Figure 10: Rates of visual errors vs. rates of semantic errors, as percentages of all word presentations, after (a) $O \Rightarrow I(0.2)$ lesions and (b) $S \Rightarrow C(0.7)$ lesions. The numbers plotted are the percentages of lesions giving rise to the rates of visual and semantic errors indicated by the position of the number.

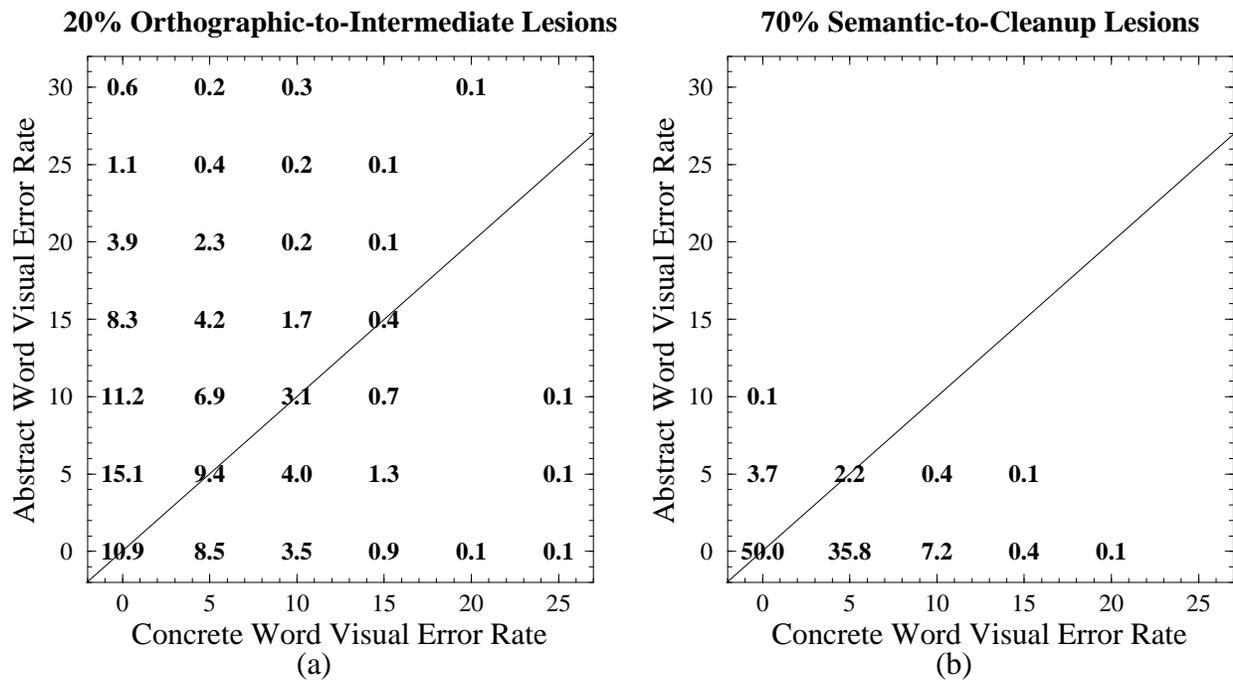


Figure 11: Rates of visual errors produced by concrete vs. abstract words after (a) $O \Rightarrow I(0.2)$ lesions and (b) $S \Rightarrow C(0.7)$ lesions. The numbers plotted are the percentages of lesions giving rise to the visual error rates on concrete words and on abstract words indicated by the position of the number. The diagonal line indicates equal rates of visual errors on the two word types.

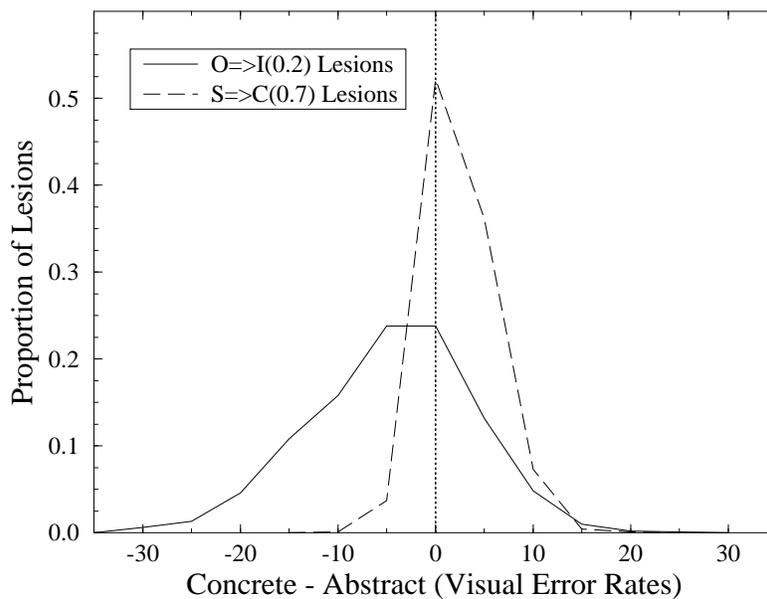


Figure 12: Distributions of the difference in visual error rates produced by concrete and abstract words after lesions of 20% of orthographic-to-intermediate ($O \Rightarrow I$) connections and lesions of 70% of semantic-to-cleanup ($S \Rightarrow C$) connections.

Summary

The current simulations investigate the behavior of the Plaut and Shallice (1993) concrete/abstract network after two particular types of lesion: 20% of the orthographic-to-intermediate connections, and 70% of the semantic-to-cleanup connections. The focus of the investigation is on the distribution, over a large number of specific instances of each lesion type, of three major effects: (1) a double dissociation in correct performance on concrete and abstract words; (2) the co-occurrence of visual errors with semantic errors; and (3) an effect of concreteness on the rates of visual errors. If the effects are averaged over lesion instances, the findings replicate those of Plaut and Shallice (1993). Specifically, (1) performance on concrete words is reliably better than on abstract words after $O \Rightarrow I$ lesions, but the opposite is true after $S \Rightarrow C$ lesions; (2) both visual errors and semantic errors occur at above-chance rates after either $O \Rightarrow I$ or $S \Rightarrow C$ lesions; and (3) abstract words produce more visual errors than do concrete words after $O \Rightarrow I$ lesions, but the opposite is true after $S \Rightarrow C$ lesions. The double dissociation, in particular, mirrors the relative performance of concrete versus abstract word reading in deep dyslexia versus concrete word dyslexia.

For each of these effects, however, there are individual lesion instances that violate the overall effect for that lesion condition. Thus, (1) some $O \Rightarrow I$ lesions yield reliably better performance on abstract words, and some $S \Rightarrow C$ lesions yield reliably better performance on concrete words; (2) some $O \Rightarrow I$ and $S \Rightarrow C$ lesions produce semantic errors but no visual errors; and (3) some $O \Rightarrow I$ lesions result in higher visual error rates on concrete words; and some $S \Rightarrow C$ lesions result in higher visual error rates on abstract words. The basic finding is clear: even when a particular location and severity of damage yields a characteristic pattern of breakdown, individual lesions may not faithfully reflect that pattern.

General Discussion

The purpose of the current investigation is to evaluate the theoretical status of double dissociations among single-case studies. Double dissociations play a central role in theorizing in cognitive neuropsychology. If each of two tasks can be selectively impaired by brain damage to two individuals, there would seem to be a basis for believing that the two tasks are subserved by separate brain mechanisms. As Shallice (1988) has pointed out, however, this logic is predicated on certain assumptions about the structure of the cognitive system—specifically, that it is composed of independent components or modules, each dedicated to performing a particular cognitive process (Chomsky, 1980; Coltheart, 1985; Fodor, 1983). If nonmodular systems can also give rise to double dissociations when damaged, the observation of such a dissociation in patients, in and of itself, does not provide evidence for a modular organization of the cognitive system.

Our first concern in this discussion is with establishing what constitutes a valid demonstration of a double dissociation in a computational simulation of a nonmodular system. Only once this is established will we take up the issue of analyzing the functional specialization in the network that gives rise to the double dissociation, and evaluating the implications of this specialization for the role of cognitive neuropsychology in understanding human cognitive processing.

The current work investigates the effects of damage in a connectionist network that has been trained to generate the pronunciations of written words via their meanings. The network forms part of a larger systematic investigation (Plaut & Shallice, 1993) of the conditions under which networks that develop attractors for word meanings, when damaged, exhibit the diverse set of symptoms found in deep dyslexia. The focus of the current simulation study is on the network's performance in reading concrete words versus reading abstract words after lesions of particular severities to two separate locations in the network: in the "direct" pathway from orthography to semantics, and in the "cleanup" pathway that forms semantic attractors. Performance of the network was evaluated after 1000 instances of each of these two lesion types, corresponding to 2000 individual brain-damaged patients. Averaging across the effects of individual lesions, the two lesion types produce a clear double dissociation between concrete and abstract word reading, corresponding to that observed between deep dyslexic patients on the one hand (see Coltheart et al., 1980), and the concrete word dyslexic patient, CAV, on the other (Warrington, 1981). Thus, on Shallice's (1988) arguments, there is evidence for functional specialization in different portions of the network, and this specialization must somehow relate to differences in processing concrete and abstract words. Before considering the exact nature of this specialization, however, we must address the implications of the effects of individual lesions.

The average effects just mentioned correspond to the means of distributions of the effects produced by individual lesions. As neuropsychological theorizing typically involves comparing the relative performance of two or more individual patients, a more direct analogy in the network would be to compare the effects after specific instances of

lesions. As would be expected from the average results, the large majority of such comparisons yield the same findings as the averages indicate. However, the occasional lesion of each type may produce effects that are exactly opposite to those produced by most quantitatively equivalent lesions. Thus, while most $0 \Rightarrow I$ lesions selectively impair concrete word reading, a specific $0 \Rightarrow I$ lesion may selectively impair abstract word reading. This state of affairs leads to the somewhat disturbing observation that two instances of quantitatively equivalent lesions can give rise to a statistically reliable double dissociation between concrete and abstract word reading (see Table 4). Furthermore, similar findings arise when we consider the patterns of errors that occur after damage—specifically, the co-occurrence of visual errors with semantic errors, and the effects of concreteness on the rates of visual errors. It would seem, then, that the observation of a double dissociation does not even indicate functional specialization, as Shallice (1988) suggests, for how can the same portion of a mechanism be “specialized” in two different ways?

One reaction to these findings might be to view them as a basis for dismissing the network, and perhaps connectionist modeling more generally, as irrelevant to the study of human cognitive neuropsychology. After all, if the same damage to a network can produce opposite results, what can the results of *any* simulation tell us? Such an argument is similar to that raised by Massaro (1988) against connectionist modeling of normal cognitive processes. Notice, however, that the *behavior* of the damaged network is no more counterintuitive than that of the relevant patients: selective impairment in reading a particular class of words. Also notice that the effects were produced by random lesions to the network and not by a subtle manipulation on the part of the experimenter designed to produce the desired effect (cf. Wood, 1978). What is perhaps unsatisfactory is not so much the behavioral results as the implied explanation for the corresponding results in patients. Thus, to understand the implications of the findings for interpreting human cognitive impairments, we must first consider in more detail how the effects of damage in a network should be related to the effects of brain damage in humans.

The critical question is, are the effects of a single lesion in a connectionist network relevant to human neuropsychology? At this stage in the development of connectionist models of cognitive processes, the answer must be no. The reason is that even the largest current-day simulations are of a vastly smaller scale than the portions of the human cognitive system to which they correspond. As a result, it is far more likely in a network than in a human that an individual lesion would give rise to idiosyncratic effects that do not reflect the general properties of the system. This effect is evident in its extreme form in simulations in which double dissociations are produced by lesions of individual units (e.g., Bullinaria & Chater, 1993; Sartori, 1988; Wood, 1978). Even though specific lesions in the current simulation involve the removal of large numbers of connections, the relatively small scale of the simulation (relative to the presumed size of the human word reading system) increases the proportion of lesions expected to have idiosyncratic effects. Thus, identifying such lesions tells us more about the limitations of the scale of the simulation than it does about the general computational properties of the system and how they might correspond to those of the human cognitive system.

Of course, no matter how large the cognitive system is assumed to be, it is always possible that specific lesions may produce effects that are not representative of the general functional specialization of the damaged portion of the system. As long as damage is characterized as complete loss of an isolable component in an information processing system (see, e.g., Figure 2), problematic questions concerning the distribution of effects caused by lesions do not arise. Unfortunately, such a characterization belies the actual complexity of effects that would plausibly be expected after damage to the cognitive system. In any sufficiently detailed neural implementation of a cognitive process, random variations in quantitatively equivalent lesions would be expected to produce a distribution of effects. Depending on the variance of this distribution, the effects of individual lesions may not be representative of the distribution.

This possibility raises concerns about the reliance in cognitive neuropsychology on single-case studies. Without some information on how the performance of a particular patient relates to the performance of other patients with equivalent deficits, effects that appear to provide insight into the functional organization of the cognitive system may simply be statistical flukes. In traditional theoretical formulations, such outliers are deemed of particular relevance because they reflect “pure” cases in which only a single component has been damaged. In alternative formulations in which the effects of damage can be graded, it becomes more important to understand the full distribution of effects across patients rather than the peculiar behavior of just a few. This proposal runs contrary to the all-too-common tendency in cognitive neuropsychological methodology to seek out and study in detail just those patients that exhibit the most unusual symptoms (recall Ellis’ comments on “word meaning deafness” quoted in the Introduction). At one level, the proposed shift amounts simply to emphasizing the need for replication of the findings in one patient in other ones with similar impairments.⁹ More fundamentally, the relevance of neuropsychological data for contributing to our understanding of the normal cognitive system depends on the degree to which the data reflect general rather than

⁹Notice that replication of single-case results is different from group studies in which patients are selected on coarse behavioral or anatomical bases (also see Caramazza, 1986; McCloskey, 1993, for related discussion).

idiosyncratic properties of the system (cf. Caramazza's, 1986, "universality" assumption).

For the same reason, investigations of the effects of damage in a connectionist network must demonstrate systematic properties that result from damage, not those that arise from nonsystematic (random) aspects of the training or testing (lesioning) procedure. On this basis, the selective impairments in concrete versus abstract word reading as reflected in the *average* effects of lesions to $O \Rightarrow I$ versus $S \Rightarrow C$ connections, respectively, constitute a valid demonstration of a double dissociation in a connectionist network. The next question to ask, then, is why do the dissociations occur in the network, and how do they help explain the corresponding dissociations in patients?

Plaut and Shallice (1993) explain the double dissociation of concrete and abstract word reading after damage to the network in the following way:

As abstract words have fewer semantic features, they are less effective than concrete words at engaging the semantic clean-up mechanism, and must rely more heavily on the direct pathway. Concrete words are read better under lesions to this pathway because of the stronger semantic clean-up they receive. . . . Under severe damage to [the clean-up pathway], the processing of most concrete words is impaired but many abstract words can be read solely by the direct pathway, producing an advantage of abstract over concrete words in correct performance. (p. 460)

At the heart of this explanation is the claim that the network develops stronger semantic attractors for concrete words than for abstract words. In support of this claim, Plaut and Shallice provide evidence that the clean-up units are driven to more extreme values (i.e., closer to 0 or 1) by the semantics of concrete words than by those of abstract words. The network learns stronger attractors for concrete words because their greater number of semantic features provide more opportunities for small subsets of features (e.g., *has-legs*, *living*, *on-ground*) to reliably predict other features (e.g., *does-run*). The role of the clean-up units is exactly to learn to implement such semantic "microinferences" (Hinton, McClelland, & Rumelhart, 1986). Attractors emerge from the collective influence of a large number of microinferences.

An important implication of this more general interpretation is that other possible differences between concepts that affect the relative strength of their attractors, if embedded in a similar network architecture, would be expected to give rise to analogous effects. For instance, McRae, de Sa, and Seidenberg (1997) suggest that the semantic features of natural kinds (e.g., animals) are more highly intercorrelated than are the features of artifacts (e.g., tools), and that this difference can account for normal subjects' faster naming latencies of pictures of natural kinds versus artifacts. As greater feature intercorrelation directly increases the available microinferences, this factor would affect the strength of attractors in much the same way as absolute numbers of features did in the current simulation. In fact, McRae and colleagues demonstrate this property in a simple attractor network (Hopfield, 1982).

Similarly, Breedin, Saffran, and Coslett (1994) suggest that semantic representations of concrete words depend more heavily on support from interactions with high-level visual representations. Thus, damage to these visual representations would selectively impair performance on semantic tasks involving concrete words, as they found in their patient DM. This proposal provides yet another alternative account for why the attractors for concrete words might be stronger than those for abstract words. In fact, Warrington and McCarthy (1987) put forth a similar account of the double dissociation in the comprehension of natural kinds versus artifacts (Warrington & Shallice, 1984), and Farah and McClelland (1991) supported and extended this account with an attractor network implementation.

Thus, critically, what would appear to be quite different proposals—numbers of semantic features, feature intercorrelations, interaction with visual representations—for various word class effects in semantic tasks can be seen as alternative versions of essentially the same explanation: word types differ in the strength of their semantic attractors. In this way, an understanding of the computational properties of attractors serves to unify what would otherwise appear to be disparate accounts of similar empirical phenomena.

One implication of the relative strength of attractors in the concrete/abstract network is that concrete and abstract words are differentially sensitive to damage at different locations in the network. How are these effects related to the standard notion of "functional specialization" (Shallice, 1988)? It would be a mistake to claim that the direct pathway is specialized for abstract words while the clean-up pathway is specialized for concrete words. Both pathways are involved in processing both types of words. However, they make different contributions in the course of this processing: the direct pathway generates an initial approximation of the semantics of the stimulus word which are gradually refined by the clean-up pathway into the exact semantics of the word. What distinguishes concrete and abstract words is not to be found in the structure of the system but rather in its *functional* properties. This is a direct consequence of the representational status of words in the network. In the general theoretical perspective, a word is not

a structural entity to be located somewhere in the system, but rather the functional consequence of the way in which different types of information (e.g., orthographic, semantic, phonological) interact.

Thus, there *is* functional specialization in the network, but the nature of the specialization does not directly correspond to the observed behavioral effects under damage (i.e., selective impairments in reading concrete vs. abstract words). In this way, the system violates Caramazza's (1986) "transparency" assumption and would seem to pose a problem for standard assumptions on how to use neuropsychological data to constrain cognitive theorizing. In particular, it raises the spectre, as expressed by Shallice's (1988) quote in the Introduction, that observed behavioral dissociations "might throw no useful light" on the nature of the underlying functional specialization. However, the implication of the current work is not that neuropsychological theorizing is fruitless (cf. Kosslyn & Intriligator, 1992; Kosslyn & Van Kleeck, 1990), but rather that it must be done in context of specific computational assumptions of how the cognitive system operates normally and under damage (see Farah, 1994, for discussion).

The modularity hypothesis has been a powerful theoretical tool in neuropsychology precisely because it provides an intuitive framework for inferring the effects of damage in an information processing system. Unfortunately, it is unnecessarily restrictive in the kinds of processes it can express, and it brings with it stringent methodological constraints on the selection criteria of patients for detailed study. Connectionist modeling, by contrast, provides a richer formalism in which to investigate the effects of damage in interactive systems. The computational principles that emerge from such systems may provide insight into the full distribution of cognitive impairments caused by brain damage in humans.

References

- Andrews, S. (1982). Phonological recoding: Is the regularity effect consistent? *Memory and Cognition*, *10*, 565–575.
- Barry, C., & Richardson, J. T. E. (1988). Accounts of oral reading in deep dyslexia. In H. A. Whitaker (Ed.), *Phonological processing and brain mechanisms* (pp. 118–171). New York: Springer-Verlag.
- Breedin, S. D., Saffran, E. M., & Coslett, H. B. (1993, November). The selective loss of concrete words: A case study. In *Proceedings of the 34th Annual Meeting of the Psychonomic Society* (p. 49). Washington, DC.
- Breedin, S. D., Saffran, E. M., & Coslett, H. B. (1994). Reversal of the concreteness effect in a patient with semantic dementia. *Cognitive Neuropsychology*, *11*, 617–660.
- Brown, W. P., & Ure, D. M. N. (1969). Five rated characteristics of 650 word association stimuli. *British Journal of Psychology*, *60*, 223–250.
- Buchanan, L., Hildebrandt, N., & MacKinnon, G. E. (1994). *Phonological processing of nonwords by a deep dyslexic patient: A rose is implicitly a rose*. Manuscript submitted for publication.
- Bullinaria, J. A., & Chater, N. (1993). Double dissociation in artificial neural networks: Implications for neuropsychology. In *Proceedings of the 15th Annual Conference of the Cognitive Science Society* (pp. 283–288). Hillsdale, NJ: Erlbaum.
- Caplan, D. (Ed.). (1992). *Language: Structure, processing, and disorders*. Cambridge, MA: MIT Press.
- Caramazza, A. (1984). The logic of neuropsychological research and the problem of patient classification in aphasia. *Brain and Language*, *21*, 9–20.
- Caramazza, A. (1986). On drawing inferences about the structure of normal cognitive systems from the analysis of patterns of impaired performance: The case for single-patient studies. *Brain and Cognition*, *5*, 41–66.
- Caramazza, A., & Hillis, A. E. (1990). Where do semantic errors come from? *Cortex*, *26*, 95–122.
- Caramazza, A., & McCloskey, M. (1988). A case for single-patient studies. *Cognitive Neuropsychology*, *5*, 517–528.
- Chomsky, N. (1980). Rules and representations. *Behavioral and Brain Sciences*, *3*, 1–61.
- Coltheart, M. (1980a). Deep dyslexia: A review of the syndrome. In M. Coltheart, K. Patterson, & J. C. Marshall (Eds.), *Deep dyslexia* (pp. 22–48). London: Routledge & Kegan Paul.
- Coltheart, M. (1980b). Deep dyslexia: A right-hemisphere hypothesis. In M. Coltheart, K. Patterson, & J. C. Marshall (Eds.), *Deep dyslexia* (pp. 326–380). London: Routledge & Kegan Paul.
- Coltheart, M. (1985). Cognitive neuropsychology and the study of reading. In M. I. Posner, & O. S. M. Marin (Eds.), *Attention and performance XI* (pp. 3–37). Hillsdale, NJ: Erlbaum.
- Coltheart, M. (Ed.). (1987). *Attention and performance XII: The psychology of reading*. Hillsdale, NJ: Erlbaum.
- Coltheart, M., Curtis, B., Atkins, P., & Haller, M. (1993). Models of reading aloud: Dual-route and parallel-distributed-processing approaches. *Psychological Review*, *100*, 589–608.
- Coltheart, M., & Funnell, E. (1987). Reading and writing: One lexicon or two? In D. A. Allport, D. G. MacKay, W. Printz, & E. Scheerer (Eds.), *Language perception and production: Shared mechanisms in listening, speaking, reading and writing* (pp. 313–339). New York: Academic Press.

- Coltheart, M., Patterson, K., & Marshall, J. C. (Eds.). (1980). *Deep dyslexia*. London: Routledge & Kegan Paul.
- Coltheart, M., Patterson, K., & Marshall, J. C. (1987). Deep dyslexia since 1980. In M. Coltheart, K. Patterson, & J. C. Marshall (Eds.), *Deep dyslexia* (pp. 407–451). London: Routledge & Kegan Paul, 2 edition.
- Coughlan, A. K., & Warrington, E. K. (1981). The impairment of verbal semantic memory: A single case study. *Journal of Neurology, Neurosurgery, and Psychiatry*, *44*, 1079–1083.
- Crick, F. H. C. (1989). The recent excitement about neural networks. *Nature*, *337*, 129–132.
- Dell, G. S. (1986). A spreading-activation theory of retrieval in sentence production. *Psychological Review*, *93*, 283–321.
- Dell, G. S., Juliano, C., & Govindjee, A. (1993). Structure and content in language production: A theory of frame constraints in phonological speech errors. *Cognitive Science*, *17*, 149–195.
- Ellis, A. W. (1987). Intimations of modularity, or, the Modularity of mind: Doing cognitive neuropsychology without syndromes. In M. Coltheart, G. Sartori, & R. Job (Eds.), *The cognitive neuropsychology of language* (pp. 397–408). Hillsdale, NJ: Erlbaum.
- Ellis, A. W., & Marshall, J. C. (1978). Semantic errors or statistical flukes? A note on Allport's "On knowing the meanings of words we are unable to report". *Quarterly Journal of Experimental Psychology*, *30*, 569–575.
- Farah, M. J. (1990). *Visual agnosia: Disorders of object recognition and what they tell us about normal vision*. Cambridge, MA: MIT Press.
- Farah, M. J. (1994). Neuropsychological inference with an interactive brain: A critique of the locality assumption. *Behavioral and Brain Sciences*, *17*, 43–104.
- Farah, M. J., & McClelland, J. L. (1991). A computational model of semantic memory impairment: Modality-specificity and emergent category-specificity. *Journal of Experimental Psychology: General*, *120*, 339–357.
- Fodor, J. A. (1983). *The modularity of mind*. Cambridge, MA: MIT Press.
- Friedman, R. B. (1996). Recovery from deep alexia to phonological alexia. *Brain and Language*, *52*, 114–128.
- Gentner, D. (1981). Some interesting differences between verbs and nouns. *Cognition and Brain Theory*, *4*, 161–178.
- Glosser, G., & Friedman, R. B. (1990). The continuum of deep/phonological alexia. *Cortex*, *26*, 343–359.
- Henderson, L. (1982). *Orthography and word recognition in reading*. London: Academic Press.
- Hillis, A. E., Rapp, B., Romani, C., & Caramazza, A. (1990). Selective impairments of semantics in lexical processing. *Cognitive Neuropsychology*, *7*, 191–243.
- Hinton, G. E., McClelland, J. L., & Rumelhart, D. E. (1986). Distributed representations. In D. E. Rumelhart, J. L. McClelland, & the PDP Research Group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition. Volume 1: Foundations* (pp. 77–109). Cambridge, MA: MIT Press.
- Hinton, G. E., & Shallice, T. (1989). *Lesioning a connectionist network: Investigations of acquired dyslexia* (Technical Report CRG-TR-89-3). Toronto, Ontario, Canada: University of Toronto, Department of Computer Science.
- Hinton, G. E., & Shallice, T. (1991). Lesioning an attractor network: Investigations of acquired dyslexia. *Psychological Review*, *98*, 74–95.
- Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Science, USA*, *79*, 2554–2558.
- Humphreys, G. W., & Evett, L. J. (1985). Are there independent lexical and nonlexical routes in word processing? An evaluation of the dual-route theory of reading. *Behavioral and Brain Sciences*, *8*, 689–740.
- Jones, G. V. (1985). Deep dyslexia, imageability, and ease of predication. *Brain and Language*, *24*, 1–19.
- Kieras, D. (1978). Beyond pictures and words: Alternative information-processing models for imagery effects in verbal memory. *Psychological Bulletin*, *85*, 532–554.
- Kinsbourne, M. (1971). Cognitive deficit: Experimental analysis. In J. L. McGaugh (Ed.), *Psychobiology*. New York: Academic Press.
- Klein, D., Behrmann, M., & Doctor, E. (1994). The evolution of deep dyslexia: Evidence for the spontaneous recovery of the semantic reading route. *Cognitive Neuropsychology*, *11*, 579–611.
- Kosslyn, S. M., & Intriligator, J. M. (1992). Is cognitive neuropsychology plausible? The perils of sitting on a one-legged stool. *Journal of Cognitive Neuroscience*, *4*, 96–106.
- Kosslyn, S. M., & Van Kleeck, M. (1990). Broken brains and normal minds: Why Humpty-Dumpty needs a skeleton. In E. L. Schwartz (Ed.), *Computational neuroscience* (pp. 390–402). Cambridge, MA: MIT Press.
- Lichtheim, L. (1885). On aphasia. *Brain*, *7*, 433–484.
- Marcel, A. J., & Patterson, K. (1978). Word recognition and production: Reciprocity in clinical and normal studies. In J. Requin (Ed.), *Attention and performance VII*. Hillsdale, NJ: Erlbaum.
- Marr, D. (1982). *Vision*. San Francisco, CA: W. H. Freeman.
- Marshall, J. C., & Newcombe, F. (1966). Syntactic and semantic errors in paralexia. *Neuropsychologia*, *4*, 169–176.

- Massaro, D. W. (1988). Some criticisms of connectionist models of human performance. *Journal of Memory and Language*, 27, 213–234.
- McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: Part 1. An account of basic findings. *Psychological Review*, 88, 375–407.
- McClelland, J. L., & Rumelhart, D. E. (1985). Distributed memory and the representation of general and specific information. *Journal of Experimental Psychology: General*, 114, 159–188.
- McCloskey, M. (1993). Theory and evidence in cognitive neuropsychology: A “radical” response to Robertson, Knight, Rafal, and Shimamura (1993). *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 19, 718–734.
- McRae, K., de Sa, V. R., & Seidenberg, M. S. (1997). On the nature and scope of featural representations of word meaning. *Journal of Experimental Psychology: General*, 126, 99–130.
- Morton, J. (1969). The interaction of information in word recognition. *Psychological Review*, 76, 165–178.
- Morton, J. (1981). The status of information processing models of language. *Proceedings of the Royal Society of London, Series B*, 295, 387–396.
- Morton, J., & Patterson, K. (1980). A new attempt at an interpretation, Or, an attempt at a new interpretation. In M. Coltheart, K. Patterson, & J. C. Marshall (Eds.), *Deep dyslexia* (pp. 91–118). London: Routledge & Kegan Paul.
- Newcombe, F., & Marshall, J. C. (1980). Transcoding and lexical stabilization in deep dyslexia. In M. Coltheart, K. Patterson, & J. C. Marshall (Eds.), *Deep dyslexia* (pp. 176–188). London: Routledge & Kegan Paul.
- Nolan, K. A., & Caramazza, A. (1982). Modality-independent impairments in word processing in a deep dyslexic patient. *Brain and Language*, 16, 237–264.
- Paivio, A. (1969). Mental imagery in associative learning and memory. *Psychological Review*, 76, 241–263.
- Paivio, A. (1991). Dual coding theory: Retrospect and current status. *Canadian Journal of Psychology*, 45, 255–287.
- Patterson, K. (1990). Alexia and neural nets. *Japanese Journal of Neuropsychology*, 6, 90–99.
- Patterson, K., Coltheart, M., & Marshall, J. C. (Eds.). (1985). *Surface dyslexia*. Hillsdale, NJ: Erlbaum.
- Patterson, K., & Marcel, A. J. (1977). Aphasia, dyslexia and the phonological coding of written words. *Quarterly Journal of Experimental Psychology*, 29, 307–318.
- Patterson, K., Seidenberg, M. S., & McClelland, J. L. (1989). Connections and disconnections: Acquired dyslexia in a computational model of reading processes. In R. G. M. Morris (Ed.), *Parallel distributed processing: Implications for psychology and neuroscience* (pp. 131–181). London: Oxford University Press.
- Peterson, C., & Anderson, J. R. (1987). A mean field theory learning algorithm for neural nets. *Complex Systems*, 1, 995–1019.
- Plaut, D. C., Behrmann, M., Patterson, K., & McClelland, J. L. (1993, November). Impaired oral reading in surface dyslexia: Detailed comparison of a patient and a connectionist network [Abstract]. In *Proceedings of the 34th Annual Meeting of the Psychonomic Society* (p. 48). Washington, DC.
- Plaut, D. C., & McClelland, J. L. (1993). Generalization with componential attractors: Word and nonword reading in an attractor network. In *Proceedings of the 15th Annual Conference of the Cognitive Science Society* (pp. 824–829). Hillsdale, NJ: Erlbaum.
- Plaut, D. C., McClelland, J. L., Seidenberg, M. S., & Patterson, K. (1994). *Understanding normal and impaired word reading: Computational principles in quasi-regular domains* (Technical Report PDP.CNS.94.5). Pittsburgh, PA: Carnegie Mellon University, Department of Psychology.
- Plaut, D. C., & Shallice, T. (1991). Effects of word abstractness in a connectionist model of deep dyslexia. In *Proceedings of the 13th Annual Conference of the Cognitive Science Society* (pp. 73–78). Hillsdale, NJ: Erlbaum.
- Plaut, D. C., & Shallice, T. (1993). Deep dyslexia: A case study of connectionist neuropsychology. *Cognitive Neuropsychology*, 10, 377–500.
- Plaut, D. C., & Shallice, T. (1994). Word reading in damaged connectionist networks: Computational and neuropsychological implications. In R. Mammone (Ed.), *Artificial neural networks for speech and vision* (pp. 294–323). London: Chapman & Hall.
- Richardson, J. T. E. (1975). The effects of word imageability in acquired dyslexia. *Neuropsychologia*, 13, 281–288.
- Rosch, E., Mervis, C. B., Gray, W., Johnson, D., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, 8, 382–439.
- Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, 323, 533–536.
- Saffran, E. M., Bogyo, L. C., Schwartz, M. F., & Marin, O. S. M. (1980). Does deep dyslexia reflect right-hemisphere reading? In M. Coltheart, K. Patterson, & J. C. Marshall (Eds.), *Deep dyslexia* (pp. 381–406). London: Routledge & Kegan Paul.

- Sartori, G. (1988). From neuropsychological data to theory and vice versa. In G. Denes, P. Bisiacchi, C. Semenza, & E. Andrews (Eds.), *Perspectives in cognitive neuropsychology*. Hillsdale, NJ: Erlbaum.
- Schwanenflugel, P. J. (1991). Why are abstract concepts hard to understand? In P. J. Schwanenflugel (Ed.), *The psychology of word meanings*. Hillsdale, NJ: Erlbaum.
- Seidenberg, M. S. (1985). The time course of phonological code activation in two writing systems. *Cognition*, *19*, 1–10.
- Seidenberg, M. S., & McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review*, *96*, 523–568.
- Seidenberg, M. S., Petersen, A., MacDonald, M. C., & Plaut, D. C. (1996). Pseudohomophone effects and models of word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*, 48–62.
- Seidenberg, M. S., Waters, G. S., Barnes, M. A., & Tanenhaus, M. K. (1984). When does irregular spelling or pronunciation influence word recognition? *Journal of Verbal Learning and Verbal Behaviour*, *23*, 383–404.
- Shallice, T. (1979). Case-study approach in neuropsychological research. *Journal of Clinical Neuropsychology*, *1*, 183–211.
- Shallice, T. (1988). *From neuropsychology to mental structure*. Cambridge: Cambridge University Press.
- Shallice, T., & Coughlan, A. K. (1980). Modality specific word comprehension deficits in deep dyslexia. *Journal of Neurology, Neurosurgery, and Psychiatry*, *43*, 866–872.
- Shallice, T., & Warrington, E. K. (1975). Word recognition in a phonemic dyslexic patient. *Quarterly Journal of Experimental Psychology*, *27*, 187–199.
- Shallice, T., & Warrington, E. K. (1980). Single and multiple component central dyslexic syndromes. In M. Coltheart, K. Patterson, & J. C. Marshall (Eds.), *Deep dyslexia* (pp. 119–145). London: Routledge & Kegan Paul.
- Strain, E., Patterson, K., & Seidenberg, M. S. (1995). Semantic effects in single-word naming. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*, 1140–1154.
- Taraban, R., & McClelland, J. L. (1987). Conspiracy effects in word recognition. *Journal of Memory and Language*, *26*, 608–631.
- Teuber, H. L. (1955). Physiological psychology. *Annual Review of Psychology*, *9*, 267–296.
- Tulving, E. (Ed.). (1983). *Elements of episodic memory*. Oxford: Oxford University Press.
- Vallar, G., & Shallice, T. (Eds.). (1990). *Neuropsychological impairments of short-term memory*. Cambridge: Cambridge University Press.
- van Gelder, T. (1990). Compositionality: A connectionist variation on a classical theme. *Cognitive Science*, *14*, 355–384.
- Van Orden, G. C., Pennington, B. F., & Stone, G. O. (1990). Word identification in reading and the promise of subsymbolic psycholinguistics. *Psychological Review*, *97*, 488–522.
- Warrington, E. K. (1975). The selective impairment of semantic memory. *Quarterly Journal of Experimental Psychology*, *27*, 635–657.
- Warrington, E. K. (1981). Concrete word dyslexia. *British Journal of Psychology*, *72*, 175–196.
- Warrington, E. K., & McCarthy, R. (1987). Categories of knowledge: Further fractionation and an attempted integration. *Brain*, *110*, 1273–1296.
- Warrington, E. K., & Shallice, T. (1984). Category specific semantic impairments. *Brain*, *107*, 829–853.
- Waters, G. S., & Seidenberg, M. S. (1985). Spelling-sound effects in reading: Time course and decision criteria. *Memory and Cognition*, *13*, 557–572.
- Williams, R. J., & Peng, J. (1990). An efficient gradient-based algorithm for on-line training of recurrent network trajectories. *Neural Computation*, *2*, 490–501.
- Wood, C. C. (1978). Variations on a theme by Lashley: Lesion experiments on the neural model of Anderson, Silverstein, Ritz, and Jones. *Psychological Review*, *85*, 582–591.