Word Reading in Damaged Connectionist Networks: Computational and Neuropsychological Implications*

David C. Plaut
Department of Psychology
Carnegie Mellon University
Pittsburgh, PA 15213–3890
plaut+@cmu.edu

Tim Shallice
Department of Psychology
University College
London, England WC1E 6BT
ucjtsts@ucl.ac.uk

In R. Mammone (Ed.) *Artificial neural networks for speech and vision* (pp. 294–323). London: Chapman & Hall, 1994.

1 Introduction

Connectionist networks are also called *neural* networks because of their abstract structural similarity to groups of neurons. Based on this similarity, many researchers believe that computation in these networks reflects important properties of neural computation. One piece of evidence often put forward in support of this claim is that, like brains, connectionist networks tend to degrade gracefully with damage. That is, if some proportion of units and/or connections are removed from a network, performance on a task is typically only partially impaired rather than completely abolished. Most demonstrations of graceful degradation in networks have used only very general measures of performance, such as total error on a task. However, the argument that connectionist computation is fundamentally similar to neural computation would be far more compelling if the *way* in which connectionist networks degraded under damage—their patterns of impaired performance—mirrored the patterns of impaired behavior observed in patients with neurological damage. To the extent that this held, a detailed investigation of the behavior of damaged connectionist networks would provide insight into both normal and impaired human cognition.

A complementary motivation for studying the effects of damage in networks is to extend our understanding of the nature of computation in the networks themselves. Here again, our concern is not just with the development of a network that accomplishes a task, but with understanding *how* the network accomplishes the task—the nature of its representations and processes. In most connectionist research, the adequacy of a network is evaluated by testing how well its performance generalizes to novel external input drawn from the same distribution as the training examples. In a

*We would like to thank Marlene Behrmann for commenting on an earlier draft. All of the simulations described in this paper were run on a Silicon Graphics Iris-4D/240S using an extended version of the Xerion simulator developed by Tony Plate. This research was supported by grant 87-2-36 from the Alfred P. Sloan Foundation.

similar way, damage to a network has the effect of generating unfamiliar activity in the remaining portions of the network . However, damage can affect internal representations in ways that cannot be directly mimicked by manipulations of the external input. Thus, the behavior of the network under damage may provide a more general, and for some purposes, more informative, indication of the nature of the representations and processes the network develops during training.

In studying patients with brain damage, the field of cognitive neuropsychology attempts to relate their patterns of impaired and preserved abilities to models of normal cognitive functioning, with the intent both of explaining the behavior of the patients in terms of the effects of damage in the model, and of informing the model based on the observed behavior of patients [Col85, EY88]. In an analogous fashion, this chapter presents an approach that might be called "connectionist neuropsychology," in which analyses of the effects of damage in connectionist networks are used both to provide a comprehensive, detailed account of the cognitive deficits of a particular class of brain-injured patients, and to clarify the nature of the representations and processes that develop in the networks themselves through learning. To illustrate this approach, we will focus on an acquired reading disorder known as "deep dyslexia," in which patients can pronounce a written word only via its meaning, and occasionally make errors in this process. The chapter begins with a summary of these patients' characteristics and a brief description of a preliminary connectionist model. Following this, results are presented from a systematic investigation of the major design decisions that entered into developing the model, relating to the task definition, the network architecture, the training procedure, and the testing procedure. In the interest of space, some results will only be summarized here; details may be found in [PS93]. The particular emphasis of this chapter will be on results, not described in that paper, that illustrate how studying damaged networks can lead to computational insights that might not arise so clearly within other methodologies. Specifically, results presented here point out some inherent difficulties with distributed output representations, and clarify differences in the computational properties of back-propagation networks and deterministic Boltzmann Machines trained with contrastive Hebbian learning.

1.1 Deep Dyslexia

Brain damage can produce selective impairments in a wide range of cognitive domains, including high-level vision, attention, speech and language, learning and memory, planning, and motor control. The class of impairments which perhaps have received the greatest theoretical attention over the last decade or so are those that involve word reading, the so called "acquired dyslexias." Of these, deep dyslexia is among the most perplexing [CPM80]. Deep dyslexic patients can only read via meaning, as evidenced by their almost complete inability to read meaningless pronounceable letter strings (e.g., MAVE). However, they also have some problems reading words—which have semantics—suggesting that the process by which words access their meanings is also impaired in these patients. The nature of this additional impairment is reflected in the errors that deep dyslexic patients typically make in oral reading—in particular, the occurrence of *semantic* errors (e.g., CAT ⇒ "dog"). However, what makes deep dyslexia such a theoretical challenge is that virtually all patients who make semantic errors also exhibit a peculiar combination of other symptoms. Central among these are other types of errors: *visual* (e.g., CAT ⇒ "cot"), mixed *visual-and-semantic* (e.g., CAT ⇒ "rat"), *derivational* (e.g., WALKED ⇒ "walk"), and *visual-then-semantic* (e.g. SYMPATHY ⇒ "orchestra", presumably via *symphony*). These patients also produce some responses that are completely unrelated to the stimulus (e.g., CAT ⇒ "mug"). Furthermore, their ability to read a

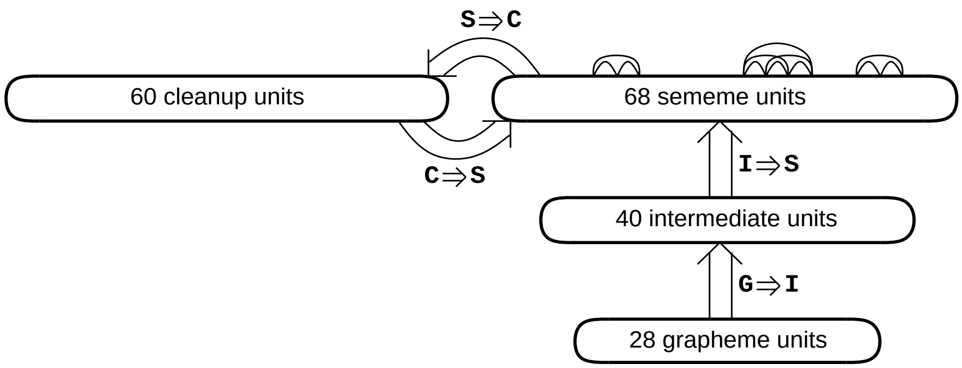

Figure 1: The network architecture used by Hinton and Shallice. Arrows represent sets of connections that were lesioned in the study—they are labeled by the initials of the source and destination layers (e.g., `G⇒I` for grapheme-to-intermediate connections). Only a randomly selected 25% of the possible connections in each of these sets were initially included in the network.

word correctly strongly depends on its part-of-speech (nouns $>$ adjectives $>$ verbs $>$ function words) and its concreteness or imageability (concrete, highly imageable words $>$ abstract, less imageable words). Strangely, the effects of concreteness—a semantic variable—interact with visual similarity in errors, such that abstract words are more likely than concrete words to produce visual errors, and the resulting responses tend to be more concrete than the stimulus (e.g., SCANDAL ⇒ "sandals"). Of these effects, the derivational errors and part-of-speech effects may be secondary to other characteristics [Fun87], but any account of the disorder needs to explain all the other apparently independent symptoms.

1.2 A Preliminary Connectionist Model

Hinton and Shallice [HS91] (hereafter H&S) put forward a connectionist account of why semantic, visual and mixed visual-and-semantic errors co-occur when the process that derives the meanings of words is damaged. Based on previous work by Hinton and Sejnowski with Boltzmann Machines [HS86], they trained a recurrent back-propagation network to map from the written form (i.e., orthography) of 40 three- or four-letter words to a simplified representation of their semantics, described in terms of 68 predetermined semantic features. The architecture of the network, shown in Figure 1, consists of two pathways: a *direct* pathway, from *grapheme* units to *sememe* units via *intermediate* units, that generates initial semantic activity; and a *clean-up* pathway, from the sememes to *clean-up* units and back to the sememes, that iteratively refines these initial semantics into the exact semantics of the presented word. Thus, in solving the task, the network learns to make the pattern of semantic features for each word into an *attractor* in the 68-dimensional space of possible semantic representations. After training, H&S systematically lesioned the network by removing proportions of units or connections, or by adding noise to the weights, and found that the damaged network occasionally settled into a pattern of semantic activity that satisfied response criteria for a word other than the one presented. These error responses were more often semantically similar to the stimulus (i.e., from the same category) and/or visually similar to the stimulus (i.e., overlapped in at least one letter) than would be expected by chance. While the network showed a greater tendency to produce visual errors with damage near the input layer and

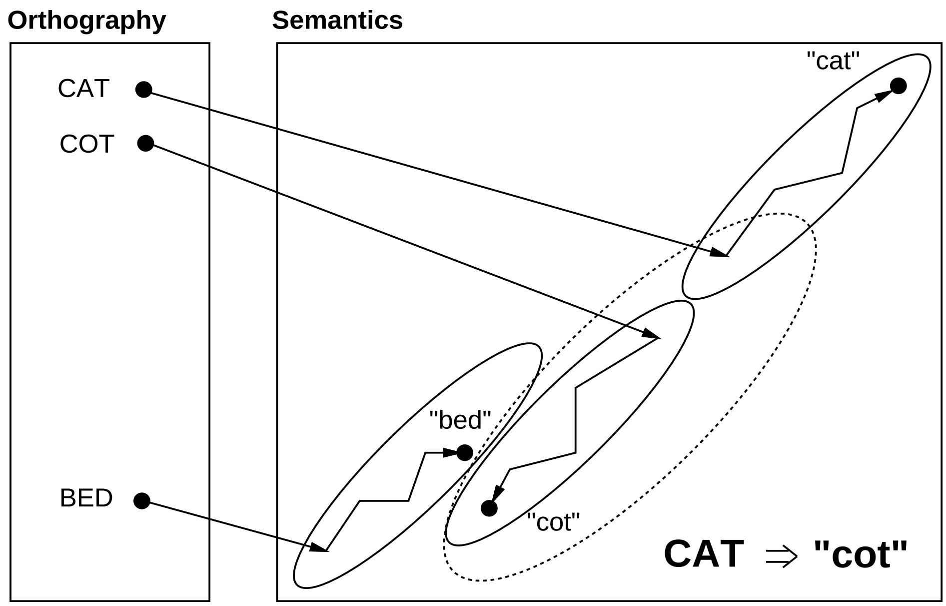

Figure 2: How semantic damage can cause visual errors. The solid ovals depict the normal basins of attraction; the dotted one depicts a basin after damage.

semantic errors with damage near the output layer, both types of error occurred for almost all sites of damage.

The occurrence of semantic errors in the model is straightforward to explain. Damage to the direct pathway corrupts the initial semantic activity caused by a word. If this corrupted pattern now happens to fall within the basin of a neighboring attractor, the operation of the clean-up pathway would cause the network to settle into the semantics of a related word. Similarly, damage to the clean-up pathway alters the layout of the basins themselves, such that the normal initial semantic pattern generated by a word might fall within a neighboring attractor.

Damage to the direct pathway would also be expected to lead to visual errors, since this pathway must rely on visual distinctions among words to generate initial semantic activity that falls within the appropriate attractor basin. What is less obvious, both in patients and in the network, is why damage within semantics should lead to visual errors. H&S provide an account in terms of the nature of the attractors that develop in mapping between two arbitrarily related domains. Connectionist networks have difficulty learning to produce quite different outputs from very similar inputs, and yet, often, visually similar words have unrelated meanings (e.g., CAT and COT). In an attractor network, visually similar words are free to generate similar initial semantic patterns as long as these patterns each fall somewhere within the correct basins of attraction. As a result, in this region of semantic space, neighboring attractors correspond to *visually* similar words (see Figure 2). Semantic damage distorts these basins, occasionally causing the normal initial semantic pattern of a word to be captured within the basin of a visually similar word. Essentially, the layout of attractor basins must be sensitive to both visual and semantic similarity, and so these metrics are reflected in the types of errors that occur as a result of damage.

H&S's simulation provides a unified account of the nature and co-occurrence of semantic, visual, and mixed visual-and-semantic errors in deep dyslexia. By contrast, most previous explanations (e.g., [MP80]) have had to resort to proposing separate, independent lesions—one producing semantic errors and the other producing visual errors. Thus, these accounts provide no principled explanation of why virtually all patients who make semantic errors also make visual errors (i.e., why patients who have one lesion almost always have the other). H&S demonstrated that this co-occurrence of error types is a natural consequence of the effects of single lesions in a network that maps between visual and semantic representations of words.

Although encouraging, H&S's work is limited in two important ways. The first is that only a few of the many characteristics of deep dyslexic patients were simulated. To constitute an adequate account of these patients, the approach would have to be extended to encompass the remaining major characteristics as well—particularly, the other error types and the effects of concreteness/imageability. The second limitation is that, although H&S attribute their results to general properties of distributed representations and attractors, they investigated only a single type of network that inevitably had many specific features. They implicitly assumed that these specific features did not significantly contribute to the overall behavior of the network under damage. Clearly it would be impossible to evaluate and improve on every aspect of the H&S model. In the following sections, each of the major design decisions that went into developing the model are systematically explored: the definition of the task of reading via meaning, the specification of a network architecture, the use of a particular training procedure, and the application of a testing procedure for evaluating the network's behavior under damage. The first issue we address is the testing procedure since its results are used in later sections.

2 The Testing Procedure

Most data on deep dyslexic reading comes from tasks in which the patient produces a verbal response to a visually presented word. Since the output of the H&S model to a letter string consists of a pattern of semantic activity, some *external* procedure is needed to convert this pattern into an explicit response so that it can be compared with the oral reading responses of deep dyslexic patients. The procedure H&S used compares the semantic activity produced by the network with the correct semantics of all known words, selecting the closest-matching word as long as the match is sufficiently good (the *proximity* criterion) and sufficiently better than any other match (the *gap* criterion). The rationale for these criteria is that semantic activity that is too unfamiliar or ambiguous would be unable to drive an output system effectively. In this way H&S's use of response criteria differs from approaches that simply take the best-matching known output as the response regardless of the quality of the match (e.g., [PSM90, SR87]).

However, these response criteria were inadequately motivated and were only indirectly verified as appropriate. In particular, while it may be reasonable that semantics which failed the criteria could not drive an output system, no evidence was given that semantics which satisfied the criteria could succeed in generating a response. Furthermore, the criteria are insensitive to the relative semantic and phonological discriminability of words and so may be inadvertently biased towards producing certain effects. Finally, a best-match procedure is a rather powerful operation, requiring considerable knowledge about the words on which network has been trained. If too much of the difficulty of a problem is solved by the assumed mechanisms for generating the input or interpreting the output, the role of the network itself becomes less interesting [LB88, PP88]. This is especially ironic as a best-match (categorization) process is exactly the sort of operation at which connectionist networks are supposed to excel [HA81, Hop82].

Thus, it would be a significant advance over the use of response criteria to extend the H&S model to derive an explicit phonological response on the basis of semantic activity. However, it turns out that developing such a network involves overcoming difficulties which are fairly general to connectionist networks and have arisen in a number of contexts (e.g., [NM91, RM86, SM89]). In the present domain, the problem is that the damaged network produces phonological responses which are inappropriate "blends" of the pronunciations of known words. In this section, we illustrate this problem and demonstrate a method for overcoming it, allowing us to replicate H&S's results using networks that map from orthography to phonology via semantics.

2.1 Phonological blends

The problems that occur in implementing an effective output system are best illustrated by describing what happens when the most straightforward procedure is used. Specifically, we develop an output network analogous to the input network, but which takes as input the semantic representation of a word and produces a phonological representation of the word. This network is then combined with an input network that maps from orthography to semantics (essentially identical to the H&S model), resulting in a much larger network that maps from orthography to phonology via semantics.

The input to the network consists of the 40 semantic representations that served as output in the H&S model. A phonological output representation was defined in terms of 33 position-specific *phoneme* units (see [PS93] for details). For each word, exactly one unit in each of three positions is active, possibly including a unit in the third position that explicitly represents the absence of a third phoneme. This representation allows the units that represent alternative phonemes in the same position to compete in a "winner-take-all" fashion.

In order to minimize the number of independent assumptions in the complete network, the architecture of the output network was designed to be as similar as possible to that of the H&S input network. The sememe (input) units were connected to a group of 40 intermediate units, which were in turn connected to the 33 phoneme units. A group of 60 clean-up units were interconnected with the phoneme units. As in the original H&S network, only a random fourth of the possible connections in each of these pathways was included. In addition, the competing phoneme units for each position were fully interconnected. The resulting network had a total of 2410 connections.

The output network was trained in exactly the same manner as the H&S network, using "back-propagation through time" [RHW86, WP90]. After about 1500 sweeps through the set of words, the network successfully activated each phoneme unit to within 0.1 of its correct state for each word over the last three of eight iterations. This output network was then combined with an input network, identical to the one H&S used, that had been similarly trained to generate semantics from graphemic input. The sememe units of the input network replaced the input units of the output network. The resulting network, shown in Figure 3, had a total of 6110 connections. This combined network was trained further by fixing the weights of the input network and running the entire network for 14 iterations on each input, allowing the output network to adapt. This additional training was required to ensure that the output network operated correctly when receiving input from the input network (which need not be correct until iteration 6) instead of being clamped throughout its operation. Fixing the weights of the input network ensured that it continued to generate the correct semantics of each word. After an additional 34 sweeps through the training

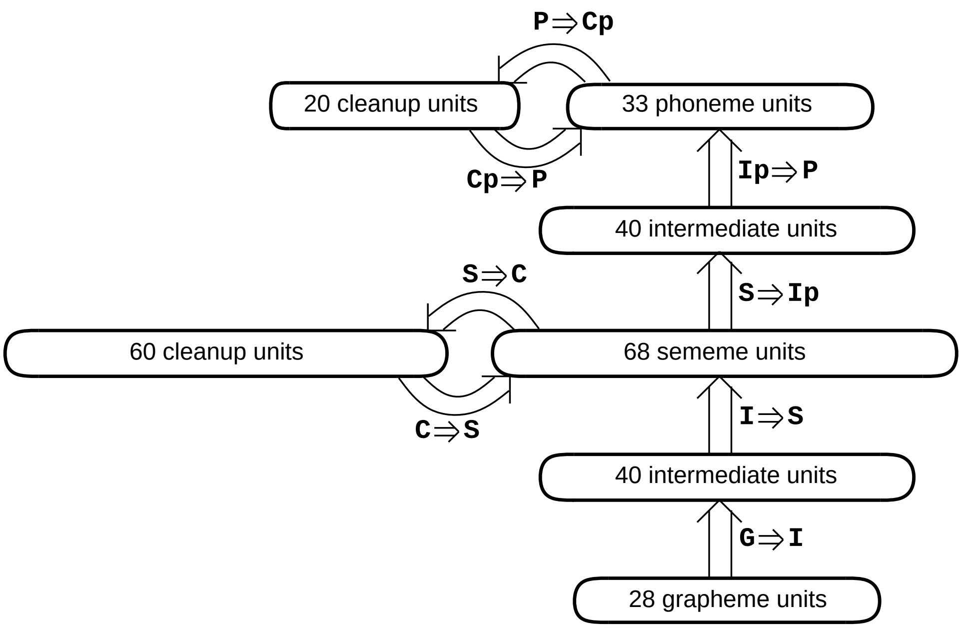

Figure 3: The architecture of a network that maps from orthography to phonology via semantics. Notice that the names of sets of connections involving the intermediate and clean-up units in the phonological output network are subscripted with a *p* to differentiate them from the corresponding sets of connections in the input network.

set, the combined network succeeded in producing the correct phonemes of each word given its graphemes as input.

Because damage will impair the ability of the network to derive the correct pronunciations of words, we need some way of deciding whether corrupted phonological activity constitutes a well-formed pronunciation. Given our phonological representation, a natural criterion is to require that exactly one phoneme unit be active in each of the three positions in order to produce a response. Since units have real-valued outputs which are rarely 0 or 1, we need a more precise definition of "active" and "inactive." The criterion we use is that the most active phoneme at each position is included in the response if its likelihood, relative to the competing phonemes at that position, exceeds a *phonological response criterion* of 0.6.[1] If, at each position, exactly one phoneme satisfies this criterion, the concatenation of these phonemes is produced as the response; otherwise, the phonological activity is considered ill-formed and the network fails to respond. It is important to point out that this type of criterion is quite different from the H&S criteria, which ensure that an output is semantically familiar (i.e., near the meaning of a known word). The criterion we employ does not rely on any knowledge of the particular words the network has been trained on—it considers only the *form* of the output representation.

Each of the four main sets of connections in the input network was subjected to "lesions" by

[1]More formally, if y_i is the output of phoneme unit i, and d_i is its smallest difference from 0 or 1 (i.e., $d_i = y_i$ if $y_i \leq 0.5$ and $1 - y_i$ otherwise), then the network produces a response if, for every position p, $\prod_{i \in p} d_i > 0.6$ and exactly one $y_i > 0.5$. The product is the probability of the most likely binary output vector at the position when the states of the phoneme units are interpreted as independent probabilities. Thus, the response procedure is closely related to the maximum-likelihood interpretation of the cross-entropy error function used to train the network [Hin89a].

choosing at random and removing a proportion of the connections. A wide range of severities were investigated: 0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.4, 0.5, and 0.7. Twenty instances of each location and severity of lesion were carried out, and correct, omission, and error responses were accumulated according to the above procedure. An error response was categorized as visually similar if it shared at least one letter in the same position with the stimulus, and was categorized as semantically similar if it belonged to the same semantic category as the stimulus.[2] In addition, the nature of the output representation and criterion creates a new type of "blend" error consisting of a literal paraphasia—a phonologically reasonable output that does not correspond to a word known to the network. Thus, each error response produced by the damaged network can be classified as visual, visual-and-semantic, semantic, blend, or *other* (unrelated).

Figure 4 presents the average rates of each error type for each lesion location. The most striking aspect of the results is the high rate of blends. These errors stand in sharp contrast to the behavior of deep dyslexics, who very rarely produce nonword responses in oral reading (see [CPM80, Appendix 2]). Table 2.2 presents some typical examples of blend errors produced by the network under various lesions. The semantic activity produced by each input is characterized by its proximity (i.e., normalized dot-product) with the semantics of the two nearest known words. It is informative to compare the phonology of these words with the response of the network. Semantic activity that is near two words often produces a phonological output that is a mixture of the words' phonemes (e.g., PIG (+RAM) ⇒ `/p a g/`), which is why these errors are called "blends." Occasionally, new phonemes are introduced under the pressure of mixed semantics (e.g., DOG (+CAT) ⇒ `/l a g/`). Interestingly, semantics that would easily satisfy H&S's criteria for a correct response may still be sufficiently corrupted for the output system to produce a blend (e.g., HOCK (*prox* `0.88`, *gap* `0.12`) ⇒ `/h u k/`). On the other hand, semantics that are quite far from any known word may still produce a response, albeit incorrect (e.g., RUM(*prox* `0.66`) ⇒ `/h aw m/`). Clearly the current output system behaves quite differently from what the H&S criteria assume about a response system.

2.2 An Explanation for Blends

In attempting to understand why blends occur, it is important to keep in mind that *any* pattern of activity that the network settles into is an attractor that has developed in the course of training.[3] We know that the network develops appropriate attractors for the 40 words since it produces correct responses when presented with their semantics. However, in the course of training the network develops other, spurious attractors. These attractors tend to be patterns that are combinations of trained patterns because, when the phonology of a word is trained as a response, other phonological patterns are also reinforced to the extent that they overlap with the trained pattern. The existence of spurious attractors is a well-known property of associative networks [Hop82] and is one way of characterizing their limited storage capacity. The existence of these additional attractors is not a problem during normal operation because inputs that would settle into them are never presented. In fact, they are not a problem for any test of generalization involving novel input that is sufficiently similar to familiar input (i.e., near in feature space or drawn from the same distribution) so as to

[2] In addition to visual and semantic similarity, errors can now be phonologically similar—that is, have overlapping phonemes. Since visual and phonological similarity are highly correlated, for the present purposes we will consider such errors to be visual—see [PS93] for more detailed discussion.

[3] Actually, it would be more accurate to say that training has produced the *potential* for this pattern to be an attractor given some input.

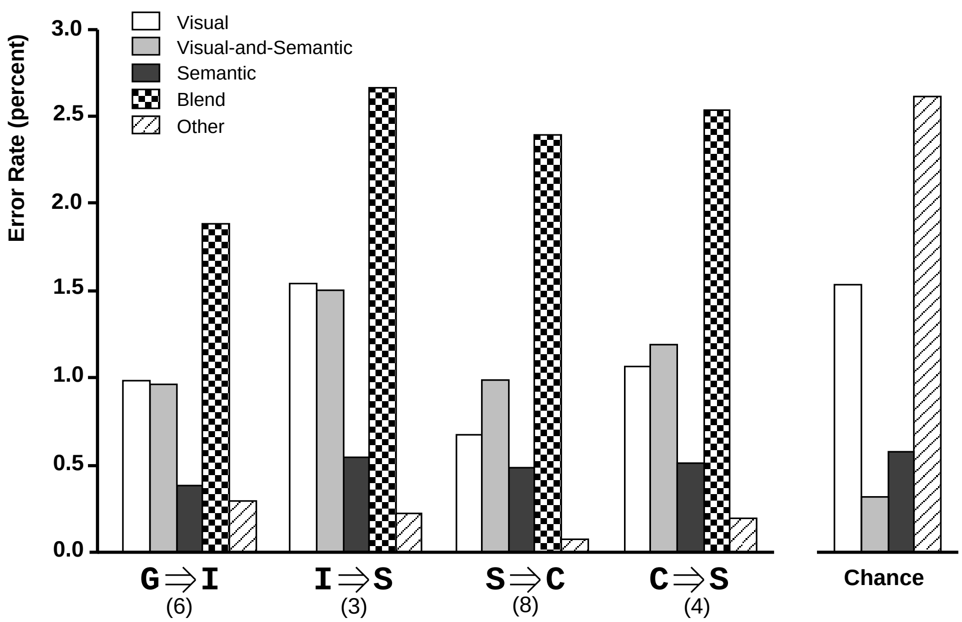

Figure 4: Error rates produced by lesions to each main set of connections in the input network. "Chance" is the distribution of error types if responses were chosen randomly from the word set. Its absolute height is set arbitrarily—only the relative rates are informative. Results are averaged over lesion densities which produced an overall correct response rate between approximately 20% and 80%. The number of lesion severities included in the calculation of error rates is indicated in parentheses below the label for each lesion location.

Table 1: Examples of nonword "blend" errors produced by the network.

Input	Response	Nearest Word	Semantics Best	*prox*	Next	*prox*	Lesion
RIB	/r u d/	MUD	RIB	0.79	GUT	0.65	G⇒I(0.15)
DOG	/l a g/	LOG	DOG	0.88*	CAT	0.79	G⇒I(0.20)
PIG	/p a g/	PIG	PIG	0.86	RAM	0.82	G⇒I(0.25)
LIP	/r a b/	RAM	RIB	0.71	LIP	0.67	G⇒I(0.50)
HOCK	/h u k/	HOCK	HOCK	0.88*	RUM	0.76	I⇒S(0.05)
RUM	/h aw m/	HAM	HAM	0.66	PORK	0.63	I⇒S(0.25)
CUP	/k a g/	CAN	CUP	0.78	CAN	0.76	I⇒S(0.40)
RAT	/r a g/	RAM	RAT	0.97*	DOG	0.73	C⇒S(0.05)
HAM	/h u m/	RUM	BUN	0.77	HAM	0.73	C⇒S(0.25)
LEG	/p o g/	LOG	POP	0.70	LEG	0.64	C⇒S(0.50)
CAN	/k u n/	CAN	CAN	0.96*	MUG	0.80	S⇒C(0.15)
DUNE	/dy o n/	DUNE	TOR	0.81	DUNE	0.81	S⇒C(0.20)
COW	/k u g/	MUG	COW	0.90*	PIG	0.80	S⇒C(0.70)

Note. "Nearest Word" is the word whose phonological representation has the closest proximity to the phonological output of the network. "Semantics" lists the best and next-best words whose semantic representations have the closest proximity *prox* to the semantic activity produced by the network. Semantics that satisfy the Hinton & Shallice response criteria are marked with an asterisk.

fall into the same attractor basins. However, damage to the input network often generates semantic activity which is quite unlike any of the inputs on which the output network has been trained. When this semantic activity consists of a mixture of the semantic features of two words (e.g., PIG and RAM), rather than fall into the attractor for one or the other of these words (either producing a correct response or a conventional error) the network occasionally settles into a spurious attractor for a combination of the phonemes of the two words (e.g., /p a g/), resulting in a blend.

Viewed another way, blends are the result of the natural tendency of connectionist networks to give similar outputs to similar inputs. This property is one of the major attractions of these networks because it enables them to generalize appropriately in many tasks when presented with novel input which is similar to trained input. However, what constitutes an appropriate generalization depends on the task. Consider Seidenberg and McClelland's model of word pronunciation [SM89], which maps from the orthography to the phonology of single-syllable words. The model generalizes to pronounce nonwords by combining the common pronunciations of subsets of its letters, producing a phonological output that is different from that of any known word. Thus, in this task a blend at the level of phonemes is the *correct* response to a novel input, and lexicalization (i.e., producing the exact pronunciation of a similar word) would be inappropriate. In fact, one of the problems with the Seidenberg & McClelland model is that, in response to a nonword, the model occasionally produces an inappropriate blend *at the level of phonemic features*. For example, when presented with the letter string VOST the network produces a blend of the vowel pronunciations of LOST and POST rather than choosing one or the other (J. McClelland, personal communication).[4] Thus, the problem of blends occurs when a network is not sufficiently constrained at the appropriate level of structure in the output: for the Seidenberg & McClelland task this is the phonemic level; for our task it is the lexical level (see also [RM86, SR87]).

2.3 Eliminating Blends

One way to eliminate blends would be to present the network with all possible patterns of semantic activity and explicitly train it to produce no response except to those patterns that correspond to known words. Such a procedure is unacceptable for both empirical and computational reasons: it involves presenting the network with far more information than is available to readers, and it would be intractable to train the network on a large fraction of the exponential number of possible semantic patterns. A better approach is to present only known words, but alter the training procedure in such a way that the network develops much larger and stronger basins of attraction for these words.[5] In this way, initial phonological patterns that are a mixture of the phonemes of two words will be much more likely to fall into the attractor of one or the other of the words, rather than into a spurious attractor for a blend. Developing strong attractors for known words is equivalent to having a strong "lexical bias" in the responses of the network.

In the original architecture with 25% connectivity density, the probability that any clean-up unit would receive connections from three particular phonemes, or receive connections from two and send to a third, is only $0.25^3 = 0.016$. Hence it is unlikely that individual clean-up units can effectively bind together the phonemes of each word—these units must work together to appropriately constraint the phoneme units. To allow clean-up units to more directly constrain combinations of phonemes, a slightly different architecture will be used from the previous one. Rather than use 60 clean-up units which are each interconnected with a random fourth of the phoneme units, only 20 clean-up units will be used, but these will be fully interconnected with all of the phoneme units. The resulting network has only about 330 more connections. Notice that, with only 20 clean-up units, the network cannot devote a single unit to each word. Nonetheless, each of these units can have a more powerful influence on phonological activity than could less-densely connected units.

Our training strategy will be to develop each output network incrementally. First, the phoneme and clean-up units will be trained on noisy versions of the pronunciations of words in order to develop strong attractors for these patterns, independent of any input from semantics. This phonological clean-up pathway will then be fixed, and a direct pathway from semantics to phonology will

[4] In general, the model often produces nonword pronunciations that differ from what normal subjects would consider the correct pronunciation ([BTMS90], but see [SM90]), suggesting that it has not sufficiently learned the appropriate regularities both between and within the phonemes of word pronunciations.

[5] The relationship between the strength of an attractor and the size of its basin of attraction is somewhat subtle. Given unlimited settling time in an undamaged network, attractors with larger basins are stronger in the sense that they pull more distant patterns to them. However, attractors with "deeper" basins (i.e., those representing activity patterns that better satisfy the constraints imposed by the input and weights) are more robust with limited settling time (as in our networks) or under damage, and are in this sense stronger than attractors with larger, more shallow basins. A later section describes simulations using contrastive Hebbian learning in a deterministic Boltzmann Machine, in which strong attractors develop naturally so that no specific training techniques are required to eliminate phonological blends under damage.

be trained, first separately, then with the phonological clean-up added, and finally with its input generated by the input network.

This training procedure differs from the standard approach in two main ways: the use of noisy input and incremental training. In generating noisy input for an example, the activity of each input unit will be moved from 0.0 or 1.0 towards 0.5 by the absolute value of a random number drawn from a gaussian distribution with mean 0.0 and fixed standard deviation. The target states for the output units are unchanged. Training on noisy input amounts to enforcing a particular kind of generalization: inputs which are *near* known patterns must give identical responses. Thus the basin of attraction for each trained pattern must be at least large enough to include the patterns that can be generated from it with the amount of noise used during training. An additional effect of training on noisy input is that there is a pressure for weights to remain small so that the effect of the noise on the rest of the network is minimized. This influence, much like "weight decay" [Hin89a], causes the knowledge of the task to be more evenly distributed across all of the connections, making the network more uniformly robust to lesions [FM91].

Incremental training has two main advantages. First, it reduces the computational demands of training, since the time to train a connectionist network with back-propagation scales much worse than linearly in the size of the network [PH87]. Second, and more important for our purposes, training parts of the network separately encourages each part to accomplish as much of the task as possible, without relying on the strengths of the other parts. Specifically, when training the complete network, if the direct pathway can generate reasonable phonology from even noisy semantics, there is less pressure on the phonological clean-up pathway to develop strong attractors for the correct patterns. Training them separately forces them each to compensate for the noise *independently* so that their combination is more robust.

The phonological clean-up pathway of the output network was trained to produce the correct phonemes of each word during the last three of six iterations when presented with these phonemes corrupted by gaussian noise with a standard deviation of 0.25. Because the phoneme units are both the input and output units for this stage of training, the phonemes cannot be presented by clamping the states of these units. Rather, these units were given an external input throughout the six iterations which, in the absence of other inputs, would produce the specified corrupted activity level. This technique is known as *soft clamping*. The direct pathway was trained to produce the phonemes of each word from the semantics of each word, corrupted by gaussian noise with standard deviation 0.1. The input units were clamped in the normal way. Each pathway was trained to activate the phoneme units to within 0.2 of their correct values for a given input. After very extensive training they accomplished this in general, but the amount of noise added to their inputs made it impossible to guarantee this performance on any given trial. For this reason, training was halted when each pathway met the stopping criteria over ten successive sweeps through the training set.

The separately trained clean-up and direct pathways were then combined into a single, complete output network. This is straightforward because the two pathways have non-overlapping sets of connections, except for the biases of the phoneme units. For these, the biases from the clean-up pathway were used. The network was then given additional training on noisy input, during which only the weights in the direct pathway were allowed to change. In this way the direct pathway adjusted its mapping to more effectively use the fixed phonological clean-up in generating correct word pronunciations.

Finally, the output network was attached to the replication of the H&S input network and given

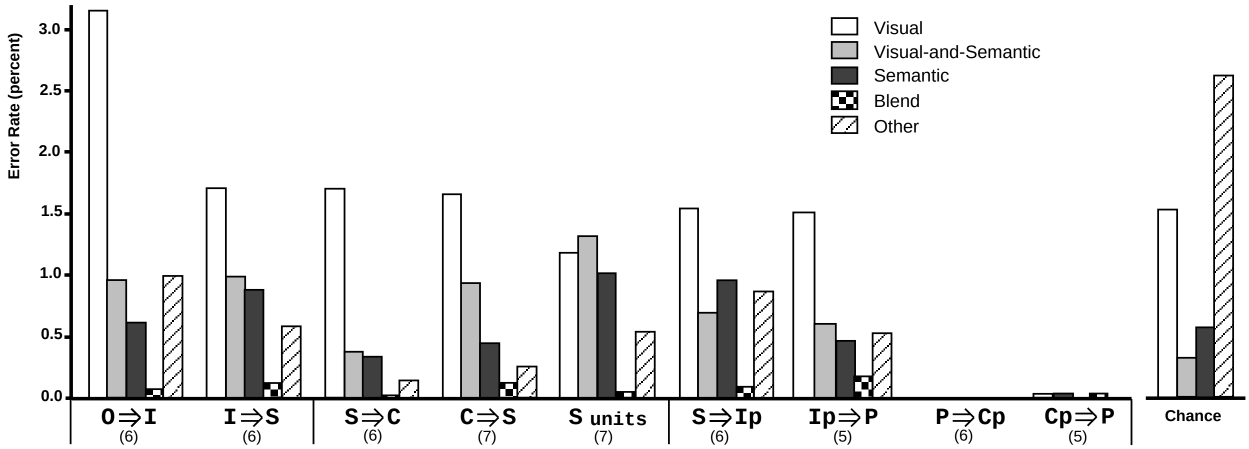

Figure 5: Error distributions for the extended back-propagation network.

a final tuning to ensure that the output network operated appropriately when its input was generated over time by an actual input network, rather than being clamped. The weights of the input network were not allowed to change, so that they continued to derive the correct semantics for each word. After this final training, which took 42 additional training sweeps, the extended network correctly derived the semantics and phonology of each word from its orthography.

Using the same random number generator seeds, the input portion of the extended network was subjected to the identical lesions as were applied to the original network. Additional lesions were applied to the semantic units themselves, and to each set of connections in the output network. For each lesion, correct, omission, and error responses were accumulated, and errors were classified according to their visual and semantic similarity to the stimulus. Figure 5 shows the distribution of error rates for all lesions of the extended network. Comparing with the results for the first extended network (see Figure 4), lesions to the input network still produce distributions of visual, semantic, and mixed visual-and-semantic errors, as well as *other* (unrelated) errors, but the rates of blend errors have been dramatically reduced by the training strategy. Notice that one result of the stronger phonological attractors for word pronunciations is that the relative rates of *other* errors have increased. When a lesion results in initial phonological activity that is highly corrupted, the new output system may still succeed in cleaning it up into a familiar response, even in cases where it bears no relation to the correct response.

Interestingly, a number of the *other* errors are actually of the visual-then-semantic type found in deep dyslexia (e.g. BOG ⇒ (dog) ⇒ "rat"). This type of error occurs when a lesion results in a semantic representation close to that of a word visually related to the stimulus, which is then mapped by the output system onto the phonology of a semantic neighbor of this visually related word. Thus, it is the *normal* operation of the output system that produces the semantic part of the visual-then-semantic error.

Lesions to the direct pathway of the output network (`S⇒Ip` and `Ip⇒P`) produce error patterns much like input lesions, although there is a slightly greater bias towards semantic errors relative to visual errors. However, most striking is the extremely low error rate for lesions within the phonological clean-up pathway (`P⇒Cp` and `Cp⇒P`). Although many words can still be read correctly with impaired clean-up—average correct performance after these lesions is 50.3%—it is very rare that phonology will be cleaned up into the pronunciation of another word. This result provides direct support for H&S's claim that attractors are critical for producing error responses.

One issue is whether the pattern of errors could have arisen by chance—that is, if error responses were related to stimuli only randomly. If the distribution of error types for a given lesion location occurred by chance, the ratios of their rates with the rate of *other* errors would approximate the corresponding ratios for the "Chance" error distribution. However, except for phonological clean-up lesions, the rates of visual, mixed visual-and-semantic, and semantic errors, relative to the rates of *other* errors, are greater for all lesion locations than predicted by chance. Specifically, the ratios with *other* error are larger than the chance value by at least a factor of 3.3 for visual errors, 11.7 for visual-and-semantic errors, and 2.9 for semantic errors. Thus, lesions anywhere along a pathway from orthography to phonology via semantics produce qualitatively similar patterns of errors. In this way, H&S's results appear to generalize to lesions all along a route from orthography to phonology via semantics.

3 The Network Architecture

The second design decision we will consider is the relevance of network architecture, by which we mean a specification of the number of units and their interconnectivity. H&S provide only a general justification for the network architecture they chose. Hidden units are needed because the problem of mapping orthography to semantics is not linearly separable. Recurrent connections are required to allow the network to develop semantic attractors, whose existence constitutes the major theoretical claim of the work. The choices of numbers of intermediate and clean-up units, restrictions on connections among sememe units, and connectivity density were an attempt to give the network sufficient flexibility to solve the task and build strong semantic attractors, while keeping the size of the network manageable. Some aspects of the design, particularly the selective use of intra-sememe connections, were rather inelegant and ad hoc.

Accordingly, we carried out a systematic comparison of the effects of damage in a range of network architectures designed to allow comparisons between basic aspects of the H&S network (see Figure 6). Versions of each of these networks were subjected to full range of lesion locations and severities, and evaluated both using the response criteria and using an output system. The results demonstrate that the qualitative error pattern after damage is surprisingly insensitive to architectural details, as long as attractors continue to operate downstream from the lesion (see [PS93] for details). When lesions are at or beyond the level at which attractors operate, the network produces very few explicit error responses, even though correct performance may be reasonable. In this way, the results in these conditions mirror those shown for lesions of the phonological clean-up pathway just described (see Figure 5). More critically for the present purposes, however, is that the similarity of error patterns produced by such a wide variety of architectures makes it highly unlikely that the basic results depend on any idiosyncratic characteristics of the H&S network.

4 The Training Procedure

Although back-propagation is quite a powerful training procedure, it uses information in ways that seem neurophysiologically implausible—a straightforward implementation of the procedure would require error signals to travel backward through synapses and axons [Cri89, Gro87]. As

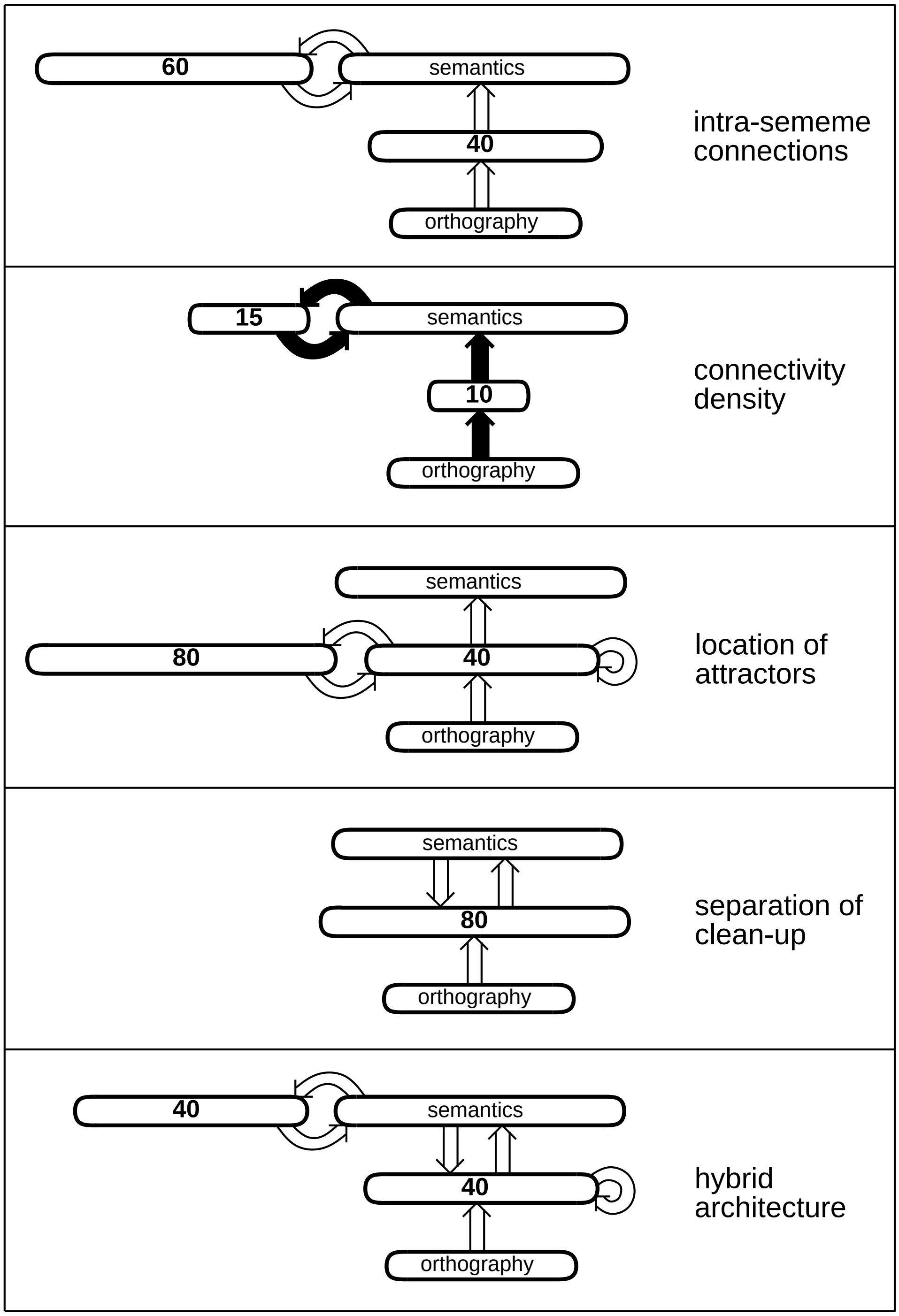

Figure 6: Five alternative network architectures for mapping orthography to semantics, and the issues they are designed to address.

such, it seems unlikely that back-propagation per se is what underlies human learning, and thus its use in modeling the *results* of human learning is somewhat suspect.

Proponents of the use of back-propagation in cognitive modeling have replied to this argument in two ways. The first is to demonstrate how the procedure might be implemented in a neurophysiologically plausible way. The more common reply, and the one adopted by H&S, is to argue that back-propagation is only one of a number of procedures for performing gradient descent learning in connectionist networks. As such, it is viewed merely as a programming technique for developing a network that performs a task, and is not intended to reflect any aspect of human learning per se. The implicit claim is that back-propagation develops representations that exhibit the same properties as would those developed by a more plausible procedure, but that it does so much more efficiently. However, this claim is rarely substantiated by a demonstration of the similarity between systems developed with alternative procedures.[6]

In this section, we replicate the main results obtained thus far with back-propagation, within the more plausible learning framework of contrastive Hebbian learning (CHL) in a deterministic Boltzmann Machine (DBM) [PA87, Hin89b]. In this framework, weights are changed in proportion to the difference in the product of unit states after settling with both inputs and outputs are clamped (the *positive* phase), and when settling after only the inputs are clamped (the *negative* phase). CHL is somewhat more biologically plausible than back-propagation because information about the correct states of output units is used in the same way as information about the input—that is, by propagating weighted unit activities, rather than passing error derivatives backward across connections. We also develop a closely-related stochastic GRAIN network [McC90, McC91] and compare it with the deterministic one.

4.1 Deterministic Boltzmann Machine

Figure 7 depicts the architecture of the DBM for mapping among the orthography, semantics, and phonology. All sets of connections are bidirectional and have full connectivity, except that no unit is connected to itself. In total, the network has 11,273 connections—about twice the number of connections in one of the back-propagation networks. This extra capacity is justified because CHL is not as efficient as back-propagation in using a small number of weights to solve a task.

In order to help the DBM learn the structure in the task (i.e., to reproduce the co-occurrences of unit states), the network was trained on three subtasks, each corresponding to a separate negative phase: (1) generate semantics and phonology from orthography, (2) generate orthography and phonology from semantics, and (3) generate semantics and orthography from phonology. Although only the first subtask is strictly required for reading via meaning, training on the other subtasks ensures that the network learns to model orthographic structure and its relationship to semantics in the same way as for phonological structure.[7] Also, learning the task in both directions should result in stronger and more robust attractors. The positive phase involved clamping the grapheme,

[6] Terry Sejnowski (personal communication) has successfully re-implemented NETtalk [SR87] as a stochastic Boltzmann Machine. However, he made no direct comparisons of the representations that the two procedures developed.

[7] Our use of a training procedure that involves learning to produce semantics from phonology in addition to producing phonology from semantics is in no way intended to imply a theoretical claim that input and output phonology are identical—it is solely a way of helping the network to learn the appropriate relationships between semantic and phonological representations.

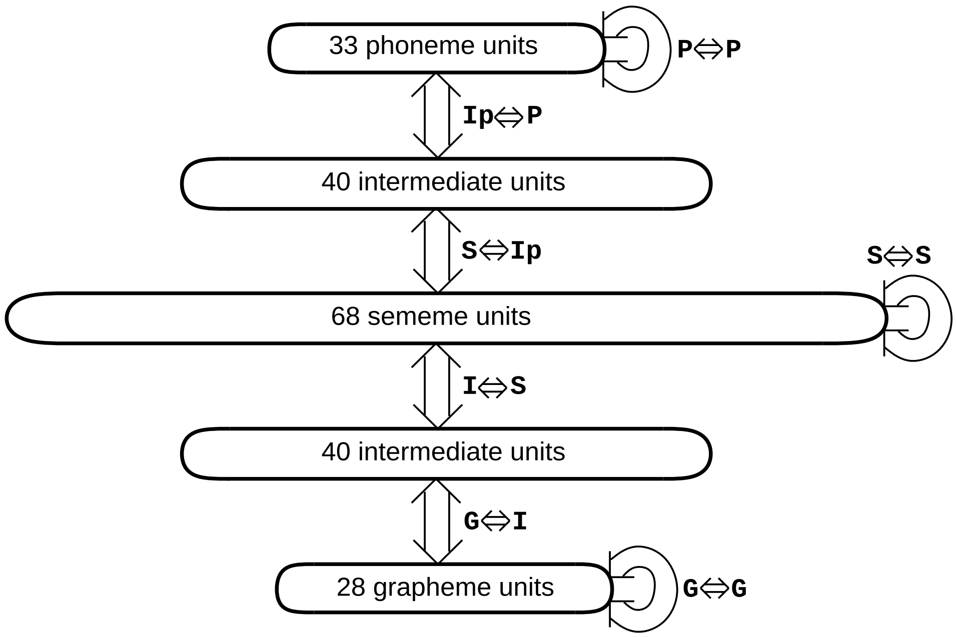

Figure 7: The DBM architecture for mapping among orthography, semantics, and phonology.

sememe, and phoneme units appropriately, and computing states for the two layers of intermediate units. In order to balance the three negative phases, the products of unit states in the positive phase are multiplied by three before being added into the pending weight changes. After slightly more than 2100 sweeps through the word set, the state of each grapheme, sememe, and phoneme unit was within 0.2 of its correct states during each of the three negative phases.

After training, each of the sets of connections in the DBM were subjected to 20 instances of lesions over the standard range of severity. We also subjected the semantic units to lesions of the same range of severity, in which the appropriate proportion of semantic units are selected at random and removed from the network. Since we are primarily concerned with the task of generating semantics and phonology from orthography, we only considered behavior in the negative phase in which the grapheme units are clamped. For each lesion, correct, omission, and error response were accumulated according to the same criteria as used for the back-propagation networks.

An interesting characteristic of the DBM is that it tends to settle into unit states that are very close to ± 1, even under damage. This results in very clean phonological output when it responds. Only 9.2% of omissions fail because of the criterion of a minimum slot response probability of 0.6 for responses. Thus, the phonological output criterion could be eliminated entirely without substantially altering the results with the DBM.

Figure 8 presents the distribution of error types for each lesion location of the DBM. Comparing with results for input lesions to the back-propagation network (shown in Figure 5), the DBM is producing about 4–8 times higher error rates. However, the distribution of error types is quite similar for the two networks. Both show a high proportion of visual errors for lesions to input pathways. Furthermore, like the back-propagation network, the DBM shows very low rates of blend responses. This is interesting because, unlike in the development of the back-propagation output network, no special effort was made to prevent blends in the design or training of the DBM. Their absence appears to be a natural and encouraging consequence of the nature of the attractors

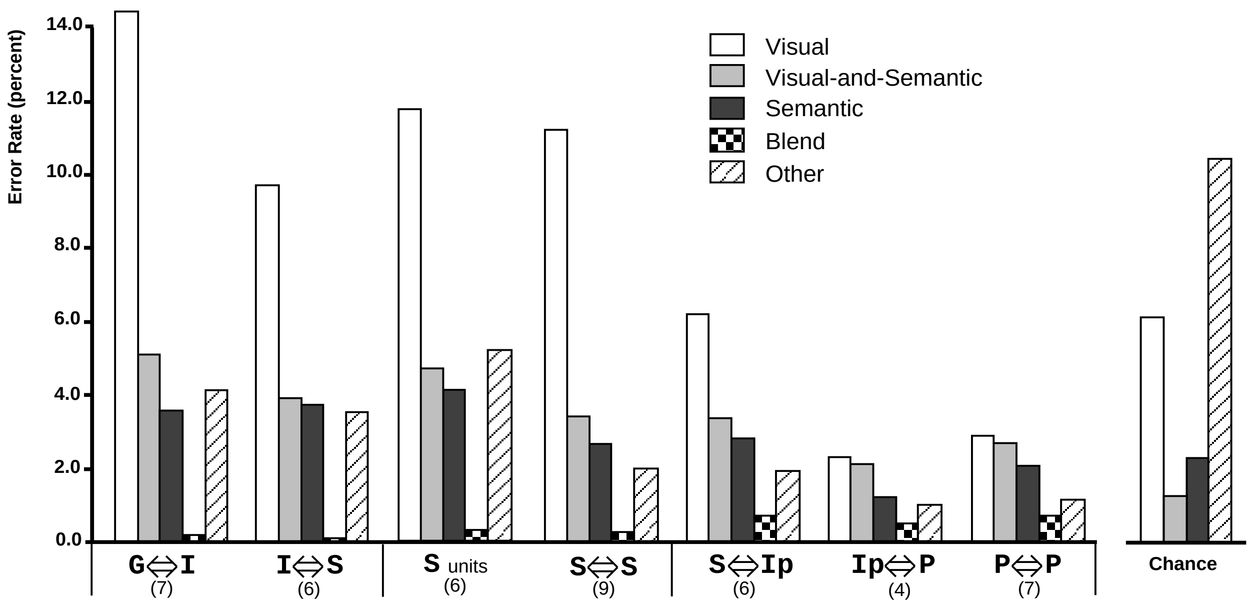

Figure 8: Error rates produced by lesions to each main set of connections, as well as to the semantic units, in the DBM. Results are averaged over severities that resulted in correct performance between 20–80%

developed by the DBM.

The error pattern for central lesions (`S⇔S` and `S` units) is quite similar to the pattern for input lesions. Lesioning the semantic units produces a higher overall error rate (25.6%) than lesioning the connections among them (19.6%), but the largest increase is among *other* errors. Also, in the DBM these lesions don't produce the same strong bias towards semantic similarity in errors as they do in the back-propagation network.

The pattern of error rates for output lesions to the DBM is quite different from that for the back-propagation network. The error rates for lesions to the direct pathway of the DBM (`S⇔Ip` and `Ip⇔P`) are lower than for input lesions, and less biased towards visual errors. In addition, the DBM produces far fewer *other* errors than the back-propagation network. Perhaps more striking, phonological clean-up lesions in the DBM (`P⇔P`) still produce significant error rates, fairly evenly distributed across type, while the analogous lesions in the back-propagation network (`P⇒Cp` and `Cp⇒P`) produce virtually no error responses. With phonological clean-up damage, the DBM can use the bidirectional interactions with the intermediate units as a residual source of clean-up.

All lesion locations in the DBM show a mixture of error types, and their ratios with the *other* error rates are higher than for randomly chosen error responses. Thus, the DBM replicates the main H&S results.

4.2 GRAIN Network

The effectiveness of noise in facilitating the development of strong attractors in the back-propagation output network suggests that it might have further benefits within the DBM framework. McClelland [McC90, McC91] has recently developed a stochastic elaboration of DBMs, called GRAIN networks (for Gradual Random Adaptive Interactive Nonlinear), that use real-valued stochastic

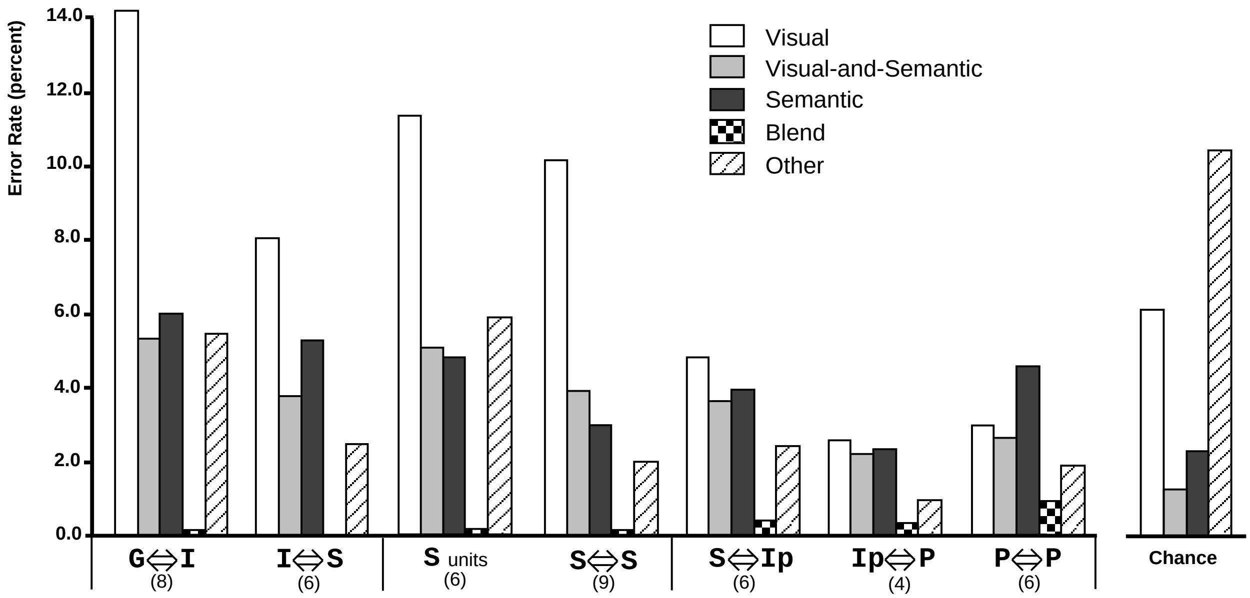

Figure 9: Error rates produced by lesions to each main set of connections in the GRAIN network.

units.[8] Although the principles of GRAIN networks can be embodied in a wide range of specific network formalisms, the type of GRAIN network we will investigate is identical to a DBM except that normally distributed noise ($\mu = 0.0, \sigma = 0.1$) is added to the input of each unit at each time step. The influence of noise is more widespread in a GRAIN network than in the back-propagation networks, because noise is applied to every unit in the network throughout settling.

A GRAIN network with the same architecture as the DBM was trained on the same task using CHL. Because the units in a GRAIN network are stochastic, the units never completely reach a fixedpoint in state space, but randomly fluctuate around it. However, if the amount of noise is small relative to the weights, the network will rarely jump out of a minimum as a result of the noise alone. In this case, all of the variation in unit states is caused by independent noise with zero mean, and so the expected value of the product of two unit states is the product of the states the units would have without noise.[9] For this reason, the final unit states at the end of settling are computed without noise before being used in the weight update rule. After 3500 sweeps through the training set, the GRAIN network could reliably generate any two of the orthography, semantics, or phonology of a word when given the third.

The GRAIN network was subjected to the same set of lesions as the DBM, and correct, omission, and error responses were accumulated. The input to units remained noisy during the gathering of data on impaired performance. Figure 9 presents the distribution of error types for each lesion location of the GRAIN network. The pattern of errors is quite similar to that of the DBM. The major difference is that the GRAIN network has significantly higher rates of semantic errors than

[8]Actually, GRAIN networks were developed as an elaboration of the Interactive Activation and Competition framework [MR81, RM82] in response to the need for intrinsic variability, as reflected by empirical limitations of the original model [Mas88]. However, the processing dynamics in a DBM are a special case of those in the IAC framework.

[9]Fluctuations in the states of two connected units due to noise will tend to be slightly correlated due to the weight between them, so that the product of their states without noise only approximates the expected value of their product with noise.

the DBM for almost all lesion locations. This makes sense in the following way. The amount of variation in input due to noise that a unit experiences increases as a function of its number of connections. Consider the input a unit j receives along a connection from unit i. Because the input to unit i has noise with zero mean added to it, its input to j can be thought of as a random variable with mean equal to what $s_i w_{ij}$ would be without noise (call it $s_i' w_{ij}$) and some variance dependent on the amount of noise. The summed input to j (before noise is added) is thus the sum of samples of a set of random variables. This sum is also a random variable, with mean equal to the sum of the means of the variables (i.e., $\sum_i s_i' w_{ij}$), and variance equal to the sum of their variances. Thus the mean of the summed input to a unit correctly approximates the true mean in a noiseless network, but the variance increases linearly with its number of connections. In the GRAIN network, semantic units have far more connections (149) than intermediate units (102) or phonological units (74), and so they are more drastically affected by the intrinsic noise in the states of other units. They must interact more effectively to compensate for this variability, resulting in stronger attractors at this level, and thus more semantic errors under damage.

Nonetheless, it is surprising that the GRAIN network and the DBM are so similar in the nature of the attractors they develop, as reflected in their behavior under damage. One explanation may come from the behavior of the DBM during learning. The mathematical justification for the learning procedure [Hin89b] assumes that only rarely will small changes to the weights cause the network to settle into a different minimum. However, in practice this appears to be more the rule than the exception. As the weights slowly change, the network samples among a large number of minima during the negative phase(s), raising their energy to the degree to which they differ from the minima of each corresponding positive phase. As the network improves on the task, fewer and fewer of these minima remain sufficiently good for the network to settle into them. Eventually the network consistently reaches the single minimum that is most similar to the positive phase minimum, and reduces the difference until the training criteria are met. This type of variability *over weight changes* in settling to minima appears to have similar effects as the variability of unit states during a single settling in a GRAIN network. Both processes force the network to explore, and hence shape appropriately, a much larger amount of the energy surface in state space than will ultimately be traversed when the network has learned. Hence, one possible explanation for why the GRAIN network is no more robust to damage than the DBM is that in both networks the attractors have been strengthened by pressure from variability, albeit from different sources.

Both the DBM and GRAIN network serve to validate the claim that the nature of the attractors developed using back-propagation have properties that are similar to those developed using these alternative, more biologically plausible formalisms.

5 The Task Domain

The final aspect of the H&S model that we investigate is the definition of the task of reading via meaning. A rather severe limitation of the H&S model is that it was trained on only 40 words, allowing only a very coarse approximation to the range of visual and semantic similarity among words in a patient's vocabulary. More critically, a distinction among words known to have a significant effect on reading in deep dyslexia—concreteness or imageability— could not be addressed using the original H&S word set because it contains only concrete nouns. In this section, we summarize our work in extending the H&S approach to account for effects of concreteness and their interactions with visual errors (see [PS91, PS93] for details).

To examine the effect of concreteness on visual errors, a set of 20 concrete and 20 abstract words were chosen such that each pair of words differed by a single letter (e.g., ROPE, ROLE). Following Jones [Jon85], Gentner [Gen81], and others, we develop a semantic representation in which concrete words have "richer" representations, in terms of number of active features, than do abstract words. Specifically, out of 98 possible semantic features, concrete words have an average of 18.2 features, while abstract words have an average of only 4.7 features. A back-propagation network was trained to map orthography to phonology via these representations, in the same manner as for the back-propagation simulations described in Section 2.

Because abstract words have far fewer features, they are less able to engage the semantic clean-up mechanism effectively, and must rely more heavily on the direct pathway where visual influences are strongest. As a result, lesions to the direct pathway of the input network reproduce the effects of concreteness and their interaction with visual errors found in deep dyslexia: better correct performance for concrete over abstract words, a tendency for error responses to be more concrete than stimuli, and a higher proportion of visual errors in response to abstract compared with concrete words.

Surprisingly, severe lesions to the clean-up pathway produce the *opposite* effect, with abstract words now being read better than concrete words, and concrete words producing more visual errors than do the abstract words. This reversal arises because, under this type of lesion, the processing of most concrete words is impaired but many abstract words can be read solely by the direct pathway.

In fact, there is a single known exception to the advantage for concrete words shown by deep dyslexic patients: patient CAV with *concrete word dyslexia* [War81]. CAV failed to read concrete words like MILK and TREE but succeeded at highly abstract words such as APPLAUSE, EVIDENCE, and INFERIOR. Overall, abstract words were more likely to be correctly read than concrete (55% vs. 36%). In complementary fashion, 63% of his visual error responses were more abstract than the stimulus. Furthermore, the hypothesis of severe clean-up damage is consistent with other aspects of his performance. His reading disorder was quite severe initially, and he also showed an advantage for abstract words in picture-word matching with auditory presentation, suggesting modality-independent damage at the level of the semantic system.

Overall, the network successfully extends the H&S approach to account for the effects of concreteness in deep dyslexia, and also offers the possibility of explaining the single, enigmatic case of concrete word dyslexia. Thus, together with extrapolations based on previous theorizing (e.g., [Fun87]), the connectionist approach offers a comprehensive, principled account of the full range of symptoms found in deep dyslexia.

6 Conclusions

Hinton and Shallice [HS91] offer a connectionist account in which the central aspects of deep dyslexia—the existence of semantic errors and their co-occurrence with visual and mixed visual-and-semantic errors—arise naturally as a result of damage to a network that builds attractors in mapping orthography to semantics. While the approach has the advantage over traditional models of being far more computationally explicit, it has the limitation that there is little understanding of the underlying principles of the model which give rise to its behavior under damage. The current research involves a set of connectionist simulation experiments aimed both at developing our understanding of these principles, and at extending the empirical adequacy of the approach on the basis of this understanding. The results demonstrate the usefulness of a connectionist approach to understanding deep dyslexia in particular, and the viability of connectionist neuropsychology in general.

Furthermore, studying the breakdown of behavior in damaged networks sheds light on their normal computational characteristics. Implementing an output system that successfully pronounces a set of words from their semantics was relatively straightforward—the limitations of the system became apparent only under damage. The tendency for distributed output representations to lead to blends under damage clarifies the need for stronger attractors that encode constraints at the appropriate level of structure in the output. The fact that contrastive Hebbian learning in a deterministic Boltzmann Machine and in a GRAIN network produces such attractors naturally, perhaps as a result of variability over weight changes, is a significant advantage of that framework.

Connectionist networks would appear *a priori* to be an appropriate formalism within which to develop computational models of neuropsychological disorders. Although the specific relationship between these networks and neurobiology is far from clear [SKC89, Smo88], the belief that representation and computation in these networks resembles neural computation at some level remains one of their strongest attractions. As the present research illustrates, the fact that the behavior of connectionist networks after damage resembles that of neurological patients supports the claim that the apparent similarity is, in fact, substantial.

References

[BTMS90] Derek Besner, Leslie Twilley, Robert S. McCann, and Ken Seergobin. On the connection between connectionism and data: Are a few words necessary? *Psychological Review*, 97(3):432–446, 1990.

[Col85] Max Coltheart. Cognitive neuropsychology and the study of reading. In Michael I. Posner and O. S. M. Marin, editors, *Attention and Performance XI*, pages 3–37. Lawrence Erlbaum Associates, Hillsdale, NJ, 1985.

[CPM80] Max Coltheart, Karalyn E. Patterson, and John C. Marshall, editors. *Deep Dyslexia*. Routledge & Kegan Paul, London, 1980.

[Cri89] Francis H. C. Crick. The recent excitement about neural networks. *Nature*, 337:129–132, 1989.

[EY88] Andrew W. Ellis and Andrew W. Young. *Human Cognitive Neuropsychology*. Lawrence Erlbaum Associates, Hillsdale, NJ, 1988.

[FM91] Martha J. Farah and James L. McClelland. A computational model of semantic memory impairment: Modality-specificity and emergent category-specificity. *Journal of Experimental Psychology: General*, 120(4):339–357, 1991.

[Fun87] Elaine Funnell. Morphological errors in acquired dyslexia: A case of mistaken identity. *Quarterly Journal of Experimental Psychology*, 39A:497–539, 1987.

[Gen81] Deidra Gentner. Some interesting differences between verbs and nouns. *Cognition and Brain Theory*, 4(2):161–178, 1981.

[Gro87] Stephen Grossberg. From interactive activation to adaptive resonance. *Cognitive Science*, 11:23–63, 1987.

[HA81] Geoffrey E. Hinton and James A. Anderson, editors. *Parallel Models of Associative Memory*. Lawrence Erlbaum Associates, Hillsdale, NJ, 1981.

[Hin89a] Geoffrey E. Hinton. Connectionist learning procedures. *Artificial Intelligence*, 40:185–234, 1989.

[Hin89b] Geoffrey E. Hinton. Deterministic Boltzmann learning performs steepest descent in weight-space. *Neural Computation*, 1(1):143–150, 1989.

[Hop82] John J. Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Science, U.S.A.*, 79:2554–2558, 1982.

[HS86] Geoffrey E. Hinton and Terrance J. Sejnowski. Learning and relearning in Boltzmann Machines. In David E. Rumelhart, James L. McClelland, and the PDP research group, editors, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Volume 1: Foundations*, pages 282–317. MIT Press, Cambridge, MA, 1986.

[HS91] Geoffrey E. Hinton and Tim Shallice. Lesioning an attractor network: Investigations of acquired dyslexia. *Psychological Review*, 98(1):74–95, 1991.

[Jon85] Gregory V. Jones. Deep dyslexia, imageability, and ease of predication. *Brain and Language*, 24:1–19, 1985.

[LB88] Joel Lachter and Thomas G. Bever. The relation between linguistic structure and theories of language learning: A constructive critique of some connectionist learning models. *Cognition*, 28:195–247, 1988.

[Mas88] Dominic W. Massaro. Some criticisms of connectionist models of human performance. *Journal of Memory and Language*, 27:213–234, 1988.

[McC90] James L. McClelland. The GRAIN model: A framework for modeling the dynamics of information processing. In D. E. Meyer and S. Kornblum, editors, *Attention and Performance XIV*. Lawrence Erlbaum Associates, Hillsdale, NJ, 1990.

[McC91] James L. McClelland. Stochastic interactive processes and the effect of context on perception. *Cognitive Psychology*, 23:1–44, 1991.

[MP80] John Morton and Karalyn Patterson. A new attempt at an interpretation, Or, an attempt at a new interpretation. In Max Coltheart, Karalyn E. Patterson, and John C. Marshall, editors, *Deep Dyslexia*, pages 91–118. Routledge & Kegan Paul, London, 1980.

[MR81] James L. McClelland and David E. Rumelhart. An interactive activation model of context effects in letter perception: Part 1. An account of basic findings. *Psychological Review*, 88(5):375–407, 1981.

[NM91] Leigh E. Nystrom and James L. McClelland. Blend errors during cued recall. In *Proceedings of the 13th Annual Conference of the Cognitive Science Society*, pages 185–190. Lawrence Erlbaum Associates, Hillsdale, NJ, 1991.

[PA87] Carsten Peterson and James R. Anderson. A mean field theory learning algorithm for neural nets. *Complex Systems*, 1:995–1019, 1987.

[PH87] David C. Plaut and Geoffrey E. Hinton. Learning sets of filters using back propagation. *Computer Speech and Language*, 2:35–61, 1987.

[PP88] Steven Pinker and Alan Prince. On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition*, 28:73–193, 1988.

[PS91] David C. Plaut and Tim Shallice. Effects of abstractness in a connectionist model of deep dyslexia. In *Proceedings of the 13th Annual Conference of the Cognitive Science Society*, pages 73–78. Lawrence Erlbaum Associates, Hillsdale, NJ, 1991.

[PS93] David C. Plaut and Tim Shallice. Deep dyslexia: A case study of connectionist neuropsychology. *Cognitive Neuropsychology*, 1993. In press.

[PSM90] Karalyn E. Patterson, Mark S. Seidenberg, and James L. McClelland. Connections and disconnections: Acquired dyslexia in a computational model of reading processes. In R. G. M. Morris, editor, *Parallel Distributed Processing: Implications for Psychology and Neuroscience*. Oxford University Press, London, 1990.

[RHW86] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning representations by back-propagating errors. *Nature*, 323(9):533–536, 1986.

[RM82] David E. Rumelhart and James L. McClelland. An interactive activation model of context effects in letter perception: Part 2. the contextual enhancement effect and some tests and extensions of the model. *Psychological Review*, 89:60–94, 1982.

[RM86] David E. Rumelhart and James L. McClelland. On learning the past tenses of English verbs. In James L. McClelland, David E. Rumelhart, and the PDP research group, editors, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Volume 2: Psychological and Biological Models*, pages 216–271. MIT Press, Cambridge, MA, 1986.

[SKC89] Terrence J. Sejnowski, Cristof Koch, and Patricia S. Churchland. Computational neuroscience. *Science*, 1989.

[SM89] Mark S. Seidenberg and James L. McClelland. A distributed, developmental model of word recognition and naming. *Psychological Review*, 96:523–568, 1989.

[SM90] Mark S. Seidenberg and James L. McClelland. More words but still no lexicon: Reply to Besner et al. (1990). *Psychological Review*, 97(3):477–452, 1990.

[Smo88] Paul Smolensky. On the proper treatment of connectionism. *Behavioral and Brain Sciences*, 11:1–74, 1988.

[SR87] Terrance J. Sejnowski and Charles R. Rosenberg. Parallel networks that learn to pronounce English text. *Complex Systems*, 1:145–168, 1987.

[War81] Elizabeth K. Warrington. Concrete word dyslexia. *British Journal of Psychology*, 72:175–196, 1981.

[WP90] Ronald J. Williams and Jing Peng. An efficient gradient-based algorithm for on-line training of recurrent network trajectories. *Neural Computation*, 2(4):490–501, 1990.