Chapter 3

Response generation: Mapping semantics to phonology

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 dyslexics. 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. Patterson et al., 1990; Sejnowski & Rosenberg, 1987).

However, satisfying the criteria only coarsely approximates the requirements of an actual output system. In particular, while it may be reasonable that semantics which failed the criteria could not drive a response system, no evidence was given that semantics which satisfied the criteria could succeed in generating a response. Also, the criteria are insensitive to the relative semantic and phonological discriminability of words and so may be inadvertently biased towards producing certain effects. In addition, by not implementing an output system H&S can consider only the "input" and "central" forms of deep dyslexia (Shallice & Warrington, 1980) and must assume that the specific nature of the output system plays no role in these patients' reading errors. Finally, a best-match procedure is rather powerful and knowledge-intensive. At a general level, if too much of the difficulty of a problem is pushed off into the assumed mechanisms for generating the input or interpreting the output, the role of the network itself becomes less interesting (Lachter & Bever, 1988; Pinker & Prince, 1988). This is especially ironic as a best-match (categorization) process is exactly the sort of operation at which connectionist networks are supposed to excel (Hinton &

Anderson, 1981; Hopfield, 1982).

For all of the above reasons, 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. 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. Nystrom & McClelland, 1991; Rumelhart & McClelland, 1986; Seidenberg & McClelland, 1989). In the domain of acquired dyslexia, the problem is that the damaged network produces responses which are inappropriate "blends" of known responses. In this chapter, we illustrate this problem and demonstrate a method for overcoming it, allowing us to develop networks that map from orthography to phonology via semantics which produce very few blends under damage. The effects of lesions to the "input" portion of these network that map from orthography to semantics are compared with those using the response criteria to provide a *post hoc* evaluation of the generality of the H&S results. Finally, we subject these networks to lesions of the "output" portions that map from semantics to phonology, and compare the resulting behavior with that produced by earlier damage.

3.1 Phonological blends

The problems that occur in realizing 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 that 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.

3.1.1 The task

The input to the network consists of the 40 semantic representations that served as output in the H&S model (described in Figure 2.6, p. 35). A phonological output representation was defined in terms of 33 position-specific *phoneme* units (see Table 3.1). 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.

3.1.2 The network

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

Phonemes allowed in each position		
b d dy g h j k l m n p r t	a ar aw e ew i ie o oa ow u	b d g k n m p t -

Phonological representation of each word									
Indoor Objects		Animals		Body Parts		Foods		Outdoor Objects	
BED	/b e d/	BUG	/b u g/	BACK	/b a k/	BUN	/b u n/	BOG	/b o g/
CAN	/k a n/	CAT	/k a t/	BONE	/b oa n/	HAM	/h a m/	DEW	/dy ew -/
COT	/k o t/	COW	/k ow -/	GUT	/g u t/	HOCK	/h o k/	DUNE	/dy ew n/
CUP	/k u p/	DOG	/d o g/	HIP	/h i p/	LIME	/l ie m/	LOG	/l o g/
GEM	/j e m/	HAWK	/h aw k/	LEG	/l e g/	NUT	/n u t/	MUD	/m u d/
MAT	/m a t/	PIG	/p i g/	LIP	/l i p/	POP	/p o p/	PARK	/p ar k/
MUG	/m u g/	RAM	/r a m/	PORE	/p aw -/	PORK	/p aw k/	ROCK	/r o k/
PAN	/p a n/	RAT	/r a t/	RIB	/r i b/	RUM	/r u m/	TOR	/t aw -/

Table 3.1: A phonological representation for words in terms of 33 position-specific phoneme units. The letter(s) used to represent phonemes are not from a standard phonemic alphabet but rather are intended to have more intuitive pronunciations. Also note that the definitions are based on British rather than American pronunciations (e.g. HAWK and PORK rhyme).

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. 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.

3.1.3 The training procedure

The output network was trained in exactly the same manner as the H&S network (described in Section 2.6.3) with one difference. The network was run for eight iterations instead of seven to allow information about the input to cycle through the phonological clean-up loop and influence the phoneme units an extra time. 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.

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 3.1, 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 the sixth iteration) 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

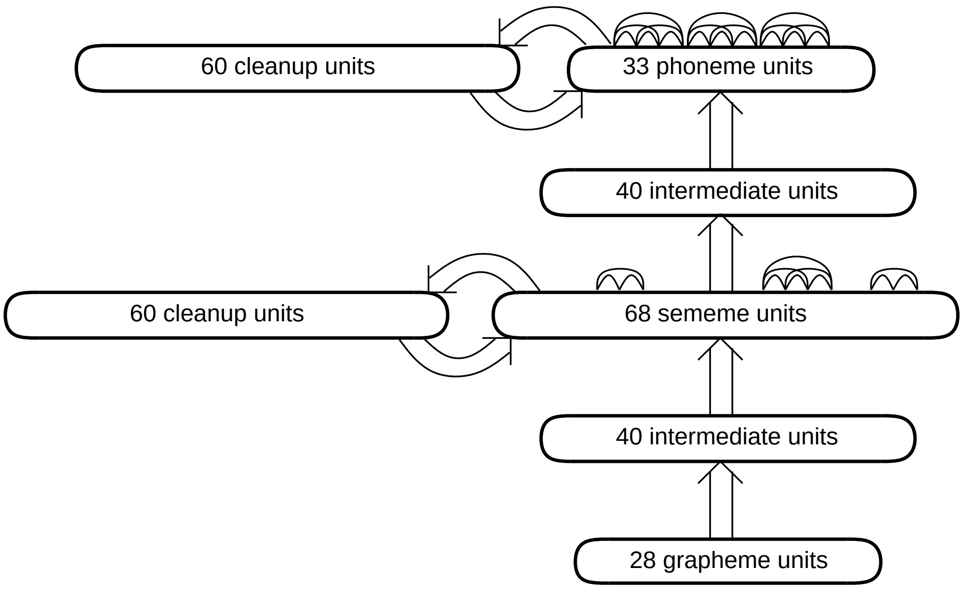


Figure 3.1: The architecture of a network that maps from orthography to phonology via semantics.

word. After an additional 34 sweeps through the training set, the combined network succeeded in producing the correct phonemes of each word given its graphemes as input.

3.1.4 The effects of lesions

After training, the complete network successfully derives the semantics and phonology of each word when presented with its orthography. In order to model the reading behavior of deep dyslexic patients, we simulate their neurological damage by removing a proportion of the connections between groups of units in the network. This damage impairs the ability of the network to derive the correct pronunciations of words. Consequently, we need some way of interpreting the corrupted output of the network as an explicit response. In addition, patients frequently produce no response to a word, or respond "I don't know." In order for the network to behave analogously, we also need a way of determining when the damaged network does *not* respond because the phonological output is ill-formed. 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*. The criterion we employ does not rely on any knowledge of the particular words the network has been trained on—it only considers the *form* of the output representation.

Given our phonological representation, a natural criterion is to require that one and only 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 0.0 or 1.0, we need a more precise definition of "active" and "inactive." In addition, we would like the definition to generalize to other types of binary output representations. Accordingly, we use the following procedure to determine if and how the network responds. For each phoneme position, interpreting the outputs of units as independent probabilities defines a probability distribution over possible binary output vectors for that position. If, for every position, the most probable output vector has exactly one active phoneme and probability greater than 0.5, the phonemes they each represent are produced as the response. More formally, if y_i is the output of phoneme unit i and

$$b_i = \begin{cases} 0 & \text{if } y_i < 0.5 \\ 1 & \text{otherwise} \end{cases}$$

is its output converted to binary, then the network produces a response if for every position p,

$$\prod_{i \in p} (1 - |y_i - b_i|) > 0.5$$

and exactly one $b_i = 1$. The response produced is the concatenation of the phonemes represented by each i for which $b_i = 1$. If the criterion is not satisfied for any position, the output activity produced by the network is considered ill-formed and it fails to respond. This procedure is closely related to the maximum-likelihood interpretation of the cross-entropy error function used to train the network (Hinton, 1989a). Notice that there are a large number of legal responses other than those the network is trained to produce. This expressiveness is one of the strengths of using a distributed output representation but it is not without its problems, as we are about to see.

Each of the four main sets of connections in the input network was subjected to “lesions” by chosing 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. Error responses were categorized in terms of their relation to the input word. In addition to visual and semantic similarity (as defined by H&S and described in Section 2.6.4), words can also be phonologically similar—that is, have overlapping phonemes. Since visual and phonological similarity typically co-occur, we considered an error to be phonological only if it was more phonologically than visually similar (e.g. HAWK `/h aw k/` and PORK `/p aw k/` using British pronunciations). In addition, some potential errors are appropriately categorized as phonological-and-semantic under this definition (e.g. DEW `/dy ew -/` and DUNE `/dy ew n/`). It should be pointed out that errors categorized as visual or mixed visual-and-semantic may actually result from phonological rather than visual influences—the current word set does not contain enough words that dissociate visual and phonological similarity to investigate the relative contribution of these two influences. We will take up the issue of distinguishing the influences of visual and phonological similarity on errors in the General Discussion.

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. Non-blend errors were divided into visual, visual-and-semantic,

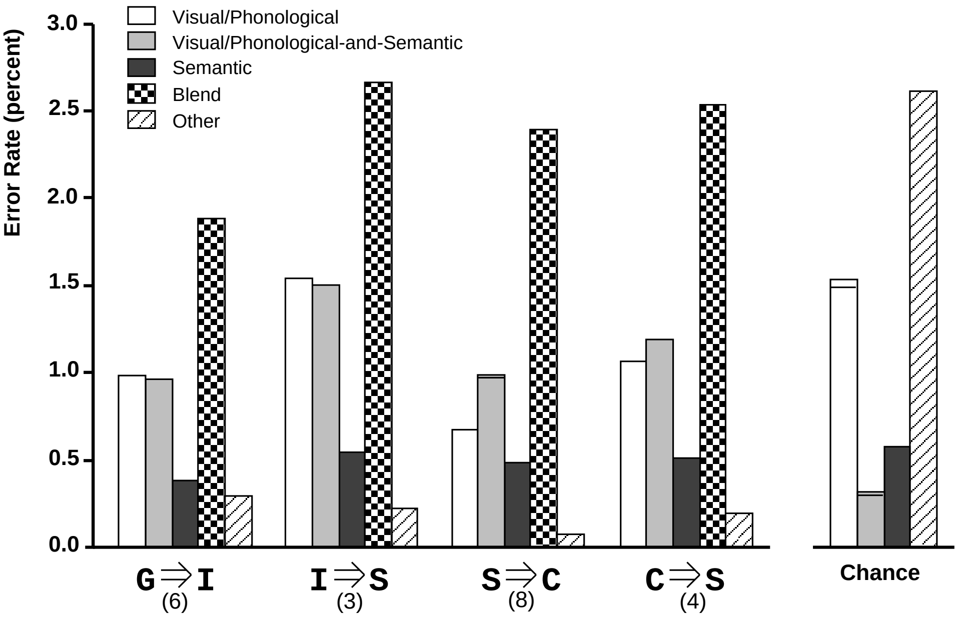


Figure 3.2: Error rates produced by lesions to each main set of connections in the input network. In this and following similar figures, phonological errors are shown as an extra bar over visual errors, and phonological-and-semantic errors are shown as an extra bar over visual-and-semantic errors. “Chance” 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.

semantic, phonological-and-semantic, phonological, and other errors. Figure 3.2 presents the average rates of each of these error types for each lesion location. The first thing to notice is that the rates of visual, mixed visual-and-semantic, and semantic errors replicate the H&S results. However, 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 to words (see Appendix 2 of Coltheart et al., 1987a). Table 3.2 presents some typical examples of blend errors produced by the network under a variety of lesions. The semantic activity produced by each input is characterized by its proximity (as defined in Section 2.6.4) 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. RIB (+HIP) ⇒ /r i p/), which is why we call these errors “blends.” Occasionally, new phonemes are introduced under the pressure of mixed semantics (e.g. ROCK (+TOR) ⇒ /r a k/). Interestingly, semantics that would easily satisfy H&S’s criteria for a correct response may still be sufficiently inaccurate for the output system to produce a blend (e.g. RAT (*prox* 0.98, *gap* 0.26) ⇒ /r a g/). On the other hand, semantics that are quite

		Nearest	Semantics				
Input	Response	Word	Best	*prox*	Next	*prox*	Lesion
RAT	/r a g/	RAT	RAT	0.98*	DOG	0.72	G⇒I(0.1)
ROCK	/r a k/	ROCK	ROCK	0.97*	TOR	0.83	G⇒I(0.2)
RUM	/h o p/	HOCK	RUM	0.84*	HOCK	0.77	G⇒I(0.5)
RIB	/r i p/	HIP	HIP	0.92*	RIB	0.83	I⇒S(0.2)
BOG	/b u k/	BUG	BOG	0.60	RAM	0.57	I⇒S(0.25)
LIME	/b aw g/	BOG	HAWK	0.63	RAT	0.59	I⇒S(0.4)
GUT	/b u t/	GUT	GUT	0.63	PORE	0.62	S⇒C(0.15)
BUN	/b o n/	BOG	BUN	0.87*	POP	0.75	S⇒C(0.3)
DUNE	/k u t/	CUP	CAT	0.61	PIG	0.60	S⇒C(0.5)
RAT	/b a g/	BOG	RAT	0.94*	BUG	0.75	C⇒S(0.1)
CAN	/k u n/	CAN	CAN	0.96*	MUG	0.80	C⇒S(0.15)
GEM	/b o m/	BOG	GEM	0.70	BUN	0.67	C⇒S(0.4)

Table 3.2: Examples of nonword "blend" errors produced by the network. "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 *p* to the semantic activity produced by the network. Semantics that satisfy the response criteria are marked with an asterisk.

far from any known word may still produce a response, albeit incorrect (e.g. BOG(*prox 0.63*) ⇒ /b u k/). Clearly the current output system behaves quite differently from what the H&S criteria assume about a response system.

In order to better understand blends, we compared correct, error, and blend responses in terms of the "goodness" of their phonological output, defined as the minimum, over phoneme positions, of the probability of the most likely output vector at that position (ignoring the 0.5 criterion for an explicit response used previously). As Figure 3.3 shows, correct, error, and blend responses differ significantly in the goodness of their phonological output (means, correct: 66.4, errors: 54.0, blends: 47.1, $t(9606) = 39.4$, $p < .001$ for correct vs. errors, $t(4025) = 16.6$, $p < .001$ for errors vs. blends). The figure reveals that increasing the minimum probability criterion to 0.6 would maximally discriminate between correct and blend responses while retaining a significant proportion of error responses. Figure 3.4 shows the distributions of the minimum probability for each error types, illustrating that increasing the criterion would increase the proportion of mixed visual-and-semantic errors but otherwise leave the relative distribution of error types essentially unchanged.[1] However, even with the higher response criterion a substantial number of blends still

[1] The lack of a difference between the goodness of visual and semantic errors suggests that, according to the model, the higher confidence that deep dyslexics have in their visual as compared with semantic errors (Patterson, 1978) does *not* arise from differences at the phonological output level. Chapter 5 investigates whether this effect can be accounted for by a "goodness" measure applied to other parts of the network.

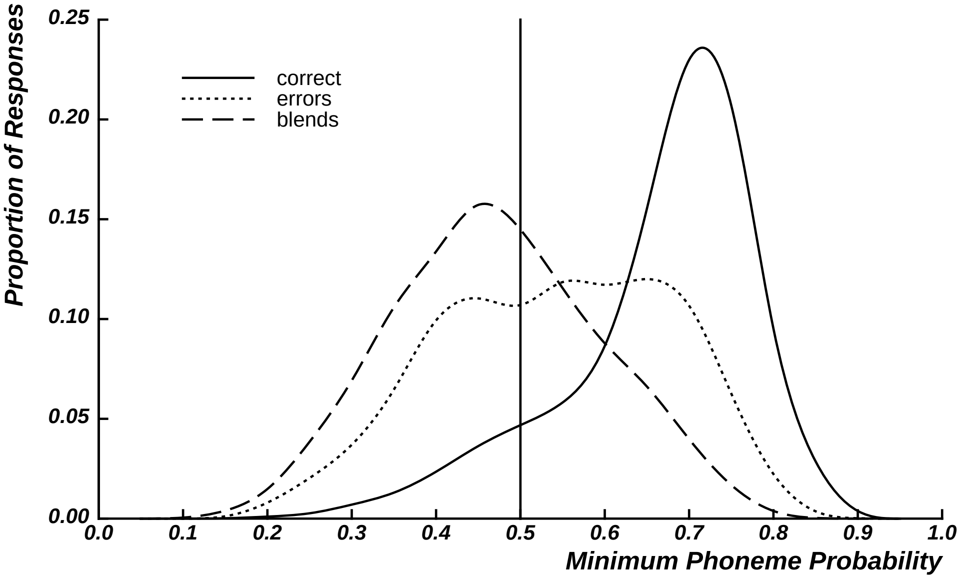


Figure 3.3: Distribution of the minimum output probability at any position for correct, error, and blend responses.

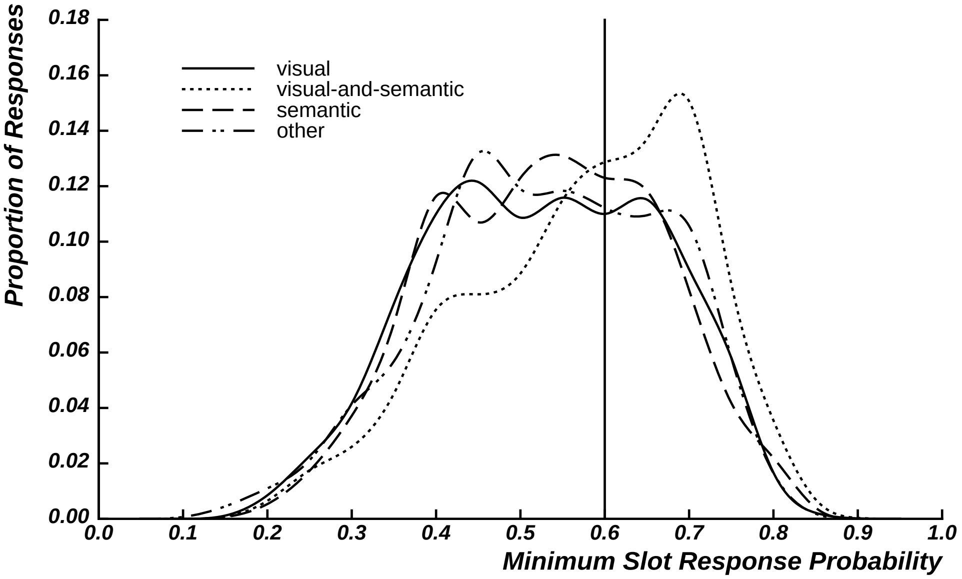


Figure 3.4: Distribution of the minimum output probability at any position for visual, mixed visual-and-semantic, semantic, and other errors.

occur. Indeed, no value for the response criterion would eliminate blends and leave a substantial number of correct responses.

3.1.5 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.[2] 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 (e.g. Hopfield, 1982) 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 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. RIB and HIP), 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. `/r i p/`), 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 & McClelland's (1989) model of word pronunciation, which maps from the orthography to the phonology of monosyllabic words. The model pronounces non-words 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 non-word, the model occasionally produces an inappropriate blend *at the level of phonemic features*. For example, when presented with the letter

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

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).[3] 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 (also see Rumelhart & McClelland, 1986; Sejnowski & Rosenberg, 1987).

We must emphasize that, while some neurological patients with more general phonological difficulties produce literal paraphasias in oral reading, the deep dyslexic patients whom the damaged model is intended to emulate do not, and hence their occurrence makes the current output system unacceptable. To some readers it may seem strange that inappropriate behavior under damage should make the *normal* network unsatisfactory. After all, the network succeeds at the task on which it was trained—pronouncing all 40 words. However, our concern is not just with accomplishing the task, but with the *way* that the network accomplishes the task—the nature of its representations and processes. Most connectionist research tests the adequacy of a network beyond its specific training by how well it generalizes to novel input. In a similar way, damage to a network has the effect of providing the remaining portions of the network with unfamiliar input. However, damage can affect internal representations in ways that cannot be directly mimicked by manipulations of the input to the network. For this reason, we suggest that the behavior of the network under damage provides a more general, and for some purposes more informative, indication of the nature of the representations and processes the network develops during training.

3.2 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.[4] In this way, initial phonological patterns that are a mixture of the phonemes of two words will

[3] In general, the model often produces non-word pronunciations that differ from what normal subjects would consider the correct pronunciation (Besner et al., 1990, but see Seidenberg & McClelland, 1990), suggesting that it has not sufficiently learned the appropriate regularities both between and within the phonemes of word pronunciations.

[4] 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. Chapter 5 describes simulations using an alternative learning procedure in which networks develop strong attractors naturally, so that no specific training techniques are required to eliminate phonological blends under damage.

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.

3.2.1 The network architecture

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. In addition, two versions of the phonological clean-up pathway will be developed, with and without interconnections among phoneme units at the same position. A comparison of these versions will allow us to evaluate the importance of direct connections in developing strong attractors. The pathway with intra-phoneme connections (IP) has a total of 1744 connections, while the other (noIP) has 1373 connections. The direct pathway from semantics to phonology still has 40 intermediate units and 25% connectivity, for a total of 1034 connections. These two pathways are depicted in Figure 3.5.

3.2.2 The training procedure

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 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

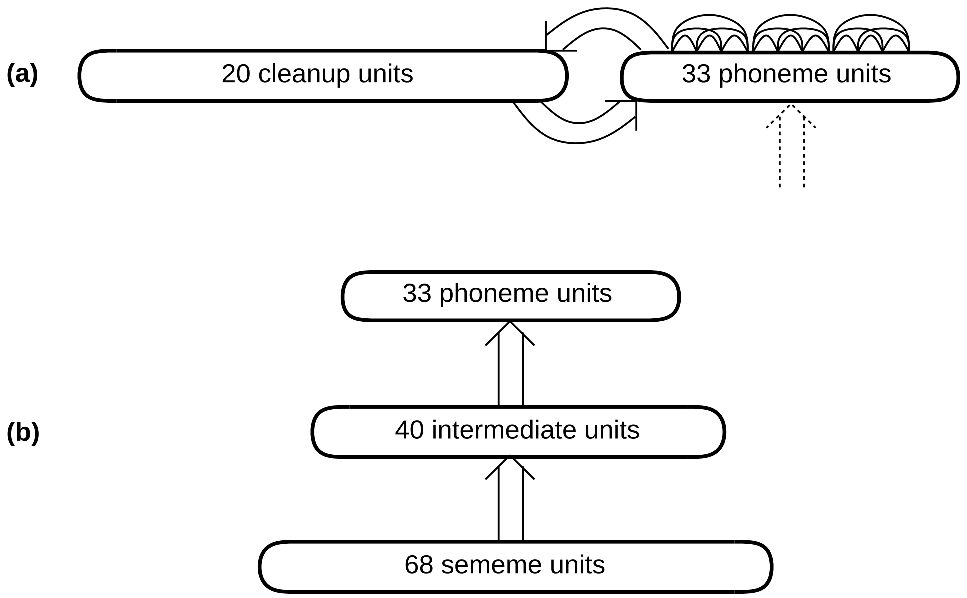


Figure 3.5: The architectures of the separately-trained parts of the output network: (a) the phonological clean-up pathway (with intra-phoneme connections), and (b) the direct semantics-to-phonology pathway.

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 explicit “weight decay” (Hinton, 1989a), causes the knowledge of the task to be more evenly distributed across all of the connections, making the network more uniformly robust to lesions (Farah & McClelland, 1991).

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 (Plaut & Hinton, 1987). 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.[5] 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. It should be mentioned that,

[5] A somewhat different use of incremental training is to enable separate parts of the network to independently specialize on *different* aspects of a task (Waibel, 1989). In fact, some recently developed connectionist learning procedures (Hampshire & Waibel, 1989; Jacobs et al., 1991; Nowlan, 1990) enable a modular network to automatically discover and carry out useful task decompositions, but the way that the outputs of separate modules can combine in such systems is typically restricted to selecting a single “expert” or a simple linear combination.

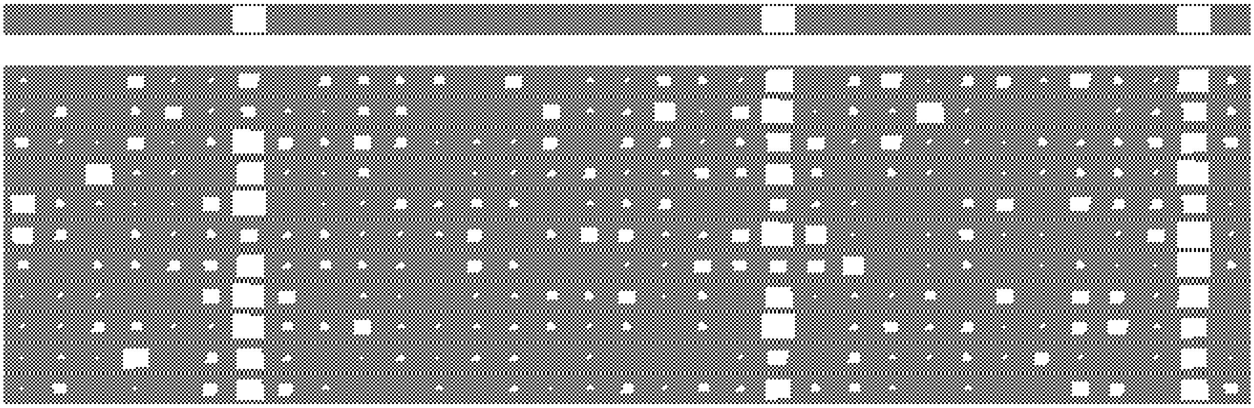
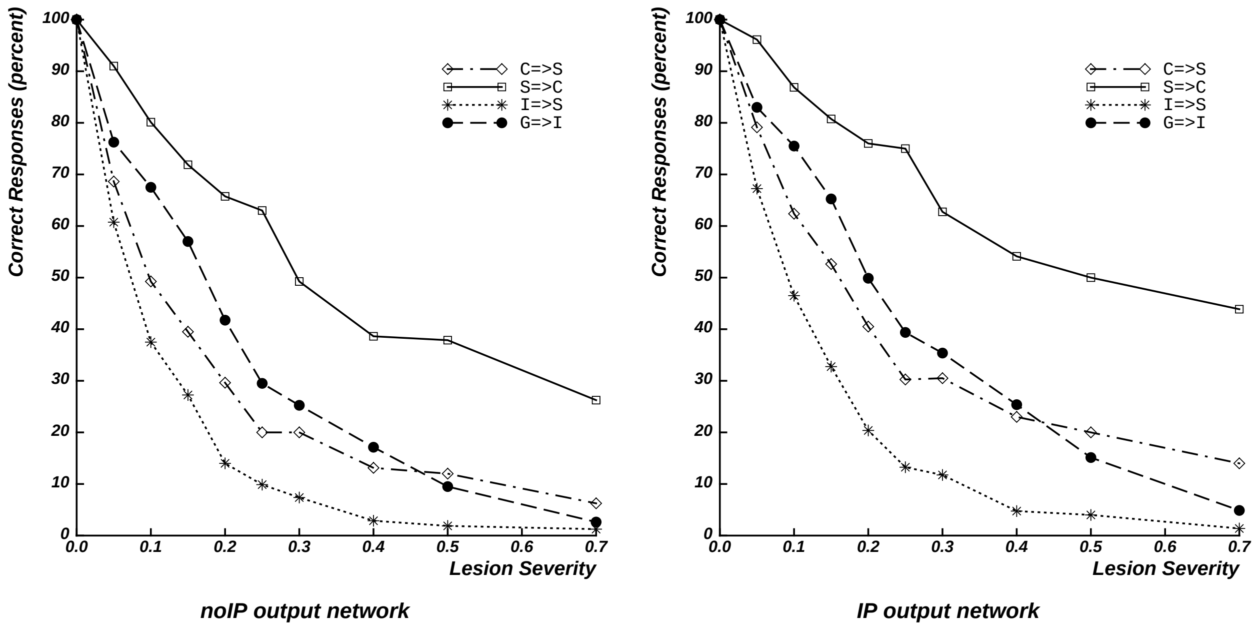


Figure 3.6: Examples of phoneme unit activities for the word COT corrupted by gaussian noise with standard deviation 0.25.

although the approach of developing phonological attractors independent of semantics is primarily computationally motivated, it is not unreasonable on empirical grounds that attractors for word pronunciations might develop as part of the process of learning to speak before these attractors would become available in reading.

Both the IP and noIP versions of the clean-up pathway were 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. Figure 3.6 provides examples of noisy inputs for the word COT. Because the phoneme units are both the input and output units for these networks, 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 (i.e. $\sigma^{-1}(y)$ where y is the activity and σ is the input-output function of the unit). 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 consistently met the stopping criterion and ceased to improve.

Two complete output networks were then formed by combining each of the two clean-up pathways with a separate copy of the direct pathway. The direct and clean-up 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 output networks with and without intra-phoneme connections have 2745 and 2374 connections, respectively. The two output networks were then given additional training on noisy input, during which only the weights in the direct pathway were



noIP output network

IP output network

Figure 3.7: Overall correct performance of the noIP and IP networks after removing various proportions of connections in each of the four main sets in the input network.

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, each output network was attached to separate copies of the input network to which the original output system was attached, and given a final tuning. In addition to the clean-up weights, the weights of the input network were also not allowed to change during this training to ensure that it continued to derive the correct semantics for each word. This final tuning ensured that each output network operated appropriately when its input was not clamped, but rather generated over time by an actual input network. Each of these final stages of training each required less than 100 sweeps through the set of words.

3.2.3 The effects of lesions

Fixing the weights of the input network during final tuning means that the IP and noIP output networks can be directly compared with the original output system, since all three output networks receive the identical semantic input. To further aid the comparison, the noIP and IP networks were subjected to the identical lesions as were applied to the original network (using the same random number generator seeds). In addition, the minimum phoneme response probability for the network to produce a response was increased from 0.5 to 0.6, as discussed in Section 3.1.4.

Figure 3.7 shows the overall performance rates of the two networks. Notice that the two patterns of correct responses across lesion locations are rather similar, but that the output network with intra-phoneme connections (IP) is more robust—that is, produces higher correct rates for equivalent lesion locations and severities (paired $t(35) = 12.9$, $p < .001$).

Figure 3.8 shows the distributions of error types for the noIP and IP networks. Although these data are roughly balanced for overall correct performance, lesions to the IP network produce much higher error rates (as opposed to omissions) compared with the noIP network. For both networks, the rate of blend errors is quite low at every lesion location, particularly for the network with intra-phoneme connections ($F(1, 54) = 9.71$, $p < .005$). In addition, the IP network has higher overall error rates ($F(1, 54) = 35.2$, $p < .001$) but also a higher proportion of "other" errors ($F(1, 54) = 19.7$, $p < .001$). These results all indicate that intra-phoneme connections contribute significantly to the development of strong attractors for words, but that one consequence of having such strong attractors is that words unrelated to the stimulus are more often produced as responses. Intra-phoneme connections also appear to influence the distribution of error types. In particular, the IP network produces a higher proportion of visual/phonological errors ($F(1, 54) = 49.2$, $p < .001$). This makes sense if the intra-phoneme connections are producing strong phonological attractors and many of the errors in this network that are categorized as visual actually result from phonological similarity. The fact that the rate of semantic errors is relatively low suggests that the damaged input network tends to produce mixtures of the semantics of words rather than the clean semantics of a single word, presumably due to the lack of sufficiently strong semantic attractors.

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 (see Figure 3.8). However, for both the noIP and IP network, the ratios for visual, mixed visual-and-semantic, and semantic errors to other errors are a number of times larger than those predicted by chance. For the noIP network, 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. For the noIP network, the ratios are larger by at least a factor of 3.2 for visual errors, 6.6 for visual-and-semantic errors, and 2.0 for semantic errors.

In addition, it is possible that mixed visual-and-semantic errors arise simply from the chance rate of semantic similarity among visual errors, and the chance rate of visual similarity among semantic errors, rather than reflecting an additional influence on errors. The expected rate M of mixed errors can be calculated from the observed rates V and S of visual errors and semantic errors assuming only a chance rate of similarity along the other dimension (Shallice & McGill, 1978):

$$M \leq V \frac{s}{1-s} + S \frac{v}{1-v}$$

where v and s are the proportions of stimulus-response pairs that are visually and semantically similar, respectively. In fact, the actual rates of mixed visual-and-semantic errors are higher than the expected rate for every lesion location using the noIP network but not the IP network. Thus,

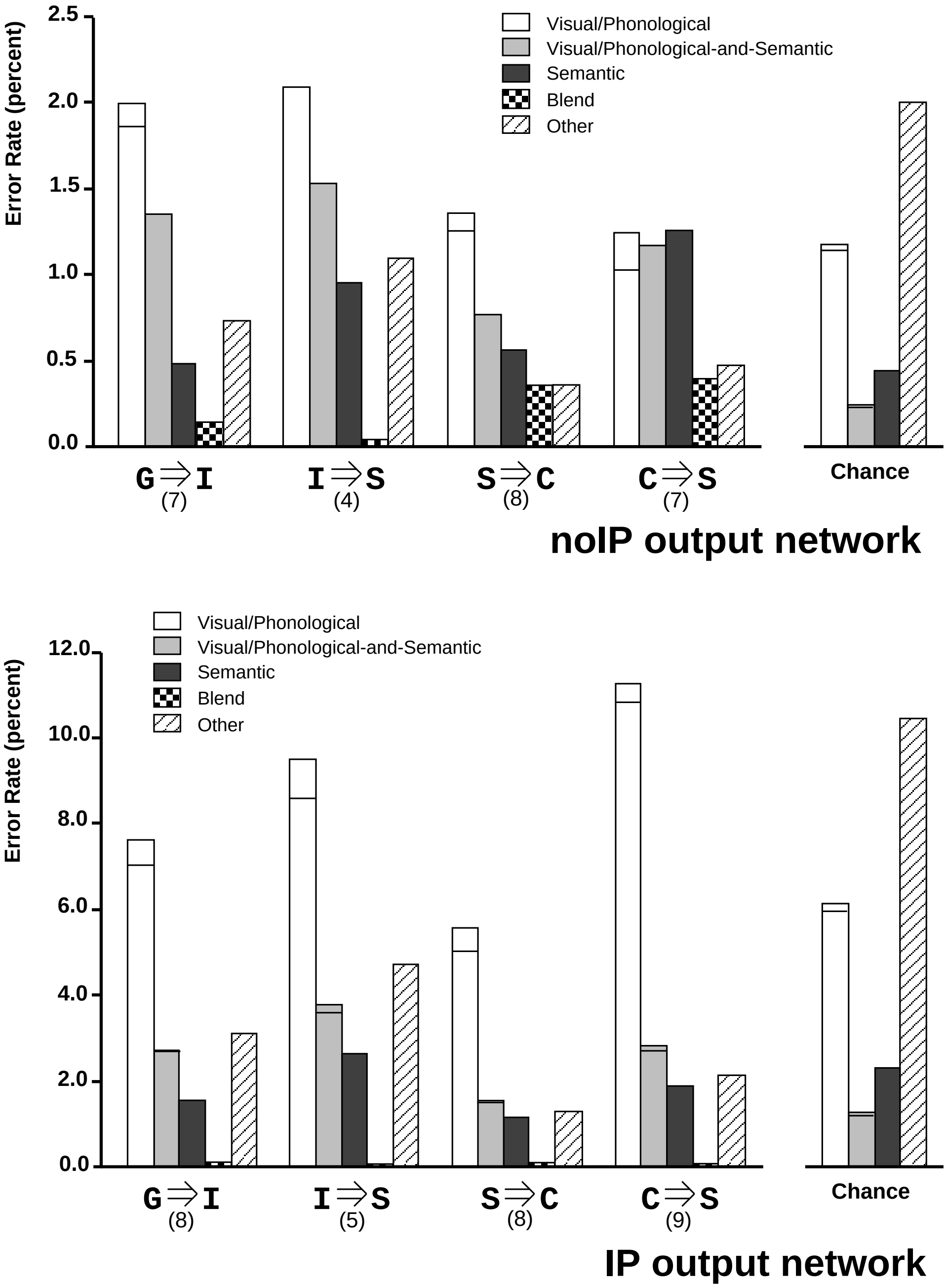


Figure 3.8: Error rates produced by lesions to each main set of connections in the input network of the noIP and IP networks. Notice that the y-axes are scaled differently in the two graphs, and the absolute heights of the "Chance" distributions are set arbitrarily.

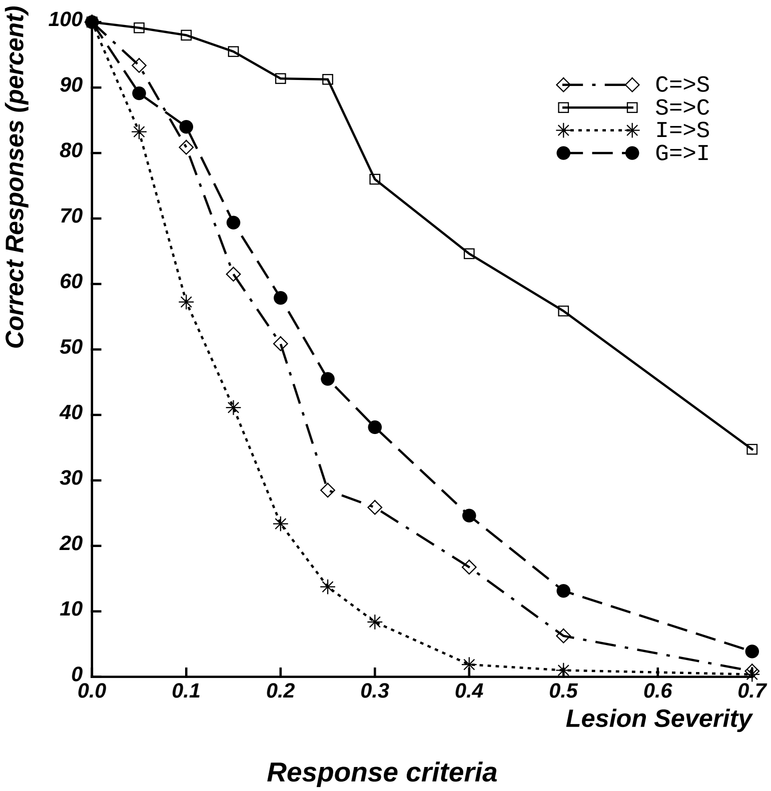


Figure 3.9: Overall correct performance using the response criteria, after removing various proportions of connections in each of the four main sets in the input network.

while both networks replicate the occurrence of visual, mixed visual-and-semantic, and semantic errors for lesions throughout the input network, the finding of higher than expected rates of mixed errors appears to be less general. We will consider the conditions under which it occurs in more detail in Chapter 4.

3.3 Comparison with response criteria

H&S approximated the behavior of a network for generating phonological output from semantics by applying *proximity* and *gap* criteria to the semantics produced by the lesioned network. They attempted to demonstrate that their results were not dependent on the exact values of these criteria, but they provided no evidence on their adequacy in approximating an actual response system. Given our success at implementing networks that map from orthography to phonology via semantics, we can now directly compare their behavior with those produced using the response criteria.

The identical set of lesions that were applied to the input network of the noIP and IP networks were now applied to input network in isolation. Correct, omission, and error responses were accumulated according to the response criteria. Figure 3.9 shows the percentage of words responded to correctly across the range of lesion densities of each of the sets of connections. In general, the

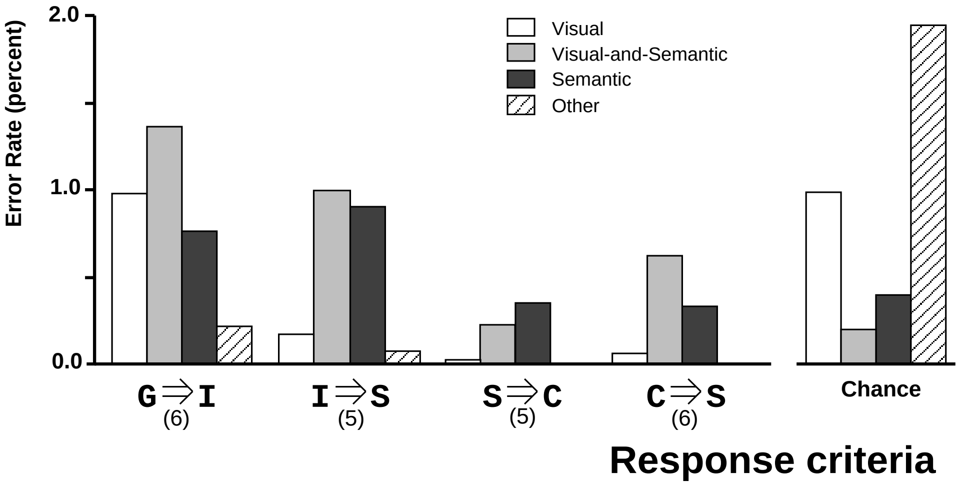


Figure 3.10: The relative proportion of error types produced by lesions to each main set of connections in the input network.

pattern of correct performance using the response criteria is quite similar to that produced using the output networks, particularly the one with intra-phoneme connections.

Figure 3.10 presents the rates of the various error types for each lesion location. The response criteria produce a lower overall error rate than either the noIP or IP networks ($F(1, 46) = 19.0$, $p < .001$ vs. noIP, $F(1, 50) = 46.4$, $p < .001$ vs. IP). Since these data are balanced for proportion of correct responses, this suggests that semantic patterns which fail the response criteria are frequently sufficient to produce (often incorrect) phonological output. The criteria also produce a lower proportion of "other" errors than either network ($F(1, 46) = 24.4$, $p < .001$ vs. noIP, $F(1, 50) = 207.1$, $p < .001$ vs. IP). While the proportion of visual errors is low for lesion locations other than G ⇒ I, their proportion relative to "other" errors is greater for all lesion locations than predicted by chance. The same applies to mixed visual-and-semantic and semantic errors, replicating the original H&S results. Furthermore, the rate of visual-and-semantic errors for each lesion location is much higher than that predicted from the rates of visual and semantic errors assuming independence. Perhaps most interestingly, the response criteria cause a much higher proportion of the errors to be semantically related to the stimulus ($F(1, 46) = 44.2$, $p < .001$ vs. noIP, $F(1, 50) = 298.5$, $p < .001$ vs. IP). As described in the previous section, the relatively weak semantic influences in the networks, particularly the one with intra-phoneme connections, suggests the attractors developed by the input network are insufficiently strong relative to those in the output networks. The use of criteria that apply directly to semantics compensate for (and therefore conceal) the limitations of the input network. Nonetheless, H&S's main results about the *qualitative* mixture of error types for lesions throughout the network stand.

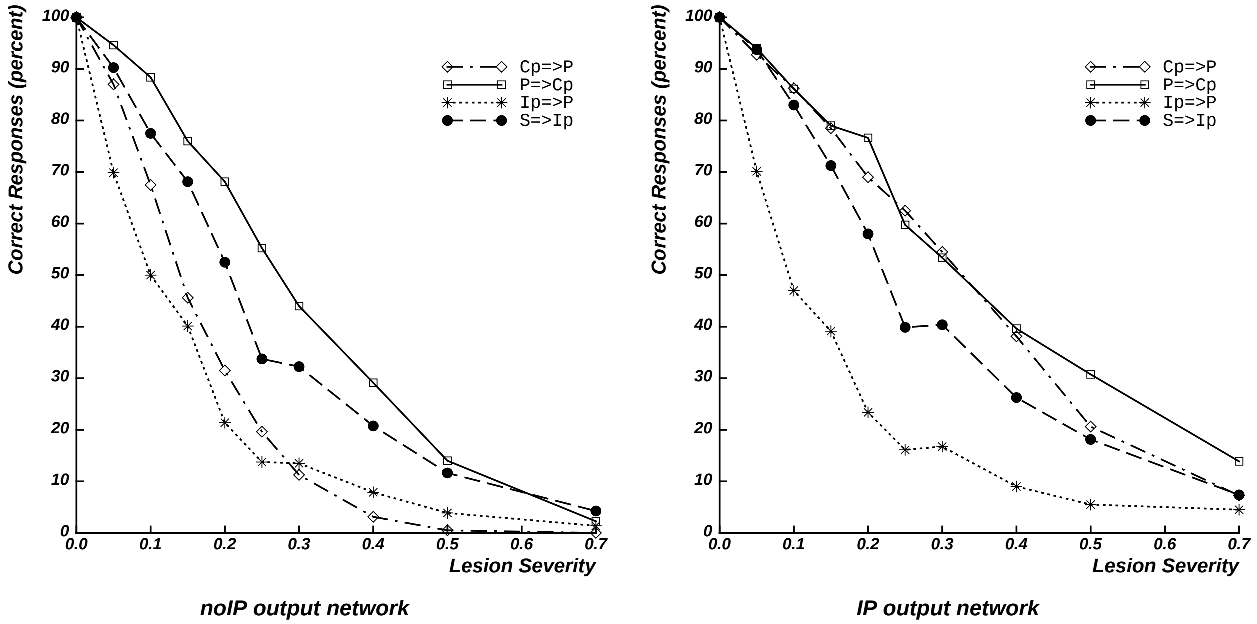


Figure 3.11: Overall correct performance of the noIP and IP networks after removing various proportions of connections in each of the four main sets in the output network. The initials for the intermediate and clean-up units are subscripted with "p" (for "phonological") to distinguish them from the intermediate and clean-up units in the input network.

3.4 Impairments in mapping semantics to phonology

Beyond revealing limitations of the original input network, implementing a phonological output system ensures that behavior under damage is due entirely to properties of the complete network and not to those of an interpretation procedure *external* to the network. Since the output system operates on the same principles as the input system, the number of independent assumptions of the entire system is minimized. In addition, a number of additional issues can be addressed in a model that maps orthography to phonology via semantics that cannot be addressed in a network that only derives semantics. In particular, it becomes possible to investigate impairments in deriving phonology from intact semantics by lesioning connections in the phonological output system. Many theories of deep dyslexic reading (e.g. Caramazza & Hillis, 1990; Coltheart et al., 1987a; Marshall & Newcombe, 1966) explain semantic errors entirely on the basis of this type of damage—implementing a complete semantic route allows us to compare how the resulting behavior compares with that produced by earlier damage.

Accordingly, we subjected each main set of connections in the output network of the noIP and IP networks to 20 instances of lesions of a variety of severity, accumulating correct, omission, and error responses. Figure 3.11 shows the overall correct performance for both networks. Two main points are of interest. The first is not particularly surprising—intra-phoneme connections make the IP net somewhat more robust to damage overall than the noIP network. The second is that intra-phoneme connections are particularly helpful for damage to connections from the clean-up units to the phoneme units. Presumably the direct interactions among phoneme units compensate to some extent for the lost (and erroneous) clean-up, and also make the network less dependent on the connections in the clean-up pathway.

Figure 3.12 presents the distributions of rates of errors categorized in terms of their visual/phonological and semantic similarity. Considering the network without intra-phoneme connections (noIP) first, lesions to the "direct" pathway (S⇒Ip and Ip⇒P) produce a mixture of visual/phonological errors and semantic errors with relatively few blends, but also a rather high proportion of "other" errors. As the lesions are subsequent to the operation of intact semantic clean-up, the high proportion of visual/phonological errors almost certainly reflects phonological rather than visual similarity.[6] 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, 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.

Lesions of the network with intra-phoneme connections (IP) produce a similar pattern of results. The additional strength of the phonological attractors in this network is evidenced by its much higher overall error rates, lower proportion of blends, and higher proportions of visual/phonological and other errors.

It is interesting to compare these effects of lesions on the "output" side of the noIP and IP networks with those produced by lesions on the "input" side (see Figure 3.8, p. 62). The error patterns for lesions to the direct pathways are quite similar, although output lesions tend to produce a somewhat stronger influence of semantic similarity and a higher proportion of "other" errors than input lesions. Not surprisingly, output clean-up lesions produce far fewer errors and far more blends than input clean-up lesions. However, for the IP network the distributions of error types other than blends for input and output lesions are fairly similar. Thus, lesions anywhere along the direct pathway from orthography to phonology via semantics produce qualitatively similar patterns of errors. In this way, the implication from H&S's results, that a patient's error pattern alone provides insufficient information for identifying lesion location, appears to generalize to lesions all along the semantic route.

[6] It is still possible that errors produced by damage after semantics would show influences of visual similarity. The output network receives input from semantics before its activity has settled correctly, and the initial semantic patterns are influenced by visual similarity (see Figure 2.10, p. 43, and the discussion in the following chapter). However, this effect on errors due to damage in the output network is likely to be small relative to the effect of phonological similarity.

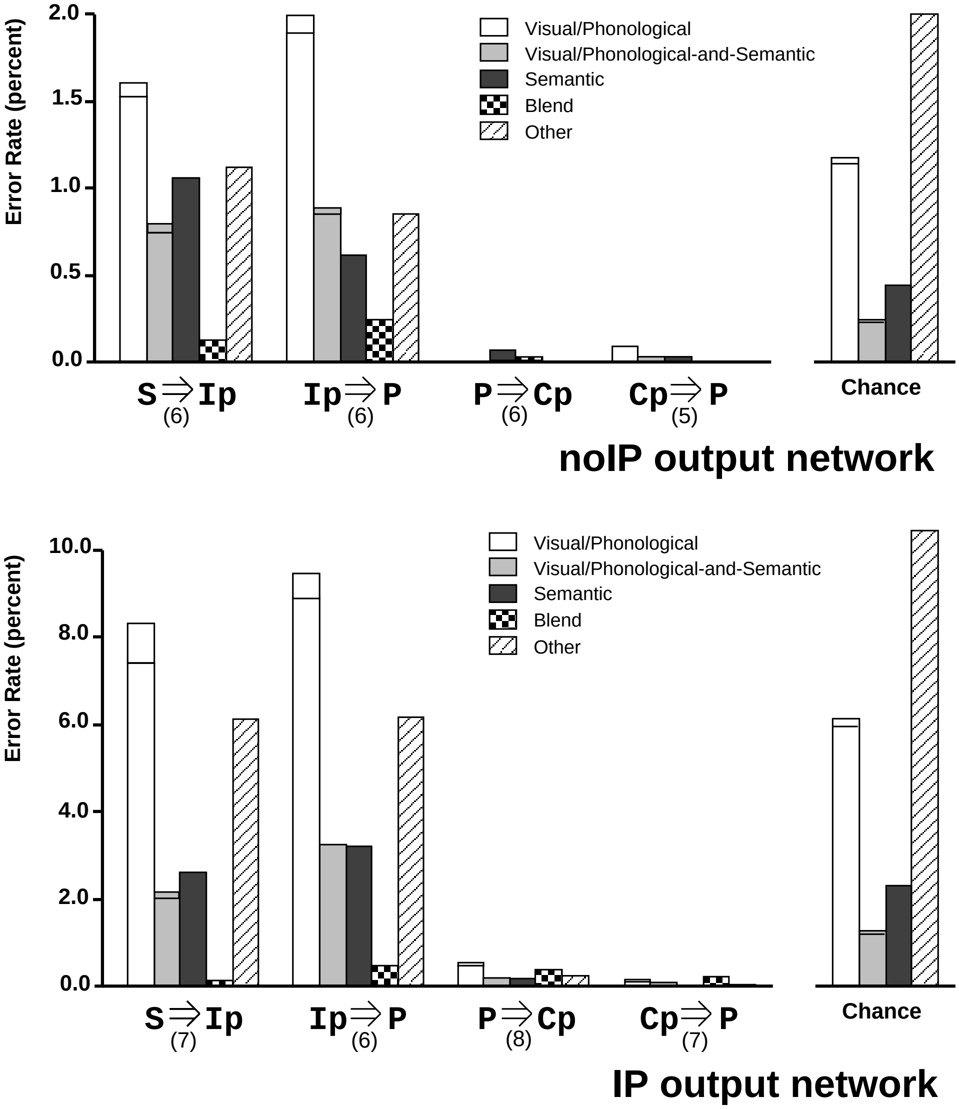


Figure 3.12: The relative proportion of error types produced by lesions to each main set of connections in the output half of the noIP and IP networks.

3.5 Summary

We have shown how the procedure that H&S used to derive explicit responses from their network can be replaced by extending the network to directly produce a phonological response on the basis of semantics. Lesion experiments with such a network replicated the main finding of a mixture of visual and semantic influences in errors for a variety of lesion locations. Lesions between semantics and phonology also produced qualitatively similar results, but with some interesting differences relating to the impact of phonological cleanup. The next chapter considers the generality of H&S's results from another perspective—the importance of network architecture.