

Chapter 4

The relevance of network architecture

Perhaps the most perplexing aspect of connectionist modeling is the design of network architecture, by which we mean choices of numbers of units and their connectivity. One reason the choices in network design often appear rather arbitrary is that they are influenced both by general connectionist principles and by the specific nature of the task at hand. Unfortunately, the general principles are rarely made explicit, and the effect of particular architectural decisions on different aspects of network behavior in a specific task is often ill-understood. H&S attempt to make explicit both the general and specific considerations that went into developing their model. The general considerations involve a tradeoff between ensuring that the network has sufficient capacity and “power” to solve the task, while keeping the network as small as possible to stay within available computational resources. The specific considerations center around attempting to facilitate the ability of the network to map between two domains, orthography and semantics, which are arbitrarily related. These two types of concerns influence the number, size, and interconnectivity of unit layers.

The number of units in the input and output layers (e.g. 28 grapheme units and 68 sememe units) is determined by the input and output representations chosen for the task. These layers are called “visible” because they make contact with the training environment. The simplest architecture would be to connect input units directly to output units, but such networks have severe computational limitations that prevent them from learning arbitrary associations (Minsky & Papert, 1969). In particular, they can only learn tasks that are linearly separable—that is, if each input pattern is viewed as a point in a state space with a dimension for each input unit, for each output unit there is some hyperplane in the space that separates the points for patterns for which the unit must be active from those for which it must be inactive. The general problem of mapping English orthography to semantics is not linearly separable using H&S’s input representation because the number of words (input patterns) so greatly exceeds the number of grapheme units (dimensionality of the hyperplanes). In general, to accomplish tasks that are not linearly separable it is necessary to add a least one layer of non-linear “hidden” units between the input and output layers (Ackley et al., 1985). Because these layers are not part of the input or output, the representations they use must be determined by a general learning procedure. Typically only one hidden layer is used because most

learning procedures slow down exponentially with the number of intervening hidden layers (see e.g. Plaut & Hinton, 1987). Such three layer networks are ubiquitous in connectionist modeling because they can learn any boolean function with enough hidden units (an exponential number in the worst case, but only a polynomial number for most “reasonable” functions; Denker et al., 1987).

An important consideration in determining the number of hidden units is the relative sizes of the input, hidden, and output layers.¹ A hidden layer that is much smaller than the visible layers must develop a compact code for the input that retains enough information to differentiate the appropriate outputs (e.g. “encoder” networks (Ackley et al., 1985) and RAAMs (Pollack, 1990)). In such an architecture the hidden units tend to accomplish this by capturing the important statistical regularities the input-output mapping (e.g. the largest principle components (Cottrell et al., 1987)), often resulting in appropriate generalization behavior when the network is presented with novel input. A hidden layer larger than the visible layers allows slight differences between input patterns to be magnified by increasing the distance between their representations (in the multi-dimensional feature space with one dimension for each hidden unit). Giving quite different representations to similar inputs is critical when the associations between inputs and outputs are arbitrary, such as mapping letter strings to meanings. Thus H&S use more intermediate units than grapheme units (but less than the number of sememe units).

In considering how units are connected, a major architectural distinction is between “feed-forward” and “recurrent” networks. In a feed-forward network, unit layers can be partially ordered such that units receive connections only from earlier layers. For a given input pattern, this restriction allows the final state of each unit to be computed in a single pass through the network, from input to output. However, for this very reason the extent that units in a feed-forward network can interact is extremely limited. In particular, feed-forward networks cannot develop attractors because each unit in the network only updates its state once—the network cannot reapply the unit non-linearities to clean-up a pattern of activity over time.² “Recurrent” networks have no restrictions on how units are connected, enabling interactions between units within a layer, and from later to earlier layers. When presented with input, units must repeatedly recompute their states, because changing the state of a unit may change the *input* to earlier units. In this way, recurrent networks can gradually settle into a stable set of unit states, called a “fixedpoint” or an “attractor,” in which unit inputs (and hence outputs) remain constant.³ Recurrent networks are particularly appropriate for temporal

¹Some recently developed connectionist learning procedures (e.g. Fahlman & Lebiere, 1990; Sietsma & Dow, 1988) avoid the issue of deciding at the outset how many hidden units to use by dynamically adding or removing units during learning, but the nature of the representations these procedures develop is not well understood.

²A feed-forward network can be thought of as developing “degenerate” attractors in the sense that, due to unit non-linearities, a set of similar input patterns may be “compressed” onto an even more similar set of output patterns. An output pattern with a neighborhood for which the compression of the input space onto it is locally maximal is analogous to an attractor in a recurrent network.

³In networks with stochastic units, a fixedpoint is achieved when the *probability distribution* of activity patterns remains constant (analogous to the notion of “thermal equilibrium” from statistical mechanics). Also, in addition

domains, such as language processing (Elman, 1990) and motor control (Jordan, 1986). They are also more effective at learning arbitrary associations because the reapplication of unit nonlinearities at every iteration can magnify initially small state differences into quite large ones. Feed-forward networks require very large weights (and hence very long training time) to map similar inputs to quite different outputs. As described in the Introduction, unit interactions in a recurrent network can fill-out and clean-up initially noisy or incomplete patterns—producing behavior in which the initial pattern of activity “moves” towards the nearest attractor state.

The existence of attractors for word meanings forms the basis for H&S’s explanation of the co-occurrence of visual and semantic errors in deep dyslexia. In order to allow such attractors to develop, H&S introduce direct connections among closely related sememe units. However, these connections only allow *pairwise* interactions—there is no way for *combinations* of sememes to have direct influences. For example, only the conjunction of “green” and “found-woods” implies “living”—neither feature alone does. These higher-order semantic “micro-inferences” (Hinton, 1981a) strengthen the attractors for words (i.e. increase the sizes and depth of their basins of attraction) by filling-out the initially incomplete semantics generated bottom-up and with only pairwise interactions. In order to implement them there must be hidden units that receive connections from some sememe units and send connections to others. While H&S could have used the intermediate units for this purpose by introducing feedback connections to them from semantics, they chose to introduce a second set of hidden (clean-up) units as an approximation to the influences of other parts of the cognitive system on semantics. In addition, separating the groups of hidden units allows them to specialize differently: one group can directly mediate between orthography and semantics; the other can make inferences among semantic features.

A final consideration in architecture design is the pattern of connectivity between layers of units. The capacity of a network is largely determined by its number of connections since the weights on these connections encode the long-term knowledge used to solve the task. For a given number of weights, there is a trade-off between using many, sparsely connected units versus using fewer, densely connected units. As described above, using many units results in a higher-dimensional representation in a layer, allowing easier discrimination between similar patterns in earlier layers. However, because each unit is only sparsely connected to layers providing input, the complexity of the distinctions it can learn is limited.⁴ In particular, as connectivity density is reduced it becomes harder for individual units to be sensitive to global structure in earlier layers and enforce global coherence in later layers.

Most connectionist networks use complete connectivity between layers, but this can result in a

to “point” attractors, recurrent networks can be trained to settle into “limit cycle” (Pearlmutter, 1989) and “chaotic” attractors (Skarda & Freeman, 1987), but this type of behavior is not directly relevant for our purposes.

⁴Specifically, if each sigmoidal unit is viewed as constructing a “soft” hyperplane decision surface in the space of activity patterns in earlier layers, sparse connectivity restricts the unit’s hyperplane to be parallel to all of the axes for units from which it receives no connection.

large number of connections for networks with even a moderate number of units. Full connectivity between layers in the H&S network would have resulted in almost 17,000 connections. Networks with far more capacity than is required to learn a task tend to approximate a “table-lookup” strategy without capturing any interesting structure in the task. Accordingly, H&S chose to include only a random quarter of the possible connections between layers, and intra-sememe connections only among related semantic features, to reduce the network to a computational reasonable size (about 3300 connections). In addition, reduced connectivity made the bottom-up input from orthography to semantics relatively impoverished, particularly because the usefulness of individual intermediate units can be significantly constrained by the absence of individual $G \Rightarrow I$ connections when input letters are represented by single grapheme units. H&S argued that impoverished bottom-up input to sememe units encouraged reliance on clean-up interactions, resulting in stronger semantic attractors.

Even among recurrent networks with hidden units that build strong attractors with a minimum number of connections, there are a vast number of possible network architectures. H&S chose one and demonstrated that its behavior under damage had interesting similarities with the reading behavior of deep dyslexics. It is clearly infeasible for computational reasons to implement every alternative architecture in order to investigate the generality of the H&S results. However, it is important to gain a better understanding of the relevance of the particular aspects of their design. In this chapter, we develop five alternative architectures which differ from the H&S model in terms of numbers of hidden units, connectivity density, existence of intra-sememe connections, location of clean-up pathway, and separation of intermediate and clean-up units. We then systematically lesion each of these networks and compare their behavior using the response criteria as well as one of the phonological output networks developed in the previous chapter, in order to better understand the impact of architectural differences on behavior under damage.

Following the separate consideration of the general behavior of each architecture, we address a number of issues regarding more detailed aspects of the pattern of correct and impaired performance shown to varying degrees by all of these networks. Among these issues are item- and category-specific effects, the impact of the definitions of visual and semantic similarity, the effects of lesion severity on the distribution of error types, and the analysis of particular errors and error types.

4.1 Alternative architectures

Figure 4.1 depicts each of the five alternative architectures for mapping orthography to semantics. We will refer to each network by using an ideographic character that expresses the essential aspects of its architecture. In these characters, the orthographic units and $O \Rightarrow I$ connections are not included as they are the same for all of the networks. The semantic layer is emphasized in bold, and the relative size of each layer is proportional to the number of units it contains. Each major connection

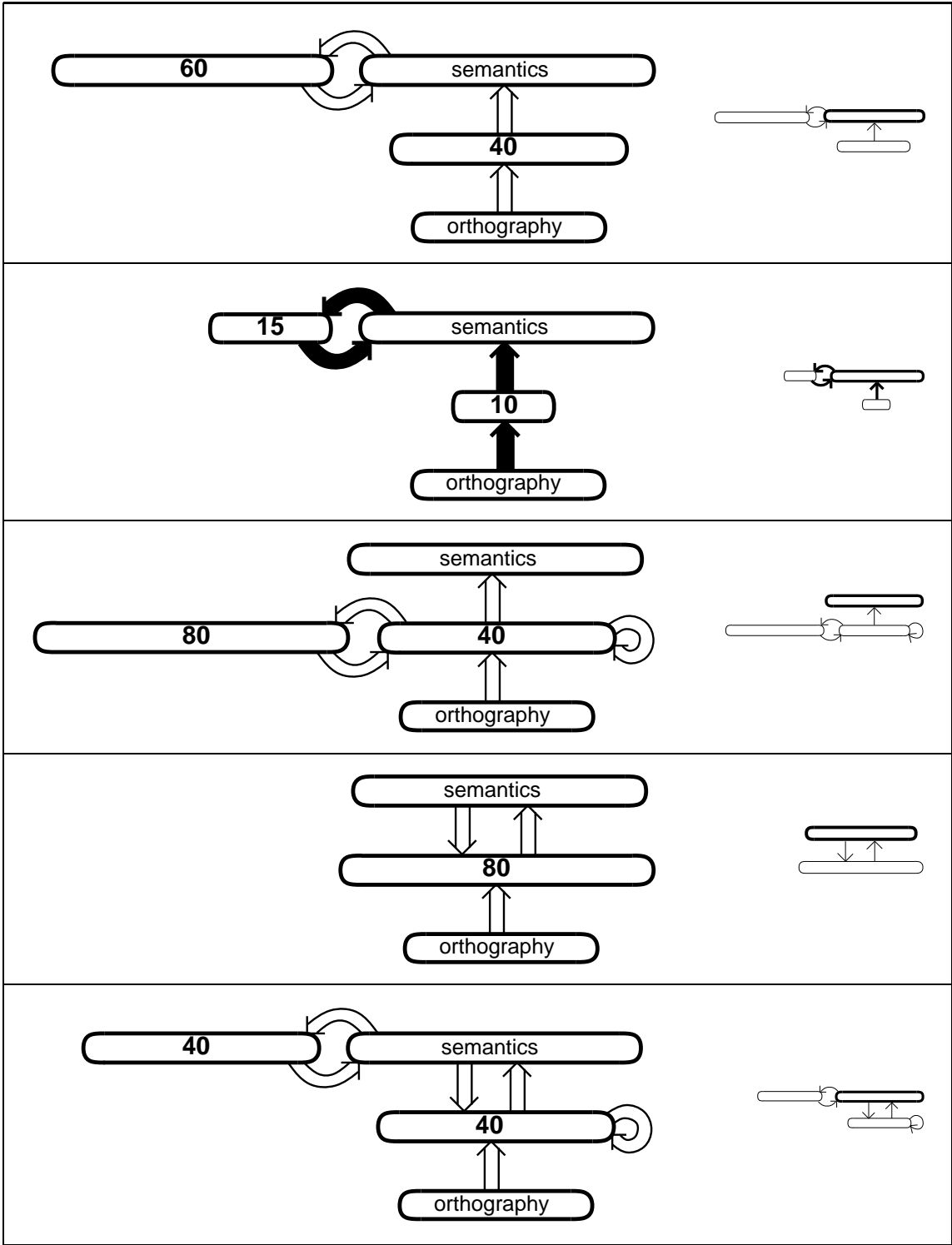
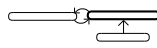
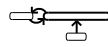
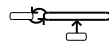
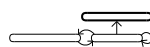




Figure 4.1: Five alternative network architectures for mapping orthography to semantics.

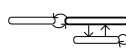
pathway between layers is depicted by a small arrow. The networks, and the main issues they are intended to address, are the following:

 *Intra-sememe connections.* This network most closely approximates the original H&S network, with 40 intermediate units, 60 clean-up units, and 25% connectivity density. However, it lacks any direct connections among sememe units, so it will allow us to investigate the importance of such connections. The network has 3252 connections.

 *Connectivity density.* Rather than using 25% connectivity density, the  network has complete connectivity between layers (as indicated by the bold arrows). Lesions to this network will allow us to evaluate the impact of connectivity density. In order to keep the number of connections approximately the same as the other networks, only 10 intermediate units and 15 clean-up units were used. The resulting network has 3134 connections.

 *Location of clean-up.* This network has clean-up *prior* to semantics, at the level of the intermediate units, rather than within semantics. We can thus evaluate the importance of the location of cleanup on behavior under damage, and whether the attractors must be *semantic* in order to produce the H&S results. Specifically, the intermediate units are reciprocally interconnected with 80 clean-up units, as well as interconnected among themselves. All connection pathways have 25% density, for a total of 3226 connections.

 *Separation of intermediate and clean-up units.* Seidenberg & McClelland (1989) propose a framework for mapping among orthography, phonology, and semantics. Although they only implement a feed-forward version of the orthography-to-phonology mapping, the  network is intended to approximate their proposed orthography-to-semantics pathway. Specifically, 80 intermediate units both send connections to the sememe units, and receive feedback connections from the sememe units. There are no separate clean-up units, and so this network allows us to evaluate the importance of having separate groups of units for this function. The network has 25% connectivity density, resulting in 3550 connections.

 *Hybrid architecture.* This network is a hybrid of the Seidenberg & McClelland architecture and the H&S architecture. The network includes both feedback connections from sememe to 40 intermediate units and a clean-up pathway with 40 units. The intermediate units are also intra-connected. Our intention in developing this network was to investigate whether having these various means

A	0 1 0 1 0 1 1 0	J	1 1 1 0 0 0 0 0	S	0 0 1 0 0 0 0 1
B	1 0 1 1 1 0 0 1	K	1 0 0 0 1 0 1 1	T	1 1 0 0 0 1 0 0
C	0 0 1 0 1 0 0 0	L	1 1 0 0 0 0 0 1	U	1 0 1 0 0 1 0 0
D	1 0 1 1 1 0 0 0	M	1 0 0 0 0 1 1 1	V	0 0 0 0 0 1 1 0
E	1 1 0 0 1 0 0 0	N	1 0 0 0 0 0 1 0	W	0 0 0 0 0 1 1 1
F	1 1 0 0 0 0 0 0	O	0 0 1 1 1 1 0 0	X	0 0 0 0 1 1 1 0
G	0 1 1 0 0 0 0 1	P	1 0 1 1 0 0 0 0	Y	1 0 0 0 0 1 1 0
H	1 1 0 0 1 1 0 1	Q	0 0 1 1 0 0 1 0	Z	0 1 0 0 0 0 1 1
I	1 1 0 0 1 1 0 0	R	1 0 1 1 0 0 1 1		

Table 4.1: The assignment of features to letters. The meanings of the features are roughly (1) contains a vertical stroke; (2) contains a horizontal stroke; (3) contains a curved stroke; (4) contains a closed part; (5) horizontally symmetric; (6) vertically symmetric; (7) contains diagonal stroke; (8) discriminator between otherwise identical letters.

of developing attractors would make them more robust. With 25% connectivity density, the network has 3626 connections.

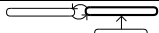
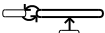
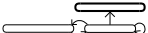

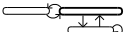
4.2 The task

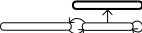
The task of each network is to generate the semantics and phonology of each of the 40 words used by H&S when presented with its orthography. The representations of semantics and phonology is the same as was described in Section 3.1.1. However, orthography is represented somewhat differently, in order to be consistent with related research using other words (described in Chapter 6). Instead of using a separate unit for each possible letter at a position, we describe each letter in terms of a distributed code of eight features, shown in Table 4.1. The set of features was designed to ensure that visually similar letters (e.g. E and F) have similar representations, while keeping the number of features to a minimum. Since the H&S word set has some four-letter words, a total of 32 “orthographic” units will serve as the input layer of each network.

4.3 The training procedure

Each input network was trained in the same way as the H&S network, with two differences. The first is that, as described in Section 3.1.3, the network was allowed to run for eight instead of seven iterations. The second difference is that the orthographic input presented to each network was corrupted by independent gaussian noise with mean 0.0 and standard deviation 0.1. Section 3.2.2 explains how training with noisy input encourages the network to develop more robust attractors. Training continued until each network could activate the correct semantic features for each word

to within 0.1 of its correct value. For each network, the following number of sweeps through the set of words was required:

Network	Sweeps
	2640
	3625
	14008
	7302
	4083

Training required a few thousand sweeps for all but the  network. The reason that this latter network took so much longer is that it lacks any interactions among sememe units, so these units cannot clean themselves up into near-binary responses. They must rely on the clean-up at the intermediate level to eliminate the influences of noise and drive them appropriately. Driving units into binary responses using only feed-forward connections typically involves traversing down the bottom of a long, shallow ravine in weight space, which requires many sweeps through the training set (see Plaut & Hinton, 1987).

Once each input network had learned to correctly map from orthography to semantics, the phonological output networks developed in Chapter 3 were combined with separate instances of each. The weights in the output networks were then allowed to tune themselves while the weights in each input network were held fixed. After this final training, which took at most a few hundred additional training sweeps, each combined network would correctly derive the phonology (and semantics) of each word from its orthography.

4.4 The effects of lesions

Twenty instances of lesions of a range of severity were applied to the main sets of connections in each input network in isolation, as well as to each network combined with the noIP and IP phonological output networks. Correct, omission, and error responses were accumulated using the response criteria for the isolated networks, and using a minimum response probability of 0.6 for the combined networks. Each error response was categorized in terms of its visual, semantic, and phonological similarity to the stimulus. The percentages of overall correct responses and distributions of error types were then determined for each network. We only present data using the output network without intra-phoneme connections because the only differences between its pattern of results and that using the IP network are those previously described in Chapter 3. For each input network, we will examine its lesion results in light of the issues that motivated its development.

4.4.1 The network

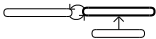
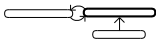
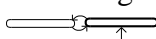
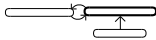
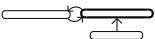
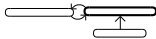
The main purpose of the  network is to evaluate the impact of intra-sememe connections, which were a rather inelegant aspect of the original H&S architecture. However, the network also differs from the original network in its use of a distributed orthographic code and in being trained with noisy input, so these manipulations must be taken into account in interpreting the data.

Figure 4.2 presents the lesion data for the  network. First consider the data using the response criteria (at the top of the figure). It is interesting to compare these results with those for the replication of the original network (Figures 3.9 and 3.10, pp. 63 and 64). In terms of correct performance, the  is more robust overall, presumably because training with noisy input fosters stronger attractors, as was shown in Chapter 3. Surprisingly, $0 \Rightarrow I$ lesions are more debilitating than $I \Rightarrow S$ lesions, and $C \Rightarrow S$ lesions are *less* debilitating than in the original network, violating the general tendency for damage closer to semantic to be more disruptive than more distant damage. This result may be due to the additional richness of the distributed orthographic code, allowing greater reliance on the direct pathway, particularly the portion nearest the input.

The distributions of error types are quite similar for the two networks. Visual errors are most prevalent for “early” lesions (to $0 \Rightarrow I$) while semantic errors occur most for “late” lesions (to $I \Rightarrow S$, $S \Rightarrow C$ and $C \Rightarrow S$). This progression can be quantified in the following way. We compute the ratio of semantically similar errors to visually similar errors and compare it with the “chance” ratio for word pairs chosen randomly from the word set (0.498). The ratio of the observed and chance ratios provides a measure of the “semantic bias” in the errors produced by lesions at each location. Ratios greater than 1.0 indicate that the network is biased towards semantic as compared with visual similarity. For the  network, the ratios are $0 \Rightarrow I$: 1.04 $I \Rightarrow S$: 2.47, $S \Rightarrow C$: 4.02, $C \Rightarrow S$: 4.82. Thus $0 \Rightarrow I$ lesions produce little bias of visual vs. semantic similarity in errors, while lesions closer to semantics show a strong bias towards semantic similarity. However, both “early” and “late” lesions produce a mixture of error types. In fact, the ratio of the observed error rates with that of “other” errors is at least three times the chance value (and often much greater) for all error types and lesion locations using both the response criteria and the noIP output network. The  network shows a stronger tendency to produce visual errors and also errors unrelated to the stimulus than the H&S replication network. While both networks produce high rates of mixed visual-and-semantic errors, the rates for the  network are not above those predicted from the independent rates of visual errors and semantic errors (except for $C \Rightarrow S$ lesions). This is mostly due to the high rates of semantic errors. Overall, intra-sememe connections (and orthographic representation) effect the quantitative but not qualitative results. A comparison of the data using the phonological output network with the corresponding data for the H&S replication network (Figures 3.7 and 3.8, pp. 60 and 62) also shows a similar pattern of results for the two input networks. While an additional replication of the H&S results using a different input representation and no intra-sememe connections is reassuring, the main role of the

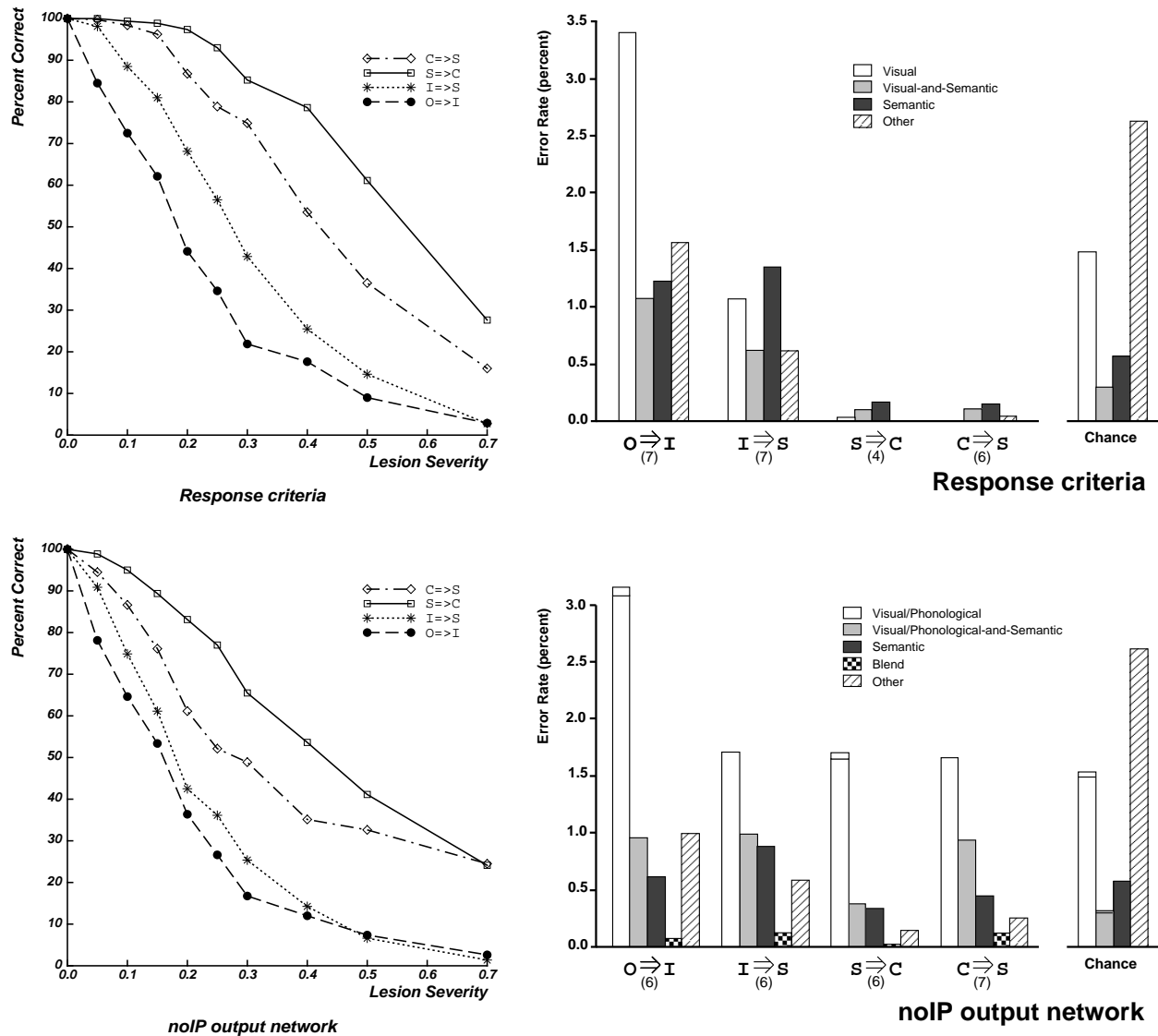
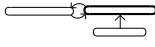
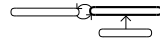
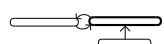


Figure 4.2: Overall correct performance and error distributions for the  network using the response criteria and the output network without intra-phoneme connections.

 network is to serve as a basis of comparison for the other networks.

One important comparison between networks is the extent that their direct pathways can convert from visual to semantic similarity by themselves. In particular, the representations at the level of the intermediate layer can be thought of as finding a compromise or “splitting the difference” between the visual similarity of the input and the semantic similarity of the output. We would like a way of comparing the nature of this compromise for different networks. More generally, it would be informative to have a measure of the extent that the representations in different parts of a network, or at different times during settling, are structured visually vs. semantically. One way to do this is to run the network and determine the pattern of activity that represents each word at a given layer and iteration. We can then compute the similarity matrix for these representations—that is, the set of proximity values for all pairs of patterns. If the representations are structured semantically, their pairwise proximities should approximate those among the actual semantic representations (shown in Figure 2.7, p. 36). Thus the degree of correlation of the two sets of proximity values provides a numeric measure of the extent that the patterns of activity for each word at a given layer and iteration are structured semantically (and similarly for visual structure). In addition, comparing graphical displays of the similarity matrices for the word representations at different layers and iterations (and with that for the semantic representations) provides a visual impression of the extent that the network has succeeded in organizing the words semantically at various points in processing.

Figure 4.3 presents the similarity matrices for the representations of words at various points in the operation of the  network: at the intermediate layer at iteration 1, and at the semantic layer at iteration 2 (when input first arrives from the direct pathway) and at the semantic layer at iteration 4 (when clean-up has its first influence). In addition, the figure displays the correlation of these matrices with those for the input (visual) and output (semantic) representations. The first thing to notice is that the network converts from visual to semantic similarity gradually. This is reflected both in the way that the visual appearance of successive matrices increasingly approximates the appearance of the matrix for the final semantic representations, and more directly in the shift in the visual and semantic correlation coefficients.⁵ Indeed, most of the off-diagonal high proximity values in the matrix for the intermediate representations (at the top) are for visually similar word pairs (e.g. PORE/PORK, RAM/PAN). Interestingly, the matrix for iteration 2 (in the middle) reveals that similarity within some categories—namely, animals, outdoor objects and, to a lesser extent, foods—already closely matches their final similarity based on the operation of the direct pathway alone. We will consider category-specific effects more thoroughly later in this chapter.

⁵The visual correlation at iteration 4 is significantly positive (near 0.3) because the diagonal terms (which are all 1.0 in both matrices) were included in calculating the matrix correlations. The correlation between the similarity matrices for the visual and semantic representations of words when these terms are excluded is actually slightly negative (−0.04), providing additional evidence that visual and semantic similarity are unrelated for the representations we use.

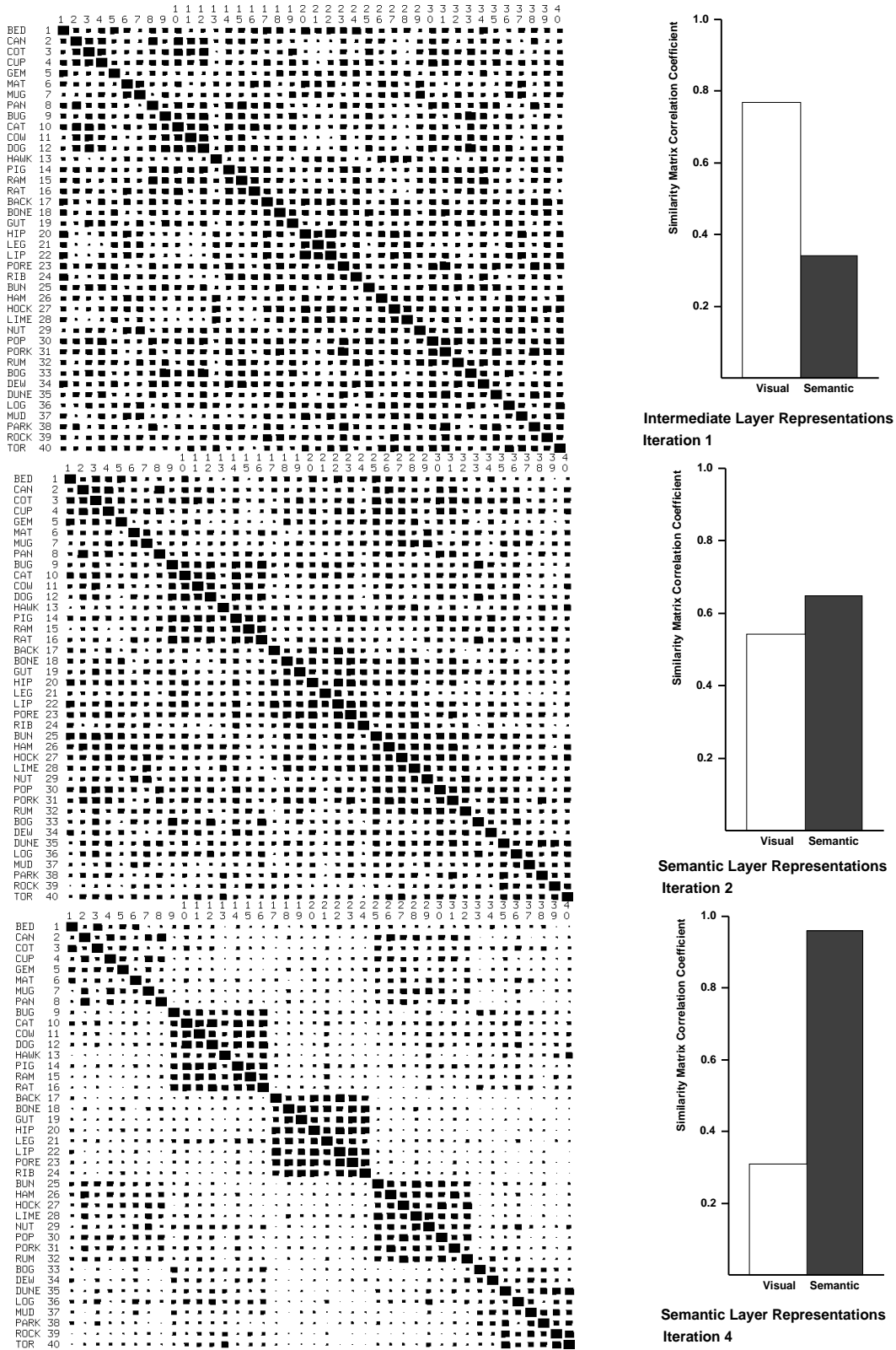
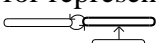


Figure 4.3: Similarity matrices and their correlation coefficients with matrices for visual (orthographic) and semantic similarity, for representations at the intermediate layer and for iterations 2 and 4 at the semantic layer of the  network.

4.4.2 The network

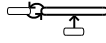
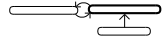
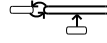
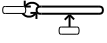
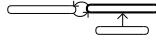
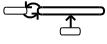
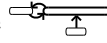
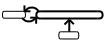
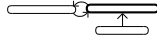
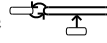
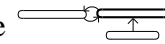
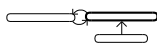
The purpose of the  network is to evaluate the impact of connectivity density on performance under damage. However, in comparing it with the  network it is important to keep in mind that, in order to balance for total number of connections, the  network also has fewer hidden units.

Figure 4.4 presents the lesion data for this network. Comparing the  and  networks in terms of overall correct performance, the  network is slightly more robust, particularly for lesions to $C \Rightarrow S$ connections. However, the relative importance of the various sets of connections are fairly similar in the two networks.

Considering the error response of the two networks, the  produces a much higher overall error rate both using the response criteria and using the phonological output network. However, since the  also has no intra-sememe connections, $C \Rightarrow S$ lesions still produce virtually no errors using the response criteria, reflecting an impairment of the only source of semantic clean-up. For the remaining locations, the ratio of the rates of all error types with that of “other” errors is at least twice the chance value. However, only clean-up lesions produce mixed visual-and-semantic errors at rates higher than expected from the independent visual and semantic rates. While the distribution of error types for lesions to the direct pathway are remarkably similar for the two networks, lesions from semantics to the clean-up units yield a weaker influence of visual similarity in the  network. In fact, the biases towards semantic vs. visual similarity in errors are lower for lesions of the  network ($O \Rightarrow I$: 0.73, $I \Rightarrow S$: 2.13, $S \Rightarrow C$: 1.95, $C \Rightarrow S$: 3.30) although still greater than the “chance” value for all but $O \Rightarrow I$ lesions. These effects are consistent with H&S’s claim that reduced connectivity in the  network impoverishes the bottom-up input to semantics, placing a stronger emphasis on semantic clean-up. However, the results may also be due to the fact that the extra intermediate units in the  network allow the input patterns to be more effectively separated, reducing the influence of their (visual) similarity in generating an output.

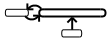
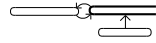
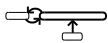
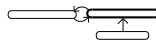
Interestingly, reduced connectivity density appears to produce a *stronger* influence of visual similarity for $S \Rightarrow C$ lesions when the phonological output network is used—however, this is most likely due to the influence of phonological rather than visual similarity. One possible explanation is that the complete connectivity of the clean-up pathway in the  network helps it build stronger semantic attractors. Less effective semantic clean-up in the  network leads to relatively inaccurate patterns of semantic activity, which are often cleaned-up into phonologically similar responses by the output network.

Figure 4.5 presents the similarity matrices for the word representations, and their correlations with the visual and semantic similarity matrices, at the intermediate layer at iteration 1, and at the semantic layer at iterations 2 and 4 in the  network. The similarity matrix for the intermediate units is much sparser than for the  network because representations over

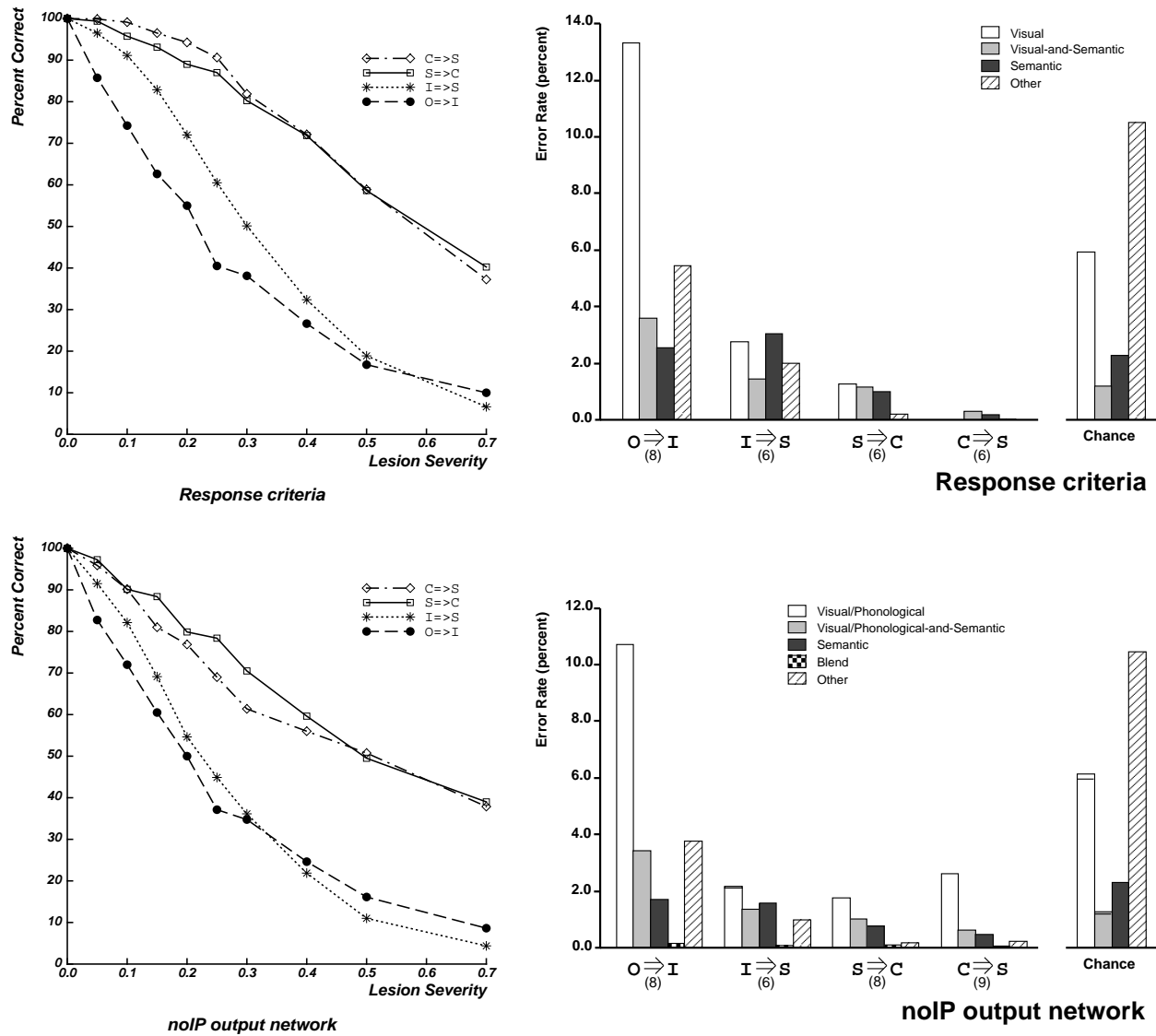
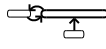


Figure 4.4: Overall correct performance and error distributions for the  network using the response criteria and the output network without intra-phoneme connections.

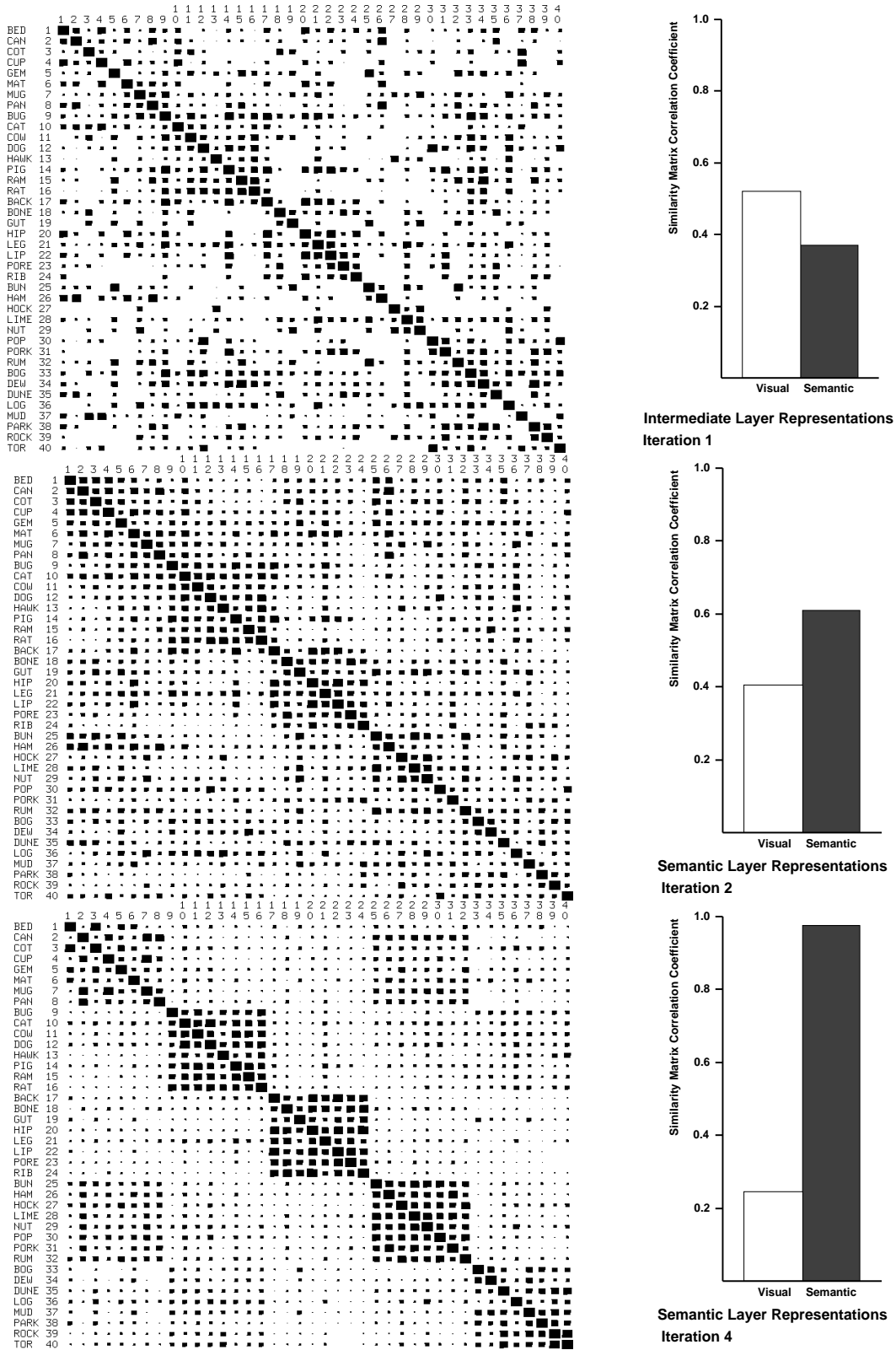
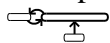


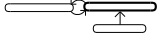
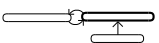




Figure 4.5: Similarity matrices and their correlation coefficients with matrices for visual (orthographic) and semantic similarity, for representations at the intermediate layer and for iterations 2 and 4 at the semantic layer of the  network.

10 units tend to be more binary than over 40 units, so that the patterns for pairs of words are either quite similar or nearly orthogonal. Interestingly, the representations are far less visually-organized, and slightly more semantically-organized, in the  network. This difference is also reflected in the semantic representations at iteration 2. Apparently, the direct pathway of the  network is more capable of converting from visual to semantic similarity on its own than in the  network. However, close inspection of the similarities within categories (e.g. indoor objects) reveals that the direct pathway is generating representations that are *more* similar than their final representations. This follows from the notion that the intermediate representations of words are more semantically organized but also more binary—the representations minimize fine semantic distinctions that must be amplified by the clean-up pathway to a greater extent than in the  network.⁶ In this sense, the  develops stronger semantic attractors.

Overall, the  network with full connectivity produces a quite similar pattern of errors (with a higher error rate) as a network with similar architecture but only 25% connectivity density. However, this network also had far fewer hidden units, and so the relative influences of connectivity density and number of hidden units have not been investigated.

4.4.3 The network

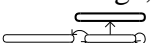
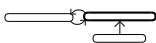
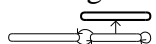
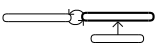
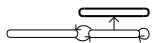
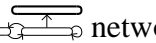
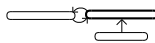

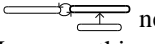
A number of results in this and the previous chapter demonstrate the importance of attractors in producing errors under damage. One question that arises is whether these attractors must be *semantic*. That is, must it be the sememe units that interact to form attractors in order for the network to produce semantic errors under damage, or would any sort of interaction suffice? To investigate this issue we developed the  network, in which attractors form at the level of the *intermediate* units prior to semantics.

Figure 4.6 presents data on the effects of lesions to this network. In terms of overall correct performance, the network shows the standard pattern that damage to the direct pathway is more debilitating than damage to the clean-up pathway. However, unlike the  network, the  network is less robust to $I \Rightarrow S$ lesions than to $0 \Rightarrow I$ lesions, reflecting the lack of semantic clean-up. On the other hand, comparing $I \Rightarrow C$ lesions with $S \Rightarrow C$ lesions in the  network suggests that the  network relies more heavily on its clean-up pathway for correct performance.

The  network produces slightly lower error rates than the  network. The ratios of the rates of each error type to that of “other” errors is at least seven times the chance value using the response criteria, and at least twice the chance value using the noIP output network (except for semantic errors after $I \Rightarrow I$ lesions, which are not above the chance value). For both

⁶According to this interpretation, the  network should be better than the  network at within-category forced-choice discriminations after clean-up damage (see Section 2.6.4). However, this prediction has not been tested.

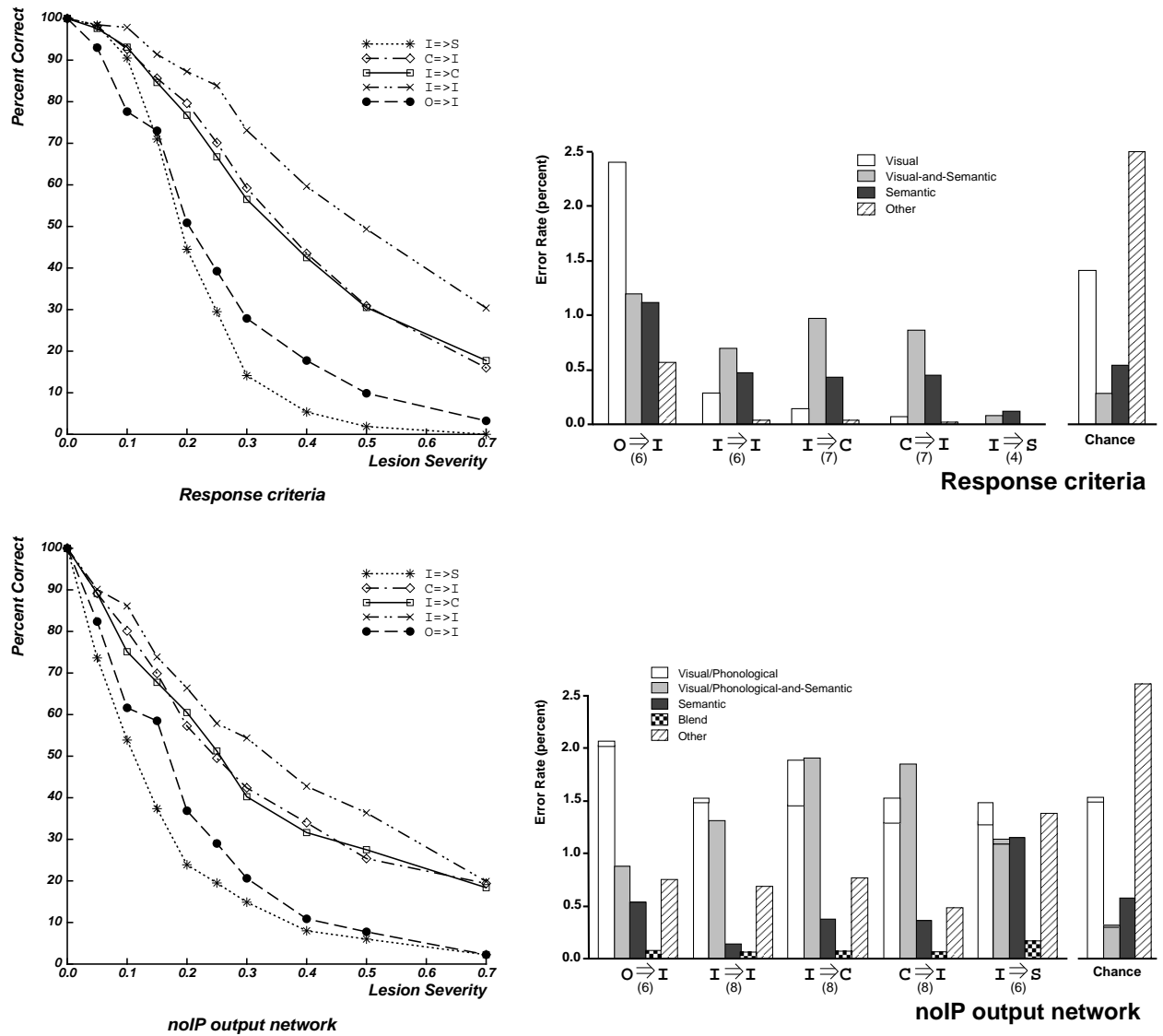
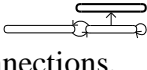
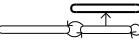
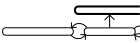
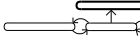
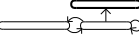
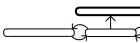
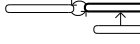
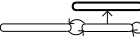
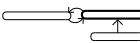
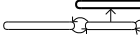
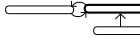
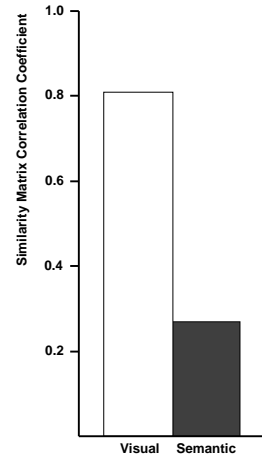
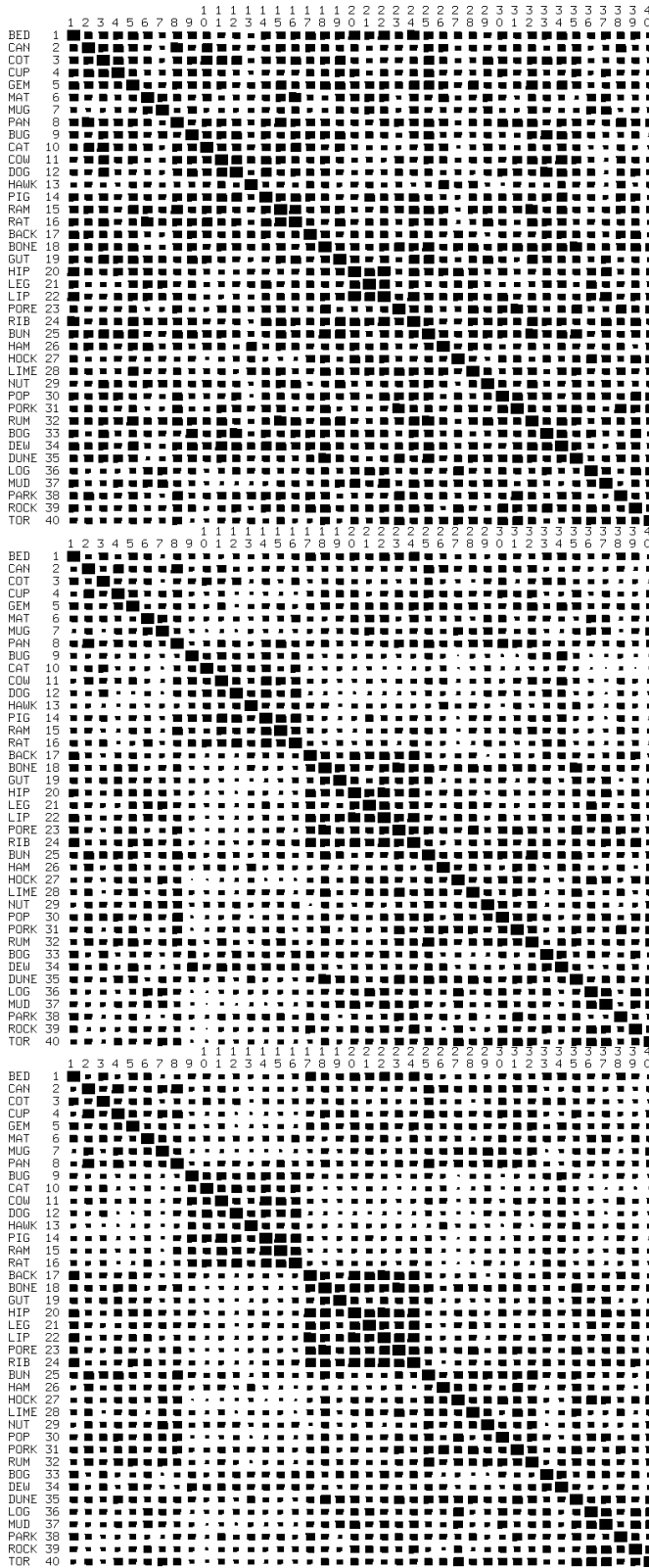


Figure 4.6: Overall correct performance and error distributions for the  network using the response criteria and the output network without intra-phoneme connections.

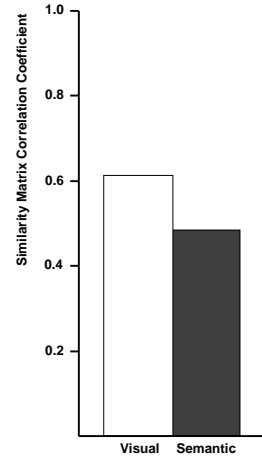
output procedures, the mixed visual-and-semantic error rates for all lesion locations are higher than expected based on the independent rates of visual and semantic errors. Thus the  network replicates the H&S results for the most part. In addition, there is an increasing bias towards semantic similarity in errors as lesions move towards semantics ($0 \Rightarrow I$: 1.30, $I \Rightarrow I$: 2.38, $I \Rightarrow C$: 2.52, $C \Rightarrow I$: 2.83, $I \Rightarrow S$: 5.02). The  network also shows a strong bias towards mixed visual-and semantic errors with damage to the attractors at the level of the intermediate units. A clearer explanation of this effect will be possible when we consider the similarity matrices for this network. Lesions to $I \Rightarrow S$ connections produce *very* few errors, with those that occur tending to be semantically related to the stimulus. These lesions do produce a significant number of errors, however, when an output network is used, due to the presence of phonological attractors.

Unlike networks with a strictly feed-forward direct pathway, the states of the intermediate units of the  network change throughout the processing of a word. This makes possible a richer comparison of the conversion from visual to semantic similarity at both the intermediate and semantic layers. Accordingly, we present two sets of similarity matrices for the  network. Figure 4.7 presents the matrices for the word representations (and their visual and semantic similarity correlations) at the intermediate layer of the  network at iterations 1, 3, and 5, and Figure 4.8 presents the same information for the semantic layer at iterations 2, 3, and 4. Considering the intermediate layer first, a comparison with the corresponding matrix for the intermediate layer of the  network (Figure 4.3, p. 80) reveals that the initial intermediate representations in the  network remain slightly more visually structured than in the  network. However, clean-up in the former network gradually adjusts these representations to eventually be slightly more semantically than visually organized. Because an additional set of connections ($I \Rightarrow S$) map these representations onto semantics, they need not become completely semantically organized at the intermediate layer. Thus, intermediate-level clean-up allows the representations to find a more even balance between visual and semantic similarity at the intermediate layer. In contrast, networks with only feed-forward direct pathways remain much more visually organized at the intermediate layer, and leave to the semantic-level clean-up more of the work of converting to semantic similarity.

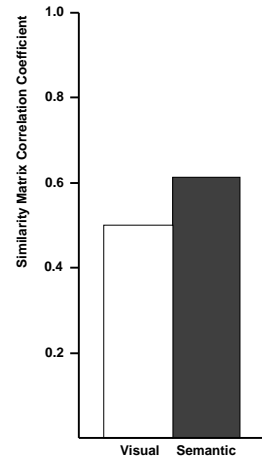
The high rate of mixed visual-and-semantic errors produced by lesions at the level of the intermediate layer in the  network (compared with, for instance, the  network) can be understood as follows. Compromising between visual and semantic similarity is least problematic for words that can produce mixed errors because the mapping between orthography and semantics is least arbitrary for these words—at least one visually similar word does map to a similar semantic representation. In a network that must accomplish the mapping via one intermediate representation, this shared visual and semantic structure makes it easiest to derive adequate representations for these words. Most words must rely on a clean-up mechanism that separates visually similar words into quite different representations. Under damage, attractor basins for



Intermediate Layer Representations
Iteration 1



Intermediate Layer Representations
Iteration 3



Intermediate Layer Representations
Iteration 5

Figure 4.7: Similarity matrices and their correlation coefficients with matrices for visual (orthographic) and semantic similarity, for representations at the intermediate layer of the  network at iterations 1, 3, and 5.

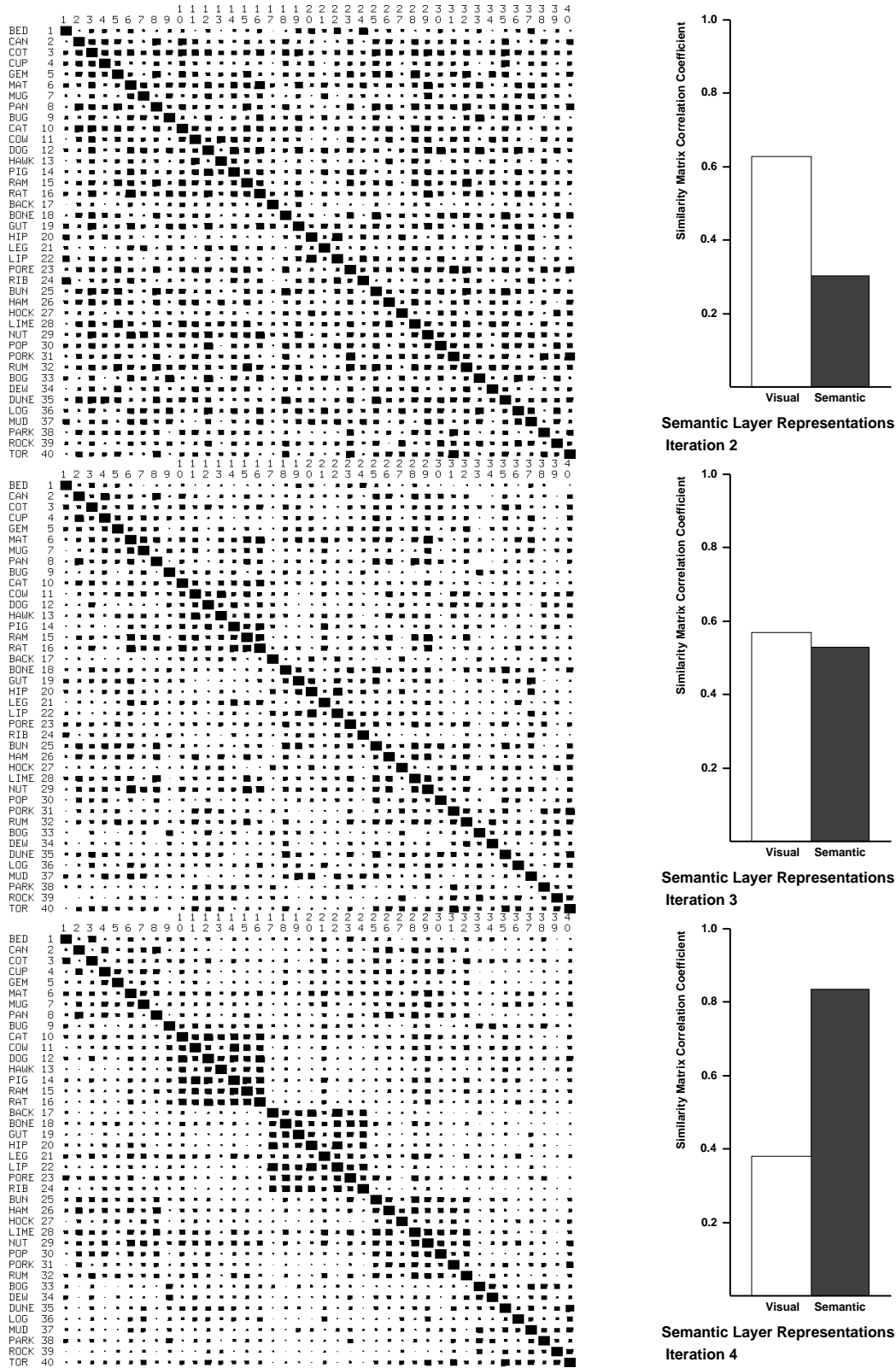
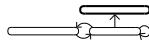
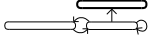
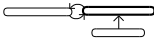
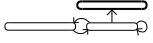
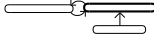
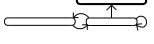
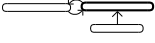
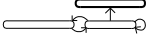
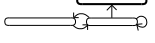
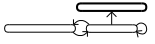



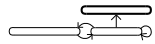
Figure 4.8: Similarity matrices and their correlation coefficients with matrices for visual (orthographic) and semantic similarity, for representations at the semantic layer of the  network at iterations 2, 3, and 4.

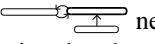
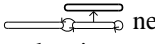
words that can lead to mixed errors tend to survive and “grow,” and are often produced as error responses. Another way of viewing the high rate of mixed errors is that, in the  network, visual and semantic similarity must operate at the same location in the network, making words that are consistent with both of these metrics most stable. In contrast, in the  network visual and semantic influences are more separated—visual similarity operates almost exclusively in the direct pathway while semantic influences occur primarily within the (semantic) clean-up pathway. This explains why the rate of mixed visual-and-semantic errors in this network is no more than predicted by the rates of visual and semantic errors occurring independently.

Comparing the similarity matrices for the semantic layer of the  network (Figure 4.8) with those for the  network (Figure 4.3, p. 80), notice that at iteration 2 the representations in the  network (like those at the intermediate layer) remain more visually organized. In fact, even by iteration 4 the representations have not become as semantically organized as in the  network, although they do by the next iteration (not shown). Thus the conversion from visual to semantic similarity at the semantic layer is more gradual in the  network. To give a more detailed sense of the gradual conversion to semantic similarity at the intermediate and semantic layers of the  network, Figure 4.9 shows the progression of each correlation coefficient across all eight iterations.

Overall, the results for the  suggest that *any* attractors that mediate in mapping between orthography and semantics will show influences of these metrics in the errors produced under damage. In addition, very few errors are produced *without* attractors, although correct performance can be reasonable.⁷ Together with the replication of the mixture of error types for lesions at or prior to the level at which attractors operate, these results clearly demonstrate the generality of the H&S results.

4.4.4 The network

H&S argue that a network must have hidden units to map from orthography to semantics, and it must form attractors to reproduce the error pattern of deep dyslexic reading. They used two separate groups of hidden units for these purposes, each carrying out a different function: the intermediate units generate an initial semantic pattern from orthography, and the clean-up units refine this into the correct semantics of the appropriate word. However, it is possible to eliminate the clean-up units and still allow the network to form attractors by introducing *feedback* connections from semantics to the intermediate units. The resulting  architecture will allow us to evaluate the impact of having a separate clean-up pathway that is not directly involved in mapping from orthography. The network is similar to the  network in that a single intermediate representation must

⁷In fact, unlike in the  network, correct performance of the  network is most impaired by lesions to the $I \rightarrow S$, suggesting that the existence of attractors is important for cleaning-up the semantics in producing correct responses as well as errors.

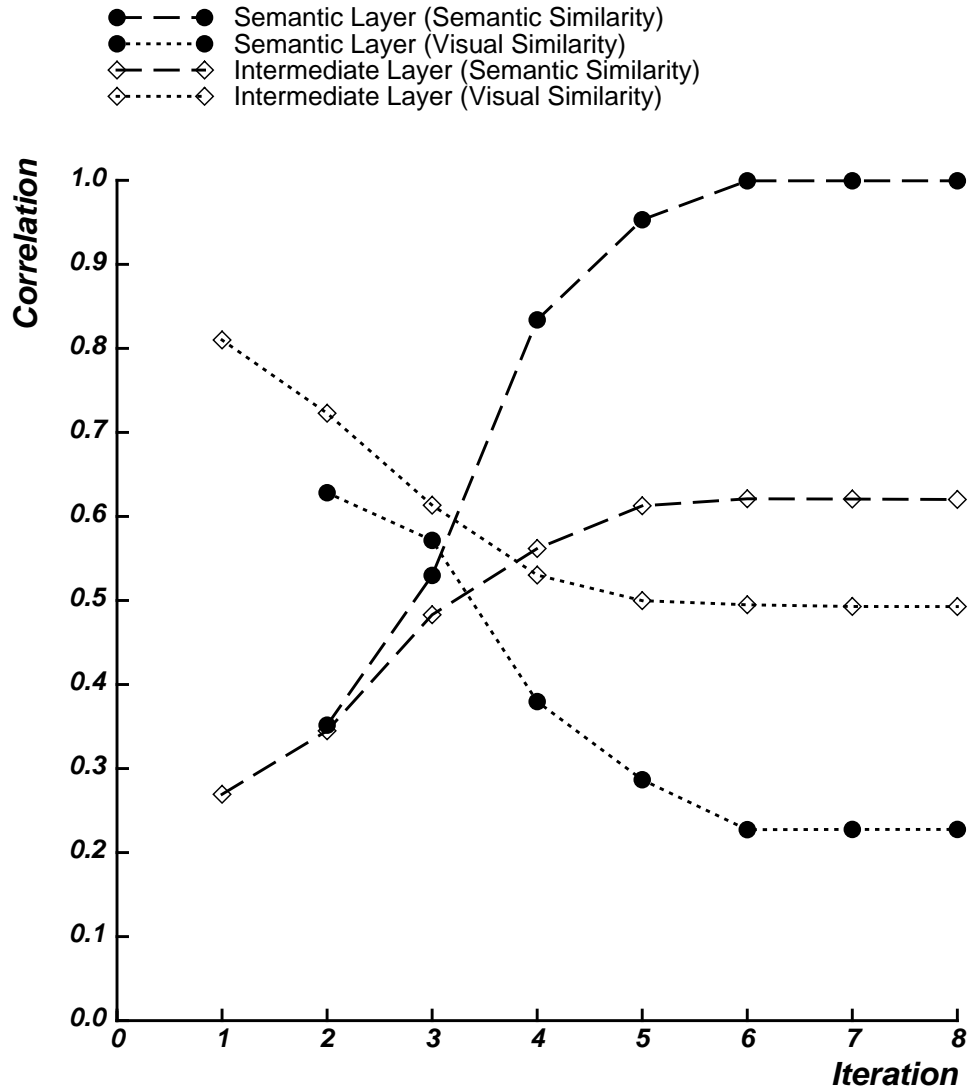
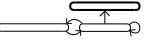



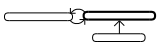




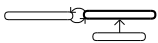








Figure 4.9: The correlation coefficients with the visual and semantic similarity matrices for the similarities among word representations at the intermediate and semantic layers of the  network for each iteration. Curves for the semantic layer begin at iteration 2 because this is when the layer first receives input from the direct pathway. Notice that a semantic similarity correlation of 1.0 for the semantic units over the last three iterations reflects the fact that the network is performing the task accurately.

mediate between orthography and semantics, but it differs by having no separate clean-up units. It is also similar to the framework for mapping among orthography, phonology, and semantics proposed by Seidenberg & McClelland (1989), and so will provide insight into how an implemented recurrent version of their system might behave under damage.

Figure 4.10 presents the lesion data for the  network. The overall correct performance of the  network under damage to the direct pathway is quite similar to that of the  network, both being more sensitive to this type of damage than the  network. In addition, the  network is like the  network (and unlike the  network) in that lesions closer to semantics ($I \Rightarrow S$) are more disruptive than more distant lesions ($O \Rightarrow I$). Lesions to the feedback connections ($S \Rightarrow I$) are much less debilitating than are lesions to any of the clean-up connection sets in the  network, and are about equally debilitating as lesions to $C \Rightarrow S$ in the  network.

All three networks show a quite similar rate and distribution of errors for $O \Rightarrow I$ lesions, except that the  network shows a somewhat lower rate of “other” errors. However, unlike the  network, the  network produces significant error rates for $I \Rightarrow S$ lesions, with the errors tending to be semantically related to the stimulus. Lesions to the $S \Rightarrow I$ feedback connections produce very few errors, with a high proportion being mixed visual-and-semantic. This is quite similar to the effect of $C \Rightarrow S$ lesions in the  network, and for clean-up lesions in the  network. For both output procedures and each lesion location, the rates of each error type (relative to that of “other” errors) is at least 2.5 times greater than for errors chosen at random (except for visual errors with $S \Rightarrow I$ lesions), and the rate of mixed visual-and-semantic errors is higher than expected from the independent rates of visual errors and semantic errors. In fact, like in the  network the relative rates of mixed errors is quite high for lesions to connections involved in building attractors ($I \Rightarrow S$ and $S \Rightarrow I$). The biases towards semantic similarity in errors for each lesion location are $O \Rightarrow I$: 1.13, $I \Rightarrow S$: 2.78, $S \Rightarrow I$: 2.92.







The similarity matrices the over iterations 1, 3, and 5 for the intermediate layer of the  network, shown in Figure 4.11, are quite similar to those of the  network. The intermediate representations in the  network ultimately become more semantically organized because they interact directly with the semantic layer, where representations become *completely* semantically organized. In this way, feedback connection from semantics make the task of the $I \Rightarrow S$ connections somewhat easier.

Figure 4.12 shows the similarity matrices for representations at the semantic layer of the  network for iterations 2 and 4.⁸ Comparing the matrix for iteration 2 with the corresponding matrix for the  network (top of Figure 4.8, p. 88), notice that the semantic activity in the  network is much more semantically organized before the influence of any clean-up. This is

⁸The matrix for iteration 3 is not displayed because it is essentially identical to that for iteration 2. Intermediate units are not influenced by semantic units until iteration 3 and so they cannot have new influences on semantic units until iteration 4.

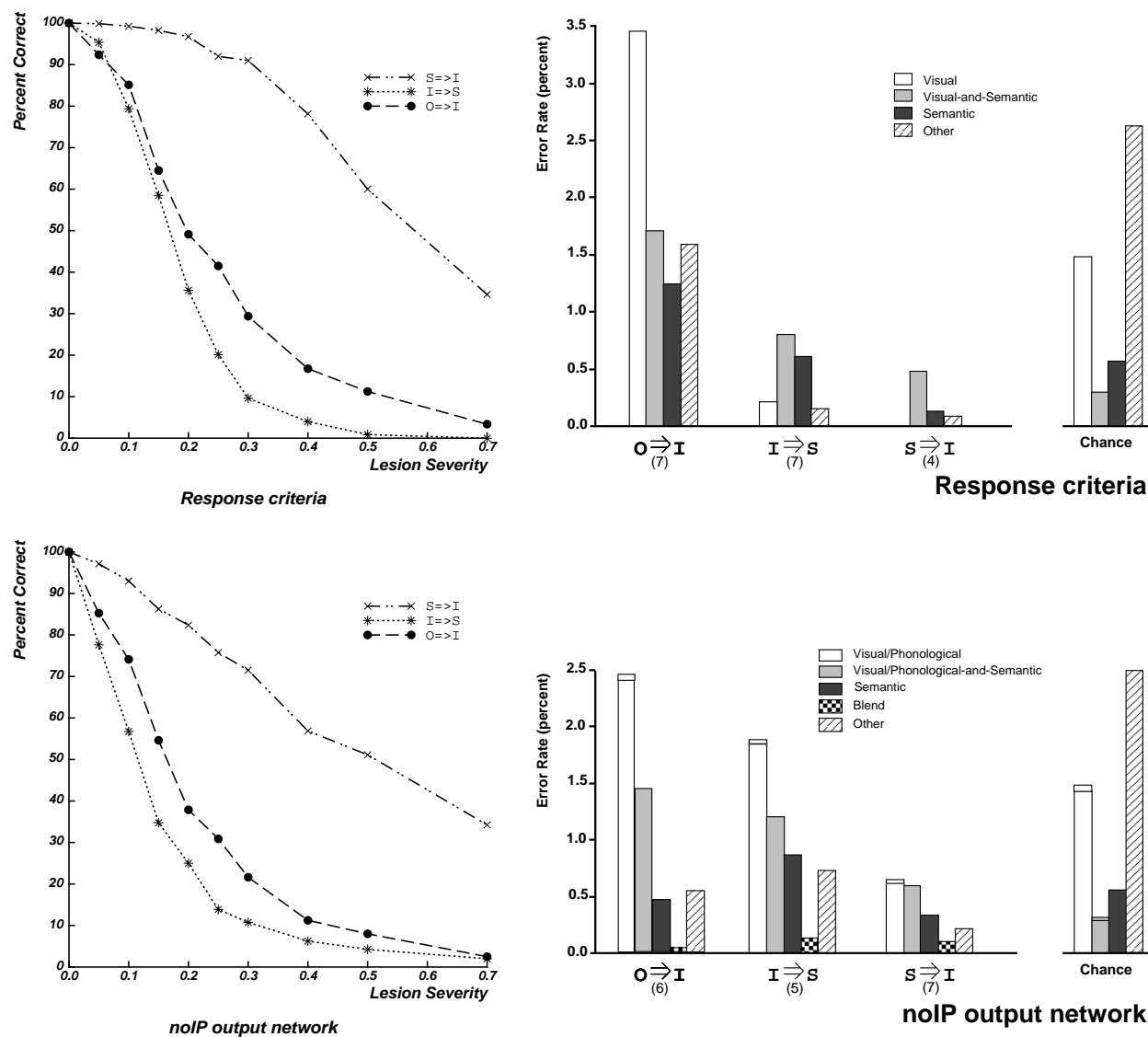



Figure 4.10: Overall correct performance and error distributions for the  network using the response criteria and the output network without intra-phoneme connections.

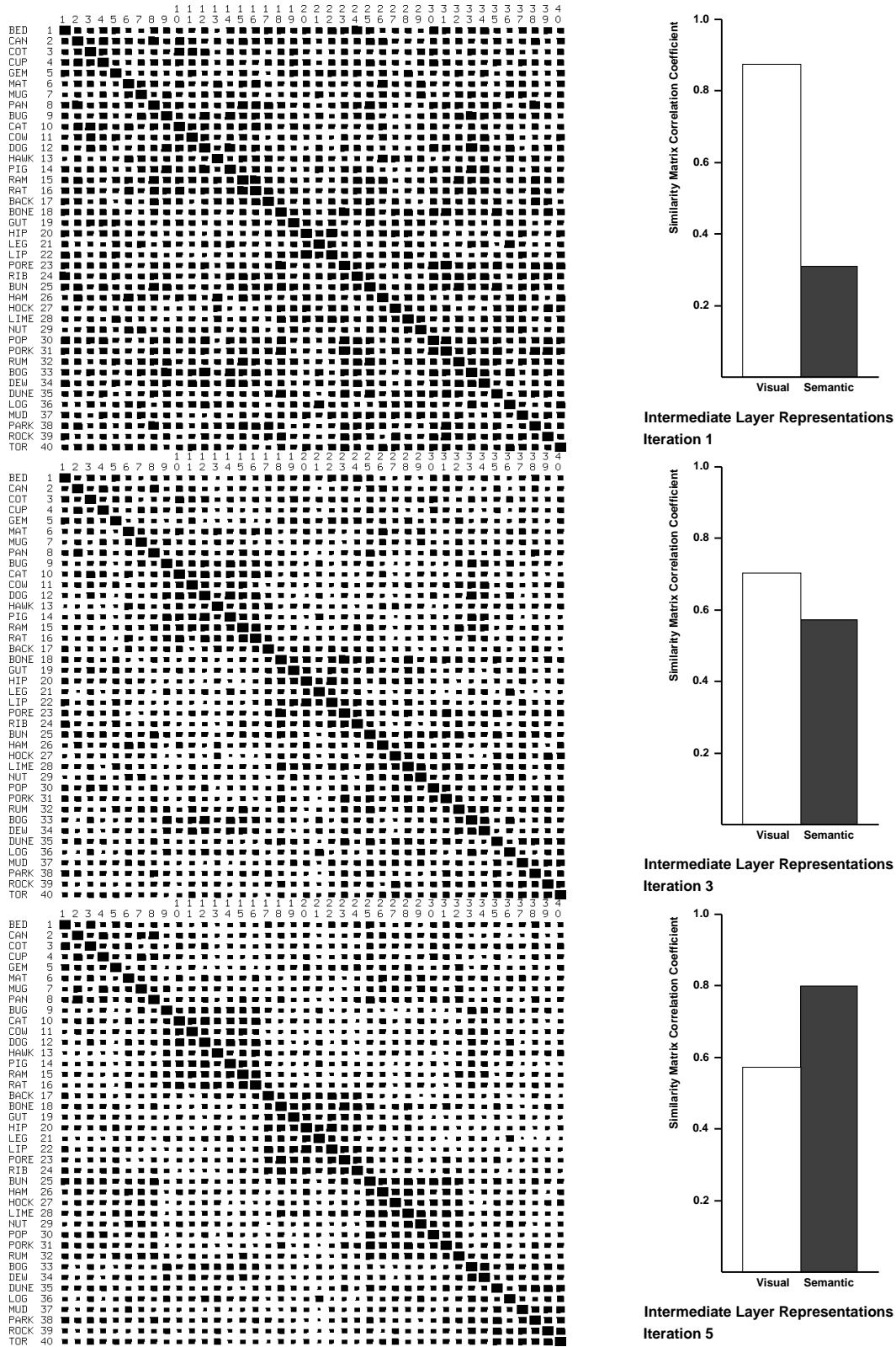



Figure 4.11: Similarity matrices and their correlation coefficients with matrices for visual (orthographic) and semantic similarity, for representations at the intermediate layer of the  network at iterations 1, 3, and 5.

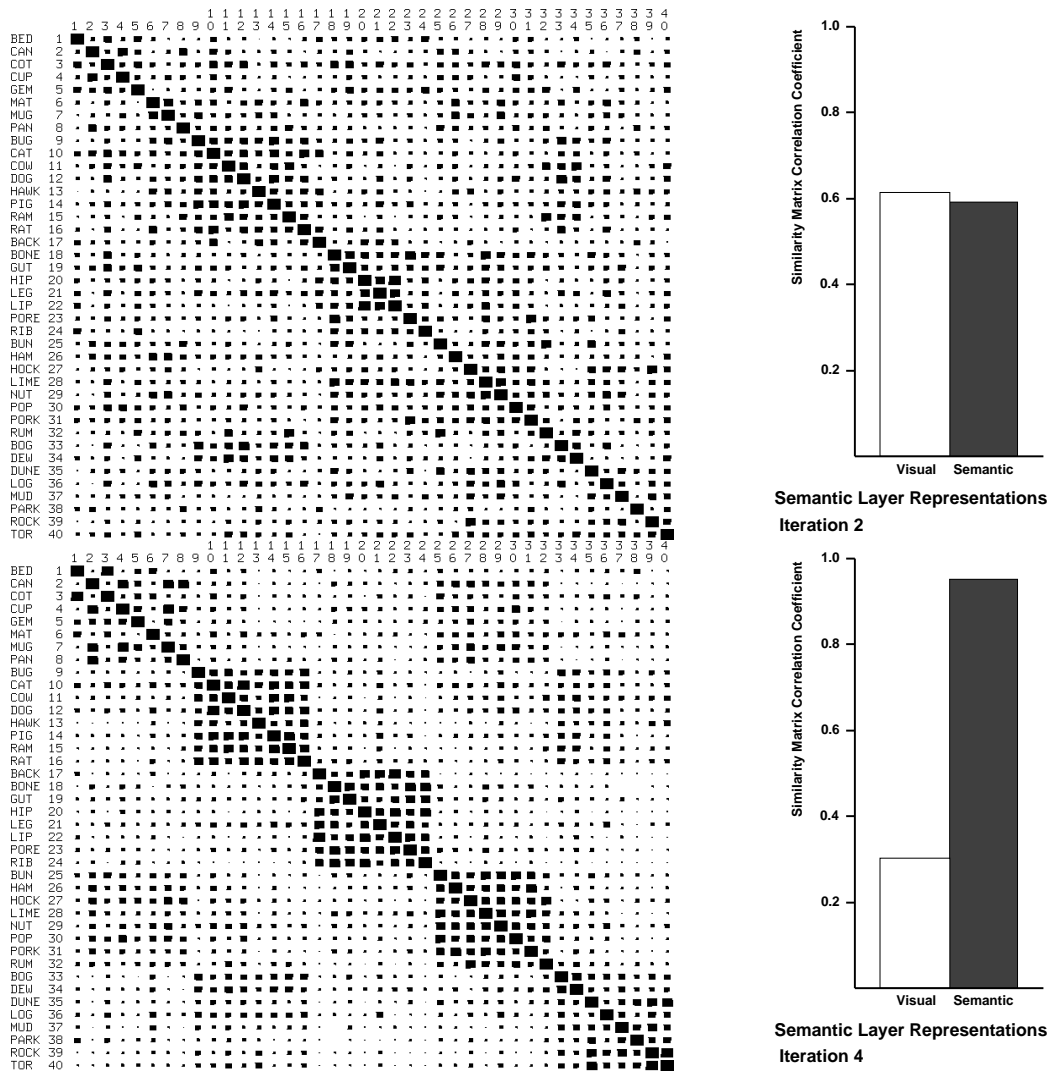


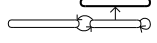
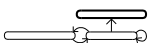



Figure 4.12: Similarity matrices and their correlation coefficients with matrices for visual (orthographic) and semantic similarity, for representations at the semantic layer of the  network at iterations 2 and 4.

true even though the intermediate representations providing input in the two networks are equally semantically organized. There are twice as many $I \Rightarrow S$ connections in the  network than in the  network (because it has twice as many intermediate units) and so intermediate representations can have more effective influence on semantics in the former network. In fact, by iteration 4 the semantic representations in this network are more accurate than at the same point in the  network, presumably because the intermediate representations are more semantically organized at the preceding iteration. Thus the semantic units serve as more effective clean-up than a separate group of clean-up units attached to the intermediate units that must *learn* to have semantic influences.

Overall, the success of the  network in replicating the H&S results demonstrates that those results do not depend on having a separate set of clean-up units to perform semantic micro-inferences. Intermediate units can learn both to convey information about orthography and to interact with semantics to form attractors for word meanings. However, using intermediate units in this way has implications for the distribution of error types—in particular, the rates of mixed visual-and-semantic errors.

4.4.5 The network

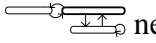
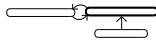

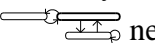
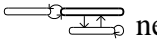
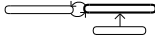
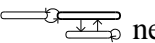
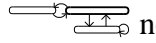
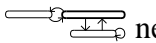
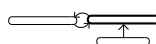
The final architecture we will consider, the  network, is a hybrid between the  and  networks. The results with the latter network demonstrated how feedback from semantics can help the intermediate layer representations become more semantically organized. Our intention in investigating the  network is to determine whether this effect, in combination with a separate set of semantic clean-up units, would produce stronger semantic attractors.

Figure 4.13 presents the lesion data for the  network. Overall correct performance after lesions to the direct pathway is much like that in the  network, with $O \Rightarrow I$ lesions being more debilitating than $I \Rightarrow S$ lesions. Interestingly, lesions to the clean-up pathway are much less disruptive in the  network. In addition, even severe lesions to the feedback connections produce a negligible deficit. This pattern suggests that the redundancy in the type of clean-up available to the network makes it more robust to this type of damage.

In terms of distributions of error types, the two networks produce quite similar patterns of errors with $O \Rightarrow I$ lesions, although the error rates of the  network are slightly higher. In contrast, $I \Rightarrow S$ lesions produce fewer errors in the  network, but these are more likely to be semantically related to the stimulus than in the  network. Both networks produce very few errors after lesions to the clean-up pathway. The semantic biases in errors for lesion locations producing a reasonable number of errors are $O \Rightarrow I$: 1.16, $I \Rightarrow S$: 3.74, $C \Rightarrow S$: 6.69. The rates of all error types relative to “other” errors are above chance for all lesion locations, except for $S \Rightarrow I$ lesions and for visual errors from $C \Rightarrow S$ lesions. Also, mixed visual-and-semantic error rates are higher than predicted by visual and semantic error rates occurring independently at each lesion

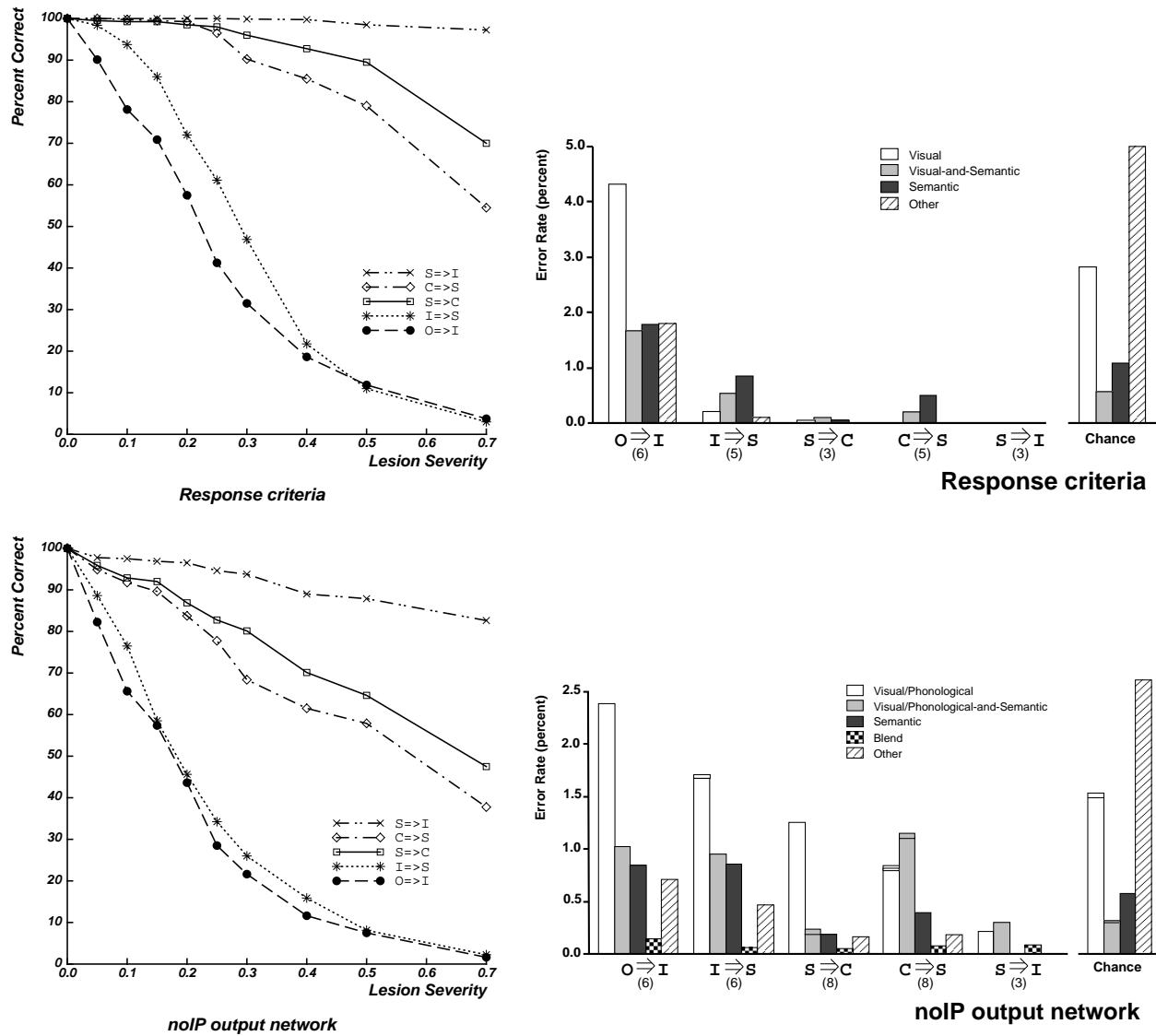
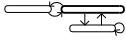
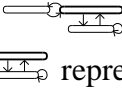

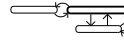
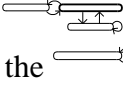
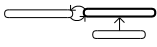
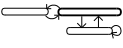
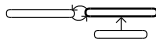
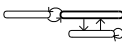
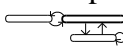
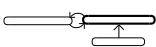
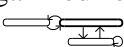
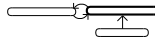
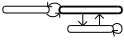
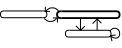
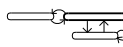


Figure 4.13: Overall correct performance and error distributions for the  network using the response criteria and the output network without intra-phoneme connections.

location except for $C \Rightarrow S$ using the response criteria, and for $S \Rightarrow C$ using the output network.

Figure 4.14 shows the similarity matrices for the intermediate layer representations at iterations 1, 3, and 5 in the  network. Comparing with those for the  network (Figure 4.11, p. 93), the  representations are less extreme—starting out less visually organized and ending up less semantically organized. This may reflect the reduced expressiveness of 40 compared with 80 intermediate units. However, the general tendency of feedback connections from semantics to induce an increasingly semantic organization at the intermediate layer is similar in the two networks.

Now consider the similarity matrices for the semantic layer representations at iterations 2, 3, and 4 in the  network (shown in Figure 4.15), and compare them with the corresponding matrices for the  network (Figure 4.3, p. 80). Interestingly, at iteration 2 the representations in the  network are less semantically organized, even though the network has the same number of intermediate units and connectivity density as the  network. Clearly the existence of additional sets of connections has caused the knowledge that produces semantic organization to be distributed somewhat differently in the  network. In particular, the impact of the clean-up units appears to be less—there is less of a difference between iterations 3 and 4 in the  network than in the  network (where iterations 2 and 3 are essentially identical). The task of inducing semantically organized representations is split more evenly between the clean-up and intermediate units in the  network. This process is also more gradual than in the  network—at iteration 4 the semantic representations are further from their final organization in the  network.

Overall, the  network, which includes aspects of a number of the other networks, behaves in ways that are similar to each of them. Feedback to the intermediate units from semantics helps these units become more semantically organized. Additional clean-up units provides a redundant mechanism for implementing semantic attractors, making the network somewhat more robust to damage. However, there seems to be little difference in the strength of the attractors that the  network develops. In this sense, the strength of attractors developed by a network may be more a function of the demands of the task and training procedure (e.g. if noise is added to the input) than of the architecture *per se*.

4.5 Summary of architecture comparisons

4.5.1 Generality of the H&S findings

There are a number of general conclusions that can be drawn from the properties of this set of networks. The most important findings are those that concern the generality of the theoretically critical results obtained by H&S. These fall into two parts. H&S's main conclusion was that all types of error—visual, semantic, and mixed—occur with all locations of lesions. With one or two

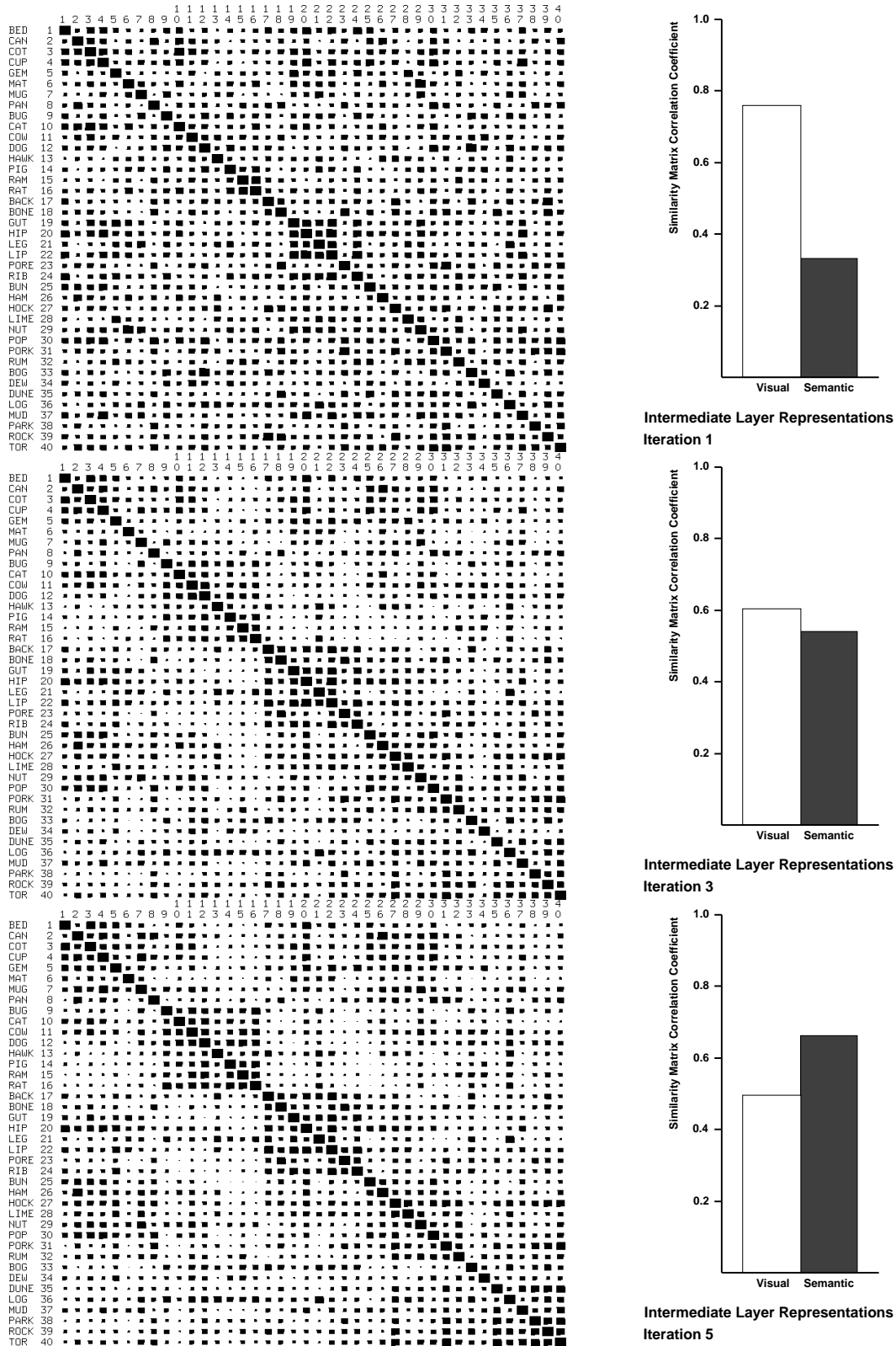
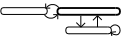


Figure 4.14: Similarity matrices and their correlation coefficients with matrices for visual (orthographic) and semantic similarity, for representations at the intermediate layer of the  network at iterations 1, 3, and 5.

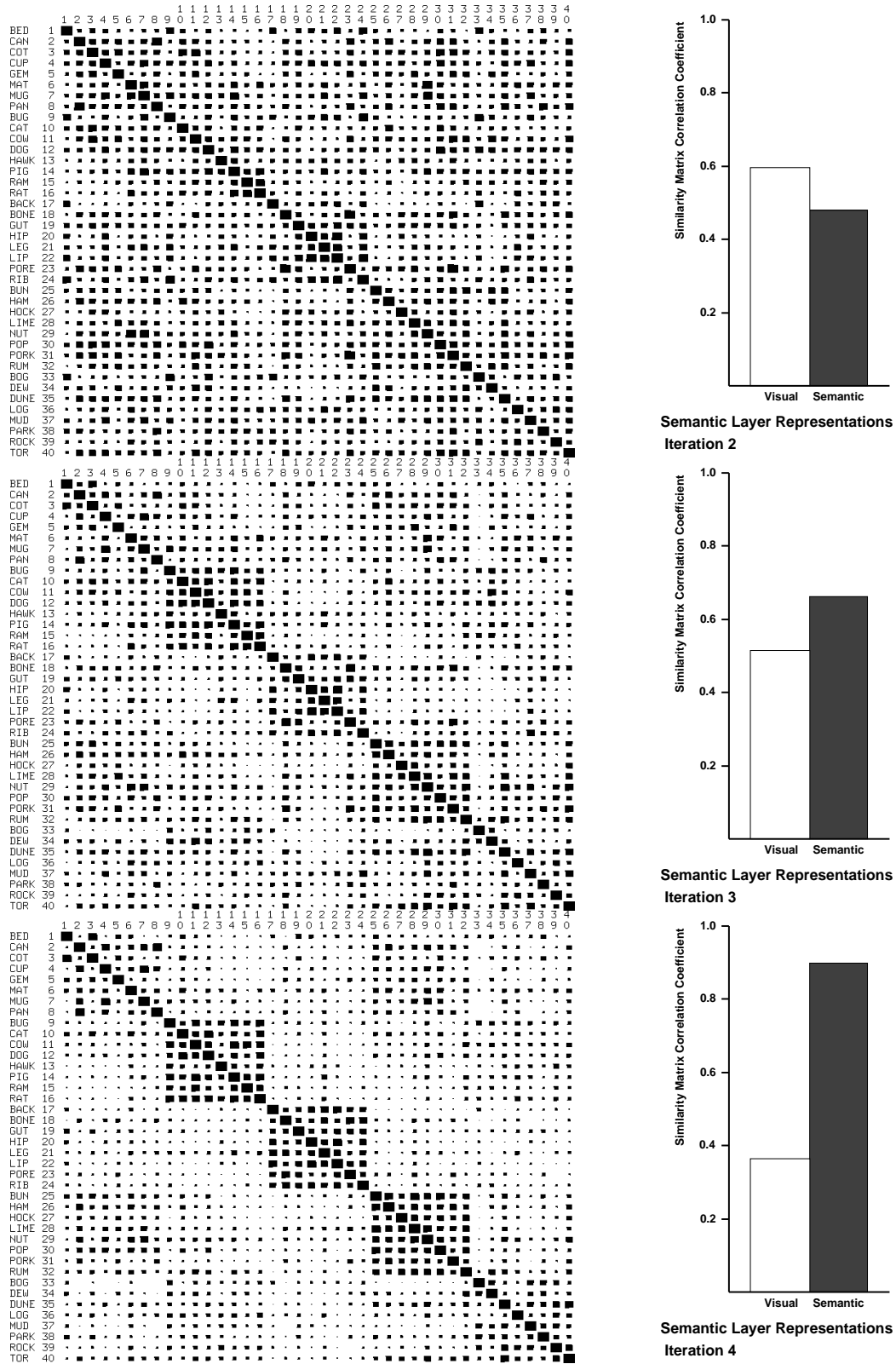
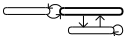
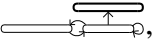

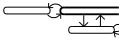
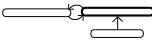

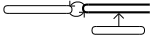
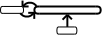


Figure 4.15: Similarity matrices and their correlation coefficients with matrices for visual (orthographic) and semantic similarity, for representations at the semantic layer of the  network at iterations 2, 3, and 4.

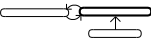
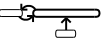
minor exceptions, concerning lesion sites that give rise to very low absolute error rates, this finding generalizes to all the other networks examined) as well as to lesions to output connections ($S \Rightarrow I_p$ and $I_p \Rightarrow P$, see Section 3.4).




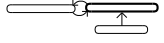

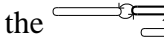
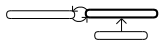
A second finding of H&S was that mixed visual-and-semantic errors occur more frequently than one would expect given the independent rates of visual errors and of semantic errors. This finding appears to be less general than the simple co-occurrence of error types. The replication of the H&S network, using the original input representation and trained without noise, also exhibits higher than expected mixed rates (except when using the IP output network, or for lesions to an output pathway—see Chapter 3). However, among networks using the distributed letter representations and trained with noise, the effect is only found when the intermediate units are directly involved in developing attractors—the , , and  networks, but not the  and  networks.

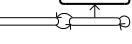

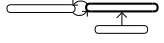
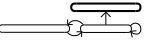

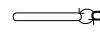
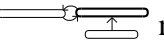
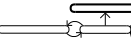
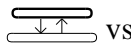
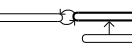
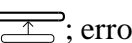
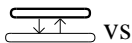
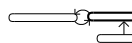
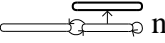
Why might these effects occur? One possibility is that the  and  networks form strong semantic attractors using the clean-up pathway, so that maximum visual similarity effects occur at a considerably earlier stage of processing than maximum semantic similarity effects. Thus the transformation from visual to semantic similarity is realized through separable stages. The networks trained without noise form weaker semantic attractors using the clean-up units, so that more of the work of mapping visual to semantic similarity is carried out by the direct pathway. This compresses the stages over which visual and semantic similarity operate, and therefore makes interactions between them in the stimulus set—the potential for mixed errors—more critical. This is also true of the networks in which intermediate units are involved in implementing attractors. In these networks, the attractors lie at a stage where visual and semantic influences cannot be separated. It should be pointed out that this account is somewhat speculative—the main point is that the mixed error findings of H&S, while narrowly robust, do not generalize to all lesion sites of all connectionist networks. It is a consequence of particular characteristics of some network architectures.

4.5.2 The strength of attractors

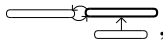

At a more general theoretical level, the argument that H&S put forward of the importance of attractors in the generation of errors is borne out. The robustness of a network to lesions of a set of connections, measured by the rate of correct performance, increases with the strength of the attractors at levels after the locus of damage. At the same time, the rates of explicit errors from lesions to these connections also rise. In essence, the attractors serve to clean-up both correct and incorrect responses, reducing the number of omissions caused by damage. In contrast, lesions at or beyond the level of the last attractors in a network produce a very low rate of overt responses, both correct and incorrect.


This effect can be seen by comparing the  network with the  network. Both

networks use the same input and output representations, were trained identically, and develop attractors at the semantic level. However, the overall correct performance and explicit error rates of the  network are higher than for the  network for both $0 \Rightarrow I$ and $I \Rightarrow S$ lesions, using both the response criteria and the noIP output network. The  network develops stronger attractors because its full connectivity between layers makes it more effective than the  network at implementing semantic micro-inferences that depend on the interaction of two or more semantic features on a third. The probability that the involved semantic features will be appropriately connected to some clean-up unit is 1.0 in the  network but quite small in the  network due to its 25% connectivity density. The replication of the H&S network, which it was argued above has weaker semantic attractors than the  network, is less robust overall to lesions of the direct pathway (although the balance between $0 \Rightarrow I$ and $I \Rightarrow S$ is reversed) and has lower explicit error rates.

For the  and  networks, correct and error rates are comparable to those of the  network for $0 \Rightarrow I$ lesions, which are before the level at which their attractors operate. However, consider lesions to $I \Rightarrow S$ connections, which are “post-attractor” for the  network, “within-attractor” for the  network, and “pre-attractor” for the  network. Both the correct and error rates are much lower (using the response criteria) for the first two networks than for the  network (e.g. $I \Rightarrow S(0.3)$, correct: 14.1%  and 9.6%  vs. 42.9% ; errors: 0.2%  and 1.8%  vs. 3.7% ).⁹ The very low error rate for the post-attractor $I \Rightarrow S$ lesions in the  network reinforces the arguments presented earlier that the occurrence of explicit errors depends on damaged input being cleaned-up into an incorrect attractor.

4.5.3 Error types

For all networks, error rates are much higher for $0 \Rightarrow I$ lesions than for $I \Rightarrow S$ ones, presumably because the output of the undamaged $I \Rightarrow S$ connections will be more likely to be closer to a word representation than will their damaged output. In addition, for the networks that have attractors only at the semantic level (H&S replication, , ), both the absolute and relative rates of visual errors drop sharply between $0 \Rightarrow I$ and $I \Rightarrow S$ lesions, and the absolute and relative rates of semantic errors climb—the absolute rise is a modest one and limited to the criteria conditions. This general trend is shown directly in the biases towards semantic vs. visual similarity in errors (as compared with word pairs chosen at random) as lesions move from “early” to “late” in the network. These findings are similar to those obtained by H&S and indicate that such networks can give rise to the quantitative differences in the distribution of error types found across deep dyslexic patients.

⁹Not surprisingly, the hybrid  network shows hybrid characteristics.

We now turn to an number of separate issues that concern more detailed aspects of the pattern of correct and impaired performance shown to varying degrees by all of these networks. These considerations serve both to verify that the general effects produced by the networks aren't due to idiosyncratic characteristics of the word set or interpretation procedure, and also to demonstrate that the networks behave like deep dyslexics in terms of the pattern of responses after individual lesions in addition to exhibiting a similar overall pattern of performance when averaged across lesions.

4.6 Item- and category-specific effects

The small size of the H&S word set raises the possibility that many of the effects arise from idiosyncratic characteristics of the word set itself, and not to any real systematic relationship between orthography and semantics. In particular, it is possible that only a handful of words account for most of the errors. In this section we address the extent that the effects we have demonstrated are distributed across the entire word set.

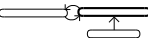
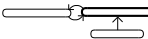
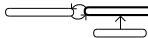

Considering correct performance first, Figure 4.16 presents the overall correct rates for individual words after lesions to the  network with using both the criteria and noIP output network to generate responses. Although there is a reasonable amount of variability among words, it is not the case that some words are always impaired or intact regardless of the type of damage. The pattern of overall correct performance is somewhat different depending on how output is generated, although the correlation between the correct rates using the response criteria and those using the noIP output network is moderate but significant ($0.47, p < .005$).

Figure 4.16 also suggests that there may be some systematic differences in correct performance across categories. H&S found that words in the “foods” category were selectively spared after $C \Rightarrow S(0.4)$ lesions in one version of their network. Figure 4.17 presents the correct performance of the  network for words in each category, separated by lesion location. With words as the random variable, there is a significant main effect of category for lesions of $0 \Rightarrow I$ ($F(4, 35) = 5.91, p < .001$), $I \Rightarrow S$ ($F(4, 35) = 3.94, p < .01$), and $C \Rightarrow S$ ($F(4, 35) = 3.35, p < .02$), but not for $S \Rightarrow C$ lesions ($F(4, 35) = 1.06$). The figure shows that “body parts” and “outdoor objects” are selectively impaired by $0 \Rightarrow I$ lesions, while “animals” are selectively preserved by $I \Rightarrow S$ lesions. $C \Rightarrow S$ lesions produce a selective deficit for “outdoor objects.” While there is certainly evidence that certain sets of connections are differentially important for reading certain categories of words, none of the selective category effects in the  network is as pronounced as that found by H&S.

However, particular lesions in some networks can produce quite dramatic category effects that are even more pronounced than those observed by H&S (see Figure 4.18). For example, $C \Rightarrow S(0.7)$ lesions in the  network produce a striking selective preservation of “animals”

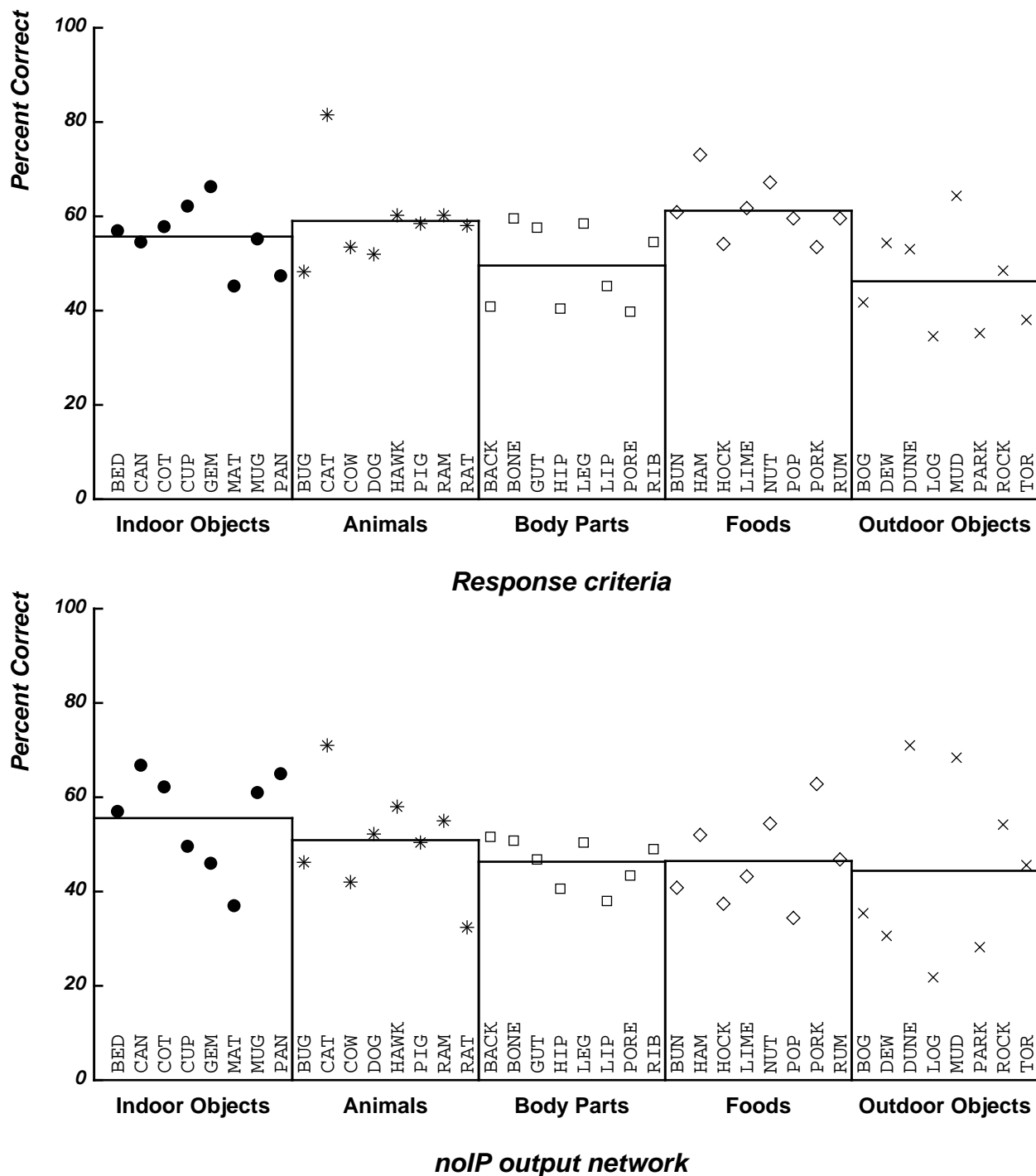
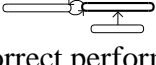


Figure 4.16: The correct performance rates of the  network for each word, averaged over all lesion locations and severities leading to correct performance between 20–80%, using the response criteria and the noIP output network.

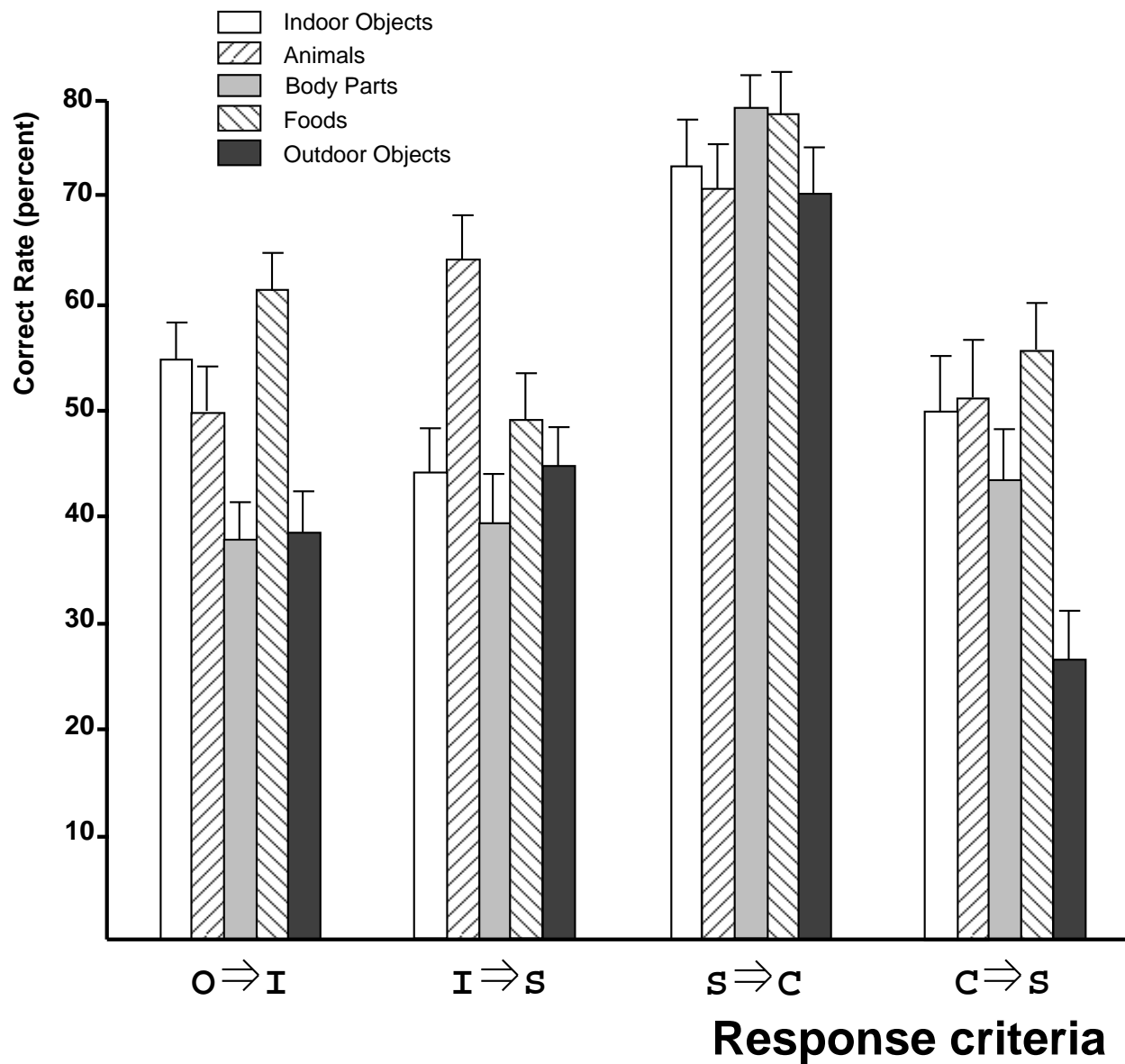
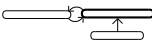


Figure 4.17: The effect of lesion location on the correct performance rates of the  network, using the response criteria, for words in each category. Error bars indicate one standard error from the mean.

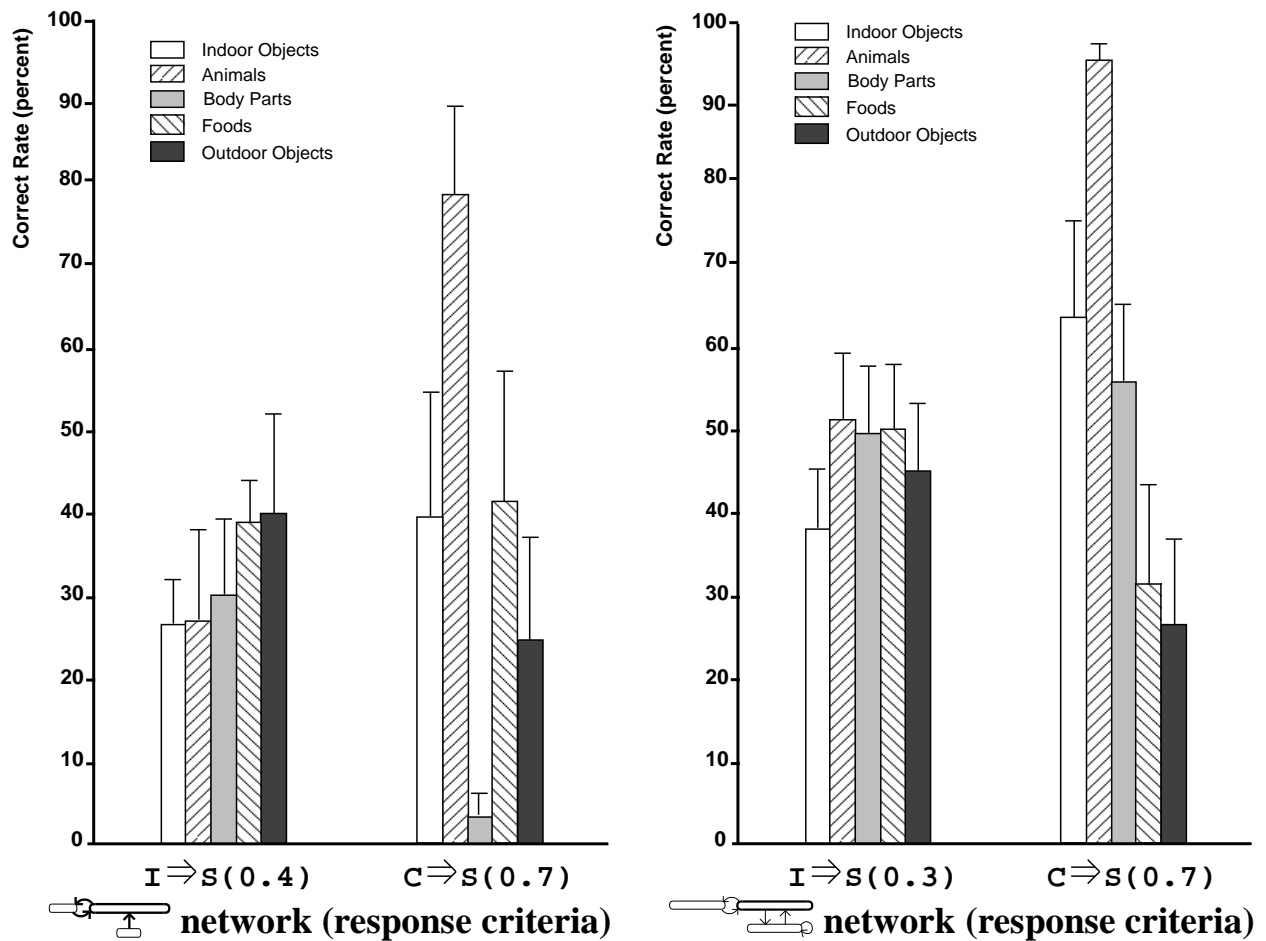




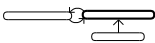
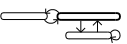
Figure 4.18: Correct performance for words in each category after lesions of $I \Rightarrow S(0.4)$ and $C \Rightarrow S(0.7)$ in the  network (left), and $I \Rightarrow S(0.3)$ and $C \Rightarrow S(0.7)$ in the  network (right).



Figure 4.19: Visual error rates the  network for each word, averaged over all lesion locations and severities leading to correct performance between 20–80%, using the response criteria.

and selective impairment of “body parts” relative to the other categories, as well as relative to other lesions yielding similar overall correct performance, such as $I \Rightarrow S(0.4)$. Interestingly, the  network also shows a selective preservation of “animals” with $C \Rightarrow S(0.7)$ lesions, but now “foods” and “outdoor objects” rather than “body parts” are selectively impaired. The nature of the selective deficits observed after damage appears to have as much to do with the particular characteristics of individual networks as with the relationships among semantic representations. In fact, the selective preservation of “foods” found by H&S did not arise in a second network that only differed from the first in its initial random weights—a type of variation typically not considered important (but see Kolen & Pollack, 1991). Clearly more research is required to understand these effects.

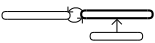
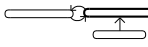
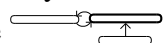
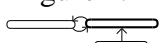
Turning to a consideration of item effects in error responses, we can also examine the distribution of each error type across words in the set. Figure 4.19 presents the rates of visual errors for each word produced by lesions to the  network using the criteria to generate responses. Only four of the words, BED, FIG, RAT, and HIP, produce no visual errors for any of the lesions. For the rest of the words there is a wide range of rates, with the highest being for COT and PORE, both having about four times the average rate. Visual errors are not arising due to only a few words but are distributed throughout the word set. In fact, there is a significant correlation ($0.49, p < .005$) between the observed visual error rates and the expected rates given the distribution of visual



Figure 4.20: Semantic error rates the  network for each word, averaged over all lesion locations and severities leading to correct performance between 20–80%, using the response criteria.

similarity throughout the word set. Thus the distribution of visual errors across words is relatively unbiased with respect to visual similarity.

Semantic errors are somewhat less uniformly distributed. Figure 4.20 presents the rates of semantic errors for each word produced by the  network. Nine of the words produce no semantic errors, while DOG produces almost twice as many as the word with the next highest rate, GEM. “Outdoor objects” have a uniformly low rate of semantic errors, while the rates for “body parts” are relatively high and distributed throughout the category. While the seven words with the highest rates account for 56% of the semantic errors, the remaining errors are spread across most of the remaining words. The correlation of the distribution semantic errors with that expected from the semantic similarity of the word set is marginally significant (0.30, $p < .06$).

In contrast, the network shows a strong bias to produce mixed errors for particular pairs of words. Figure 4.21 presents the rates of mixed visual-and-semantic errors for each word produced by the  network. Almost half (18) of the words do not produce any mixed errors. Of the remaining words, the top three (PAN, HIP, and LIP) account for 45% of the errors; the top six, over 65%. There is virtually no correlation (0.09) between the distribution of mixed errors across words and the distribution of visual-and-semantic similarity.

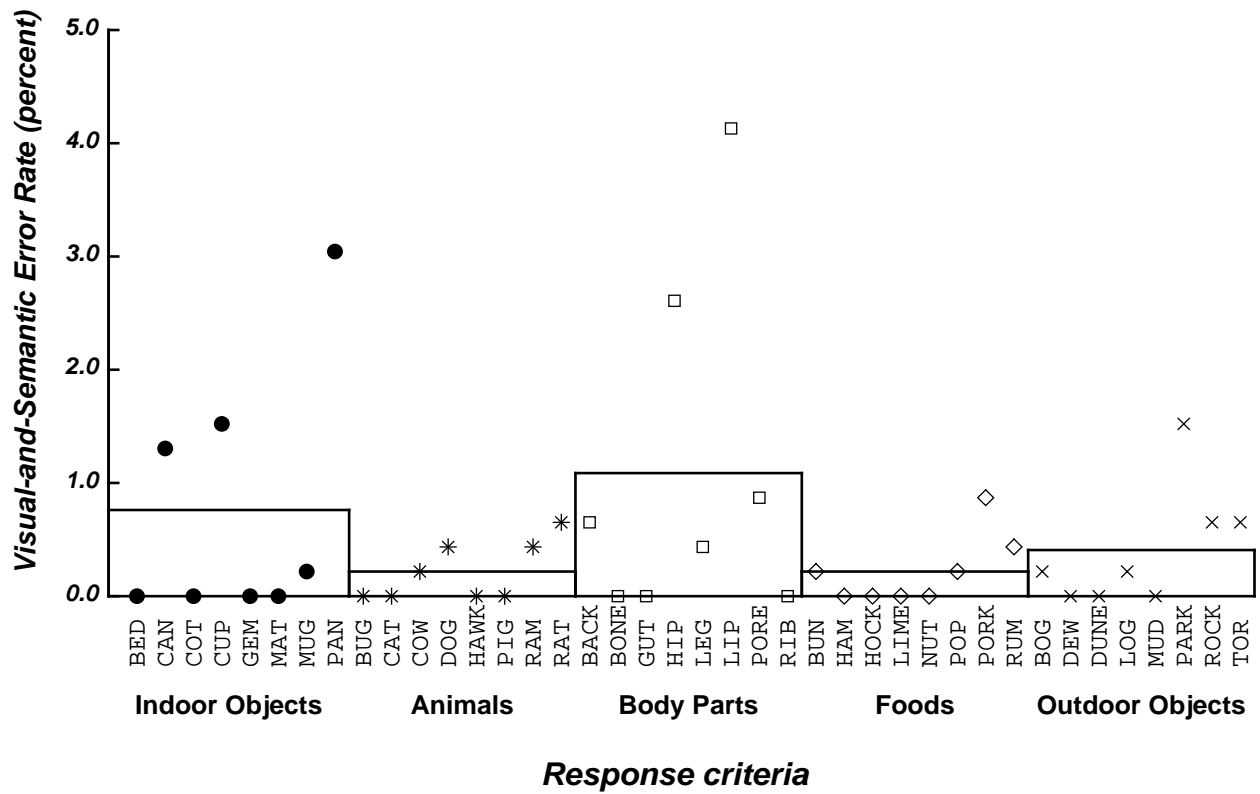
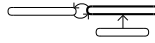


Figure 4.21: Mixed visual-and-semantic error rates the  network for each word, averaged over all lesion locations and severities leading to correct performance between 20–80%, using the response criteria.

To provide a more complete picture of the nature of error responses in the network, Figure 4.22 displays the frequency of each possible error in the form of a “confusion matrix.” Semantic and mixed visual-and-semantic errors consist of squares in one of the 8-by-8 blocks along the main diagonal. Thus the most common semantically related errors among “indoor objects” are PAN \Rightarrow “can” and BED \Rightarrow “cot”. In general, common errors are not symmetric, although the frequencies of HIP \Rightarrow “lip” and LIP \Rightarrow “hip” are about equal. Visual and other errors are represented in other regions of the matrix. Interestingly, the two most common incorrect responses to BOG are CAT and RAT—these appear to be visual-then-semantic errors via DOG. We will consider visual-then-semantic errors more extensively below.

Overall, the variation of the rates of various types of errors across words demonstrates that the effects in error patterns produced under damage do not arise from idiosyncratic characteristics of a few words. A possible exception is the mixed visual-and-semantic errors—the one theoretically important topic where the original H&S findings did not generalize consistently. However, the considerable degree of variability of error types across categories raises a concern about the use of these categories in defining semantic similarity. In the next section we address this issue directly.

4.7 Definitions of visual and semantic similarity

Following H&S, and as described in Section 2.6.4, we have considered a pair of words to be visually similar if they overlap in at least one letter, and semantically similar if they come from the same category. These definitions are intended to approximate the criteria used in categorizing the reading responses of patients. However, they are at best only coarse approximations. Our definition of visual similarity is somewhat more lax than that used for patients, where typically a stimulus and response must share at least 50% of their letters to be considered a visual error (Morton & Patterson, 1980). It is more difficult to compare the definitions of semantic similarity as there is no accepted formal measure of semantic relatedness that can be applied to patient responses—it is somewhat a matter of opinion whether a particular incorrect response should be considered a semantic error. Nonetheless, category membership is as good an approximation to the informal criteria of semantic relatedness used with patients as the limited word set used in the simulations allows.

However, it is important to realize that letter overlap and category membership, while providing intuitive groupings of words, only approximate the actual similarities among the orthographic and semantic representations used in the simulations. This was seen in the matrix of proximities among semantic representations (Figure 2.7, p. 36). Words like CUP and MUG are “indoor objects” and so are considered semantically related to words like BED and MAT, even though their semantic representations are actually much more similar to those of most “foods.” Similarly, some word pairs that overlap in a letter have quite different representations (e.g. LIME and RAM) while others

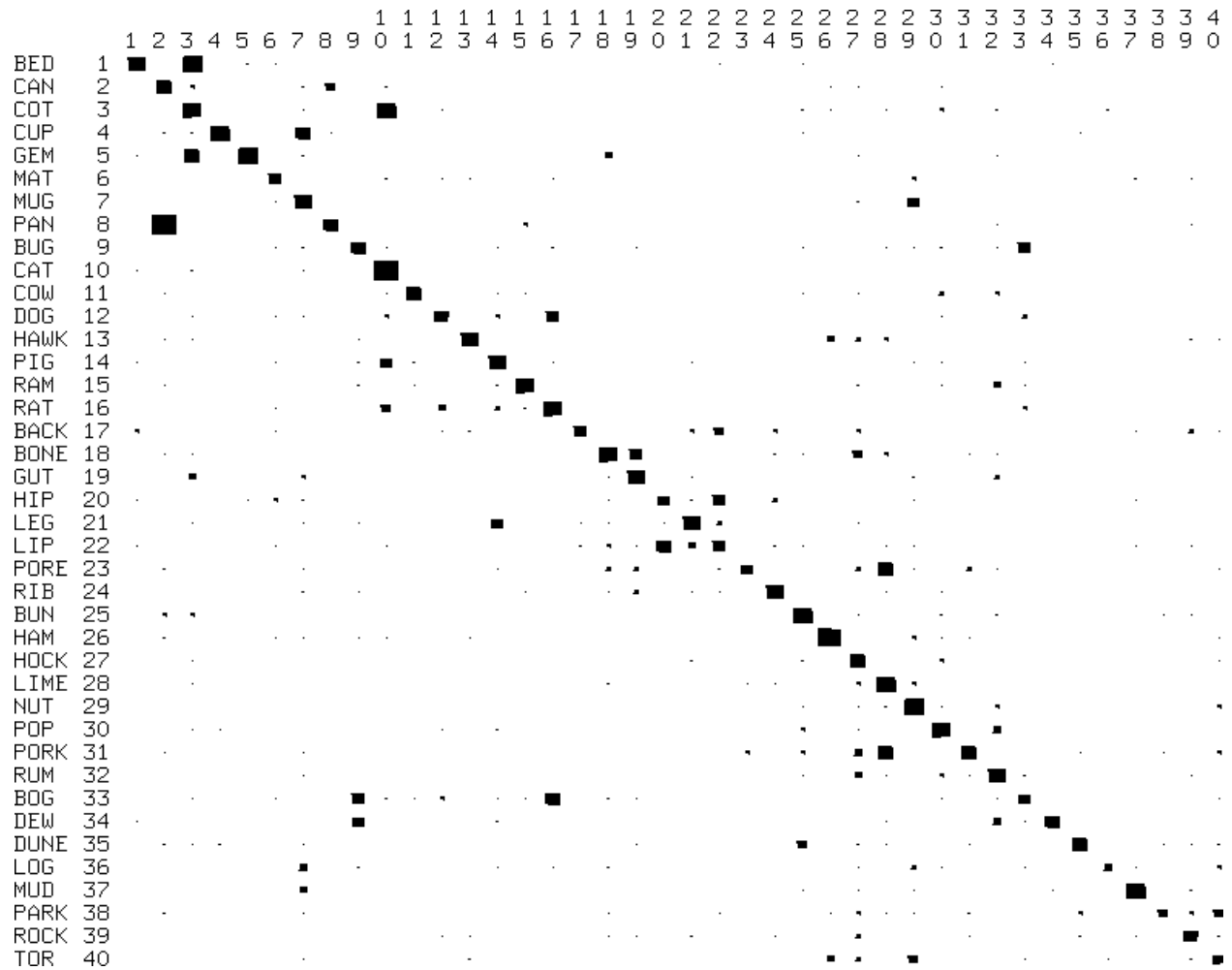
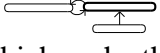


Figure 4.22: The confusion matrix for responses of the  network. Each row corresponds to a particular word as stimulus. The frequency with which each other word is given as the response is represented by the size of the square in the column with the corresponding number (listed at the top). The off-diagonal squares represent errors—their size is scaled relative to the most frequent error (PAN ⇒ “can”). The diagonal squares represent correct responses—since correct responses are much more common than any particular error, their sizes are scaled relative to the most frequently correct word (CAT).

that do not are quite similar (e.g. BED and RIB). Because a network can only be sensitive to the actual similarity among orthographic and semantic representations, it is possible that the distributions of error types produced under damage are biased by the inadequacies of the use of letter overlap and category membership as definitions for visual and semantic similarity.

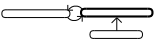
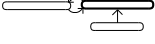
In order to ensure that our results are not biased by the particular definitions of similarity we used, we reclassified the errors produced by the  network using criteria for visual and semantic similarity based on the actual proximity values of each stimulus-response pair. For ease of comparison, the values of these criteria were defined so that the incidence of error types among all word pairs occurring by chance approximated that for the original definitions. Specifically, a pair of words were considered visually similar if the proximity of their orthographic representations was greater than 0.55, and semantically similar if the proximity of their semantic representations was greater than 0.47. While these criteria result in only a 0.5% decrease in the incidence of visual similarity and a 1.3% increase in the incidence of semantic similarity, they significantly change the distributions of these similarities over word pairs. This is because proximity is based on shared features, so that letters can resemble other letters without being identical, and words can be semantically related without being in the same category. As a result, there is only a 0.64 correlation between the assignment of visual similarity using letter overlap and using the proximity criterion. The correlation for semantic similarity is only 0.72. For both, only about three-fourths of the word pairs that are similar using the original definitions remain so using the proximity criteria.

Figure 4.23 shows the distribution of error types for lesions to the  network using the definitions of visual and semantic similarity based on proximity. Comparing with the corresponding results using the original definitions (shown in the right side of Figure 4.2, p. 78), there is remarkable similarity in the pattern of results. When the response criteria are used, the only significant difference is that the proximity-based definitions result in a lower rate of “other” errors for lesions of the direct pathway. Thus many of the error responses that are considered unrelated to the stimulus when using the original definitions do actually reflect the influences of visual or semantic similarity when measured more accurately. However, it should be noted that “other” errors still occur, as they do in patients. This effect is not apparent when using the noIP output network, although $0 \Rightarrow I$ lesions do produce a slightly higher rate of semantic errors with the proximity-based definitions. Overall, the similarity of the pattern of results indicates that the use of the original definitions for visual and semantic similarity, in terms of letter overlap and category membership, does not significantly bias the results.

4.8 Visual-then-semantic errors

In addition to producing error responses that are directly related to the stimulus either visually or semantically, deep dyslexics occasionally produce errors in which the relationship between stimulus

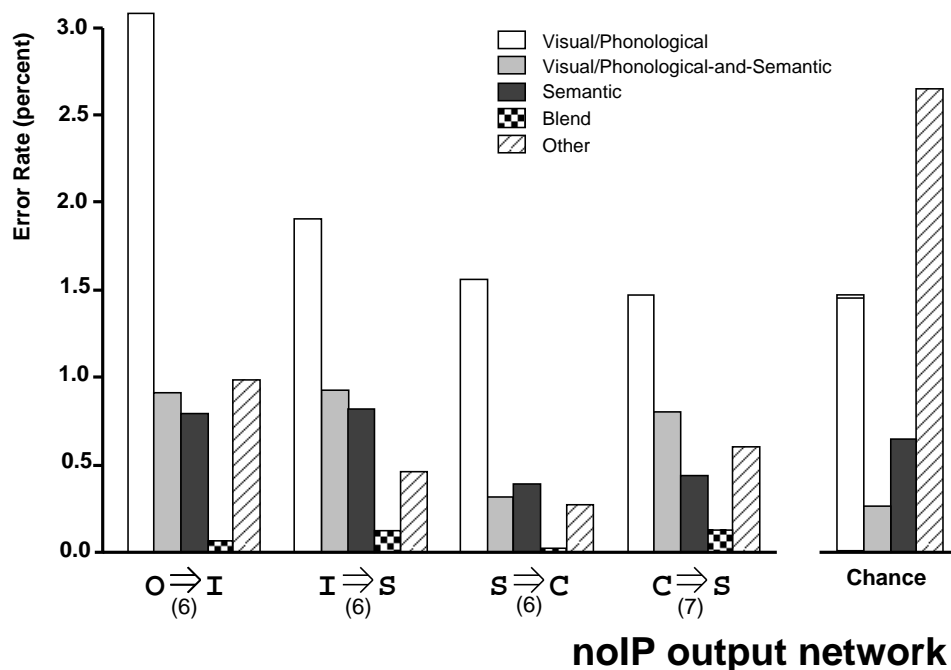
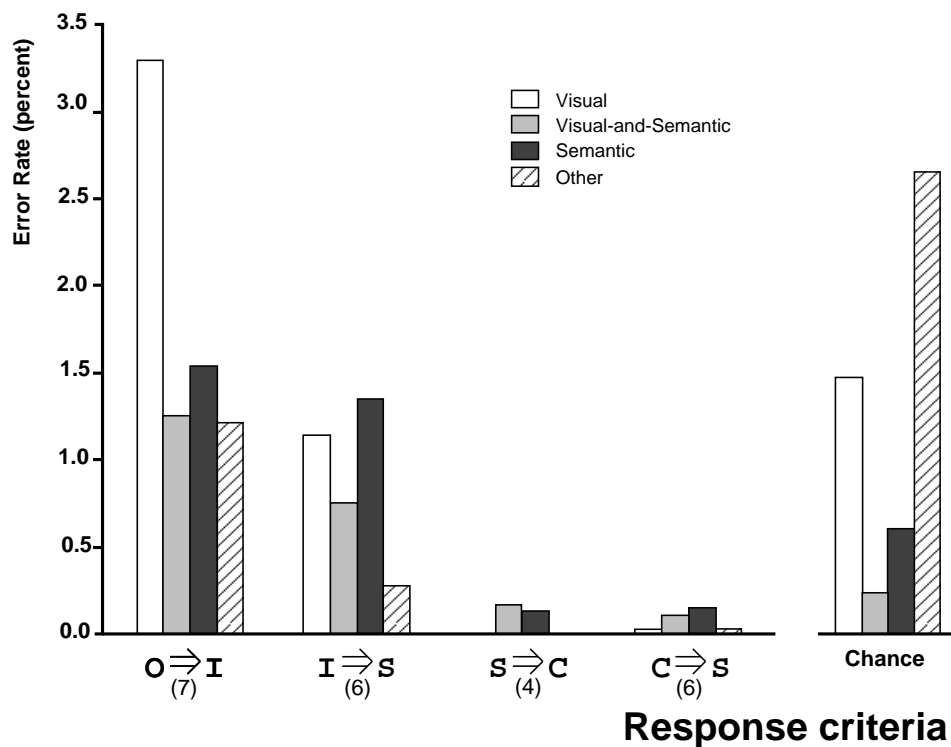
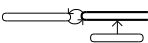
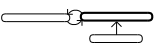
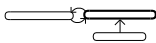


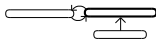
Figure 4.23: Error distributions for the  network, using the response criteria and the output network without intra-phoneme connections, when visual and semantic similarity is defined directly in terms of proximity.

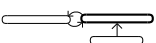
and response is more complex. For example, Marshall & Newcombe's (1966) patient G.R. read SYMPATHY as "orchestra." They considered this a visual error, SYMPATHY \Rightarrow "symphony", followed by a semantic error, SYMPHONY \Rightarrow "orchestra", and so termed it a "visual-then-semantic" error. Subsequently, this type of error has been observed in a number of other deep dyslexia patients (see Coltheart, 1980a)—other examples include STREAM \Rightarrow (steam) \Rightarrow "train" by H.T. (Saffran et al., 1976), FAVOUR \Rightarrow (flavour) \Rightarrow "taste" by D.E. and COPIOUS \Rightarrow (copies) \Rightarrow "carbon" by P.W. (Patterson, 1979). Although visual-then-semantic errors are quite rare, the possibility of their occurrence at all is rather perplexing, and certainly theoretically relevant. We know of no attempt to give an explanation of them other than Marshall & Newcombe's (1973) remark that they are "compound mistakes which are a function of misperception plus semantic substitution" (p. 186). They seem to be generally assumed to arise from combining two separate errors.

Given that visual-then-semantic errors are an acknowledged characteristic of deep dyslexic reading, the question arises as to whether they occur after lesions to our networks. Because the stimulus and response of a visual-then-semantic error are neither visually nor semantically related, up until now we would classify such errors as "other." Hence, we analyzed the "other" errors produced by the  network to determine whether some of them are more appropriately classified as visual-then-semantic. A visual-then-semantic error occurs when the stimulus and response are unrelated, but there is a third word, which we will call the "bridge," that is visually related to the stimulus, semantically related to the response, *and was directly involved in producing the error*. This last point is assumed for patient errors because the likelihood of a response being appropriately related to the stimulus by chance is assumed to be negligible. However, in the simulations the small size of the word set and high chance rate of visual and semantic similarity make it necessary to demonstrate that the relation of the presumed bridge word to the stimulus and response does not arise merely by random selection from the word set.

When using the criteria to generate responses, for each "other" error we identified the potential bridge word as the one whose semantics had the second-best match to those generated by the network under damage (the best matching word is the response). If this word was visually related to the stimulus and semantically related to the response, we considered the error to be visual-then-semantic. Of the 114 "other" errors produced by the  network, 49 (43.0%) satisfied these criteria (42.0% when using the proximity-based definitions of visual and semantic similarity). The chance rate of visual-then-semantic errors can be calculated by estimating how often the next-best matching word would meet the criteria even if it had no influence on the error. This rate is just the chance rate that the bridge is visually related to the stimulus times the chance rate that it is semantically related to the response, given that the response is neither visually nor semantically related to the stimulus. The first term is just the overall rate of visual similarity for word pairs other than the stimulus and response (29.9%). The rate that the bridge and response are semantically related by chance is much higher than the overall rate of semantic similarity because the bridge

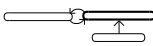
word was selected on the basis of how well its semantics match those generated by the network (which match the response best). We can use as an estimate the rate at which the response and bridge words are semantically related over *all* “other” errors produced by the network, which is 83.3%. Thus the chance rate of visual-then-semantic errors is approximately 24.9%, which is only slightly more than half the observed rate.

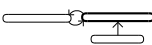
When using an output network, it is possible for the response generated at the phonological layer to differ from the best matching word at the semantic layer (even with the output network intact). Under these conditions we can apply a more conservative, but also more informative, definition of visual-then-semantic errors. Specifically, for each error in which the stimulus and response are unrelated, we can use the best-matching word at the semantic layer as the potential bridge word. If this word is visually related to the stimulus and semantically related to the response (but not identical or it would be a visual error), the “other” error is considered to be visual-then-semantic. It is clear that the bridge word is playing a role in the error because the phonological response is based solely on the generated pattern of semantic activity, which is most similar to that of the bridge word. Of the 97 “other” errors produced by lesions to the  network with the noIP output network generating responses, 12 (12.4%) satisfy the criteria for visual-then-semantic errors. In contrast, only four of the “other” errors (4.1%) involve semantic similarity followed by visual/phonological similarity (e.g. COW \Rightarrow (pig) \Rightarrow “pan”). Although the chance rate of this type of error is the same as for visual-then-semantic errors, it is observed much less frequently, both in patients and in the network.

The 12 visual-then-semantic errors produced by the network are listed in Table 4.2. Notice that for some of the errors (e.g. BOG \Rightarrow (pig) \Rightarrow “ram”) the generated semantics match those of the bridge word well enough to satisfy the response criteria (for a *visual* error). Even so, the semantics are sufficiently inaccurate that the (intact) output network produces a semantic error. All but one of the visual-then-semantic errors were caused by damage to the direct pathway, with most arising from $O \Rightarrow I$ lesions. This distribution across lesion locations very closely approximates the distribution of visual errors for the  network when using the response criteria (see top of Figure 4.2, p. 78).¹⁰ This makes sense given that, under our definition, visual-then-semantic errors consist of a visual confusion in the input network followed by a semantic confusion in the output network. In a sense, we interpret visual-then-semantic errors as visual errors gone awry under semantic influences. Because the damaged input network fails to clean up the visual error completely, the output network is given somewhat corrupted input. Even though it is intact, it may misinterpret this input as a semantically related word.

¹⁰In contrast, when using an output network to generate responses clean-up lesions produce about the same number of visual errors as lesions to the direct pathway. Since in this situation visual-then-semantic errors must consist of a visual error *followed* by a semantic error, their relative distribution across lesion locations may provide a more direct measure of the distribution of true visual errors. Thus, the pattern of visual-then-semantic errors provides further evidence that many of the visual error produced when using an output network are actually due to influences of phonological similarity.

Visual-then-Semantic Errors (noIP output network)				
Stimulus	Bridge	<i>prox</i>	Response	Lesion
BOG	DOG	0.85	RAT	$0 \Rightarrow I(0.1)$
BUG	NUT	0.68	LIME	$0 \Rightarrow I(0.15)$
BOG	PIG	0.91*	RAM	$0 \Rightarrow I(0.2)$
BONE	HOCK	0.94*	HAM	$0 \Rightarrow I(0.25)$
TOR	HOCK	0.78	HAM	$0 \Rightarrow I(0.25)$
CAT	PARK	0.64	ROCK	$0 \Rightarrow I(0.25)$
COW	CAN	0.79	PAN	$0 \Rightarrow I(0.25)$
GEM	LEG	0.79	BONE	$0 \Rightarrow I(0.3)$
HIP	HAWK	0.69	RAT	$I \Rightarrow S(0.25)$
BED	BONE	0.80*	HIP	$I \Rightarrow S(0.3)$
RAT	GUT	0.65	LEG	$I \Rightarrow S(0.4)$
HOCK	BONE	0.70	RIB	$S \Rightarrow C(0.5)$

Table 4.2: The visual-then-semantic errors produced by the  network with responses generated by the noIP output network. “Bridge” is the word whose semantics match those produced by the network best (with proximity *prox*). Bridge words with proximities satisfying the response criteria are marked with asterisks—these would have resulted in visual errors.

This interpretation raises the possibility of reclassifying some errors into three additional compound types based on the potential relationships of the bridge word to the stimulus and response: visual-then-phonological (e.g. $PARK \Rightarrow (pork) \Rightarrow \text{“pore”}$), semantic-then-phonological (e.g. $PARK \Rightarrow (bog) \Rightarrow \text{“bug”}$), and semantic-then-semantic (e.g. $PARK \Rightarrow (bog) \Rightarrow \text{“dune”}$). Visual-then-phonological errors would typically be classified as visual errors, but could be “other” (e.g. $GUT \Rightarrow (cat) \Rightarrow \text{“can”}$), semantic-then-phonological errors would usually be “other,” and semantic-then-semantic errors would be, by definition, semantic. The classification of errors in which the bridge word is both visually/phonologically and semantically related to the stimulus or response is somewhat ambiguous. To avoid confusion we restrict our analysis to errors that do not involve bridge words with mixed similarity. Applying these definitions to errors produced by lesions to the  network with the noIP output network, 5.0% of visual errors and 7.8% of other errors are visual-then-phonological, 3.0% of other errors are semantic-then-phonological, and 3.8% of semantic errors are semantic-then-semantic. Because these error types are defined in terms the derivation of a semantic representation that differs from both the stimulus and response, it is difficult to apply them to the reading responses of patients. In fact, many of them (e.g. visual-then-phonological, semantic-then-semantic) cannot be distinguished from more simple errors on the basis of the stimulus and response alone. It might be possible to use detailed tests of comprehension to determine when a patient derives semantics that are different from both the stimulus and response—the model predicts that this should occur occasionally, although rarely.

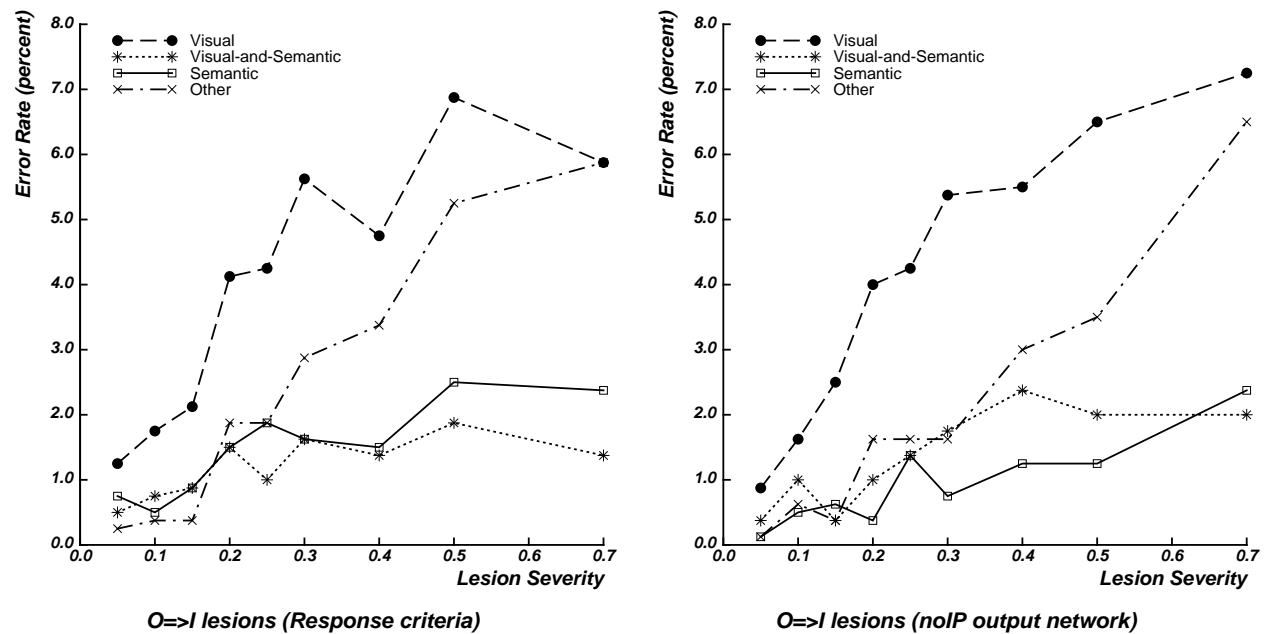
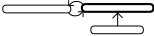


Figure 4.24: Rates of each error type across lesion severities for $0 \Rightarrow I$ lesions in the  network, using the response criteria and the noIP output network.

4.9 Effects of lesion severity on error type

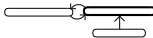
To this point, all of the data we have presented on the relationship between types of errors have averaged over a range of lesion severities, typically over those producing correct performance between 20–80%. However, it is possible that the distribution of error types changes with lesion severity. In addition, the extent of this effect may be influenced by the nature of the output system employed. Rather than address these issues for all of the network architectures, we present data from only the  network. Similar results obtain for the other networks.

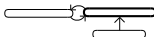
Figure 4.24 presents the rates of each type of error as a function of lesion severity for $0 \Rightarrow I$ lesions in the  network, using both the response criteria and the noIP output network. The plots are somewhat difficult to interpret due to the variability of the data—however, a number of overall effects are present. The first most obvious effect is that error rates increase with lesion severity. Our main motivation for averaging only over lesions producing 20–80% correct performance in previously reported results is that otherwise the results would be dominated by effects from the most severe lesions, which often do not show the typical distribution of error types. It is also the case that the correct performance of most of the patients we are considering falls within this range. The most interesting effect is that the rates of visual and other errors rise more quickly with increasing lesion severity than the rates of semantic and mixed visual-and-semantic errors.

Figure 4.25 shows the same data replotted in terms of the *proportion* of each error type as a function of incorrect performance.¹¹ The proportion of error responses that are unrelated to the

¹¹Incorrect rather than correct performance is used in order to correspond more directly with lesion severity, with

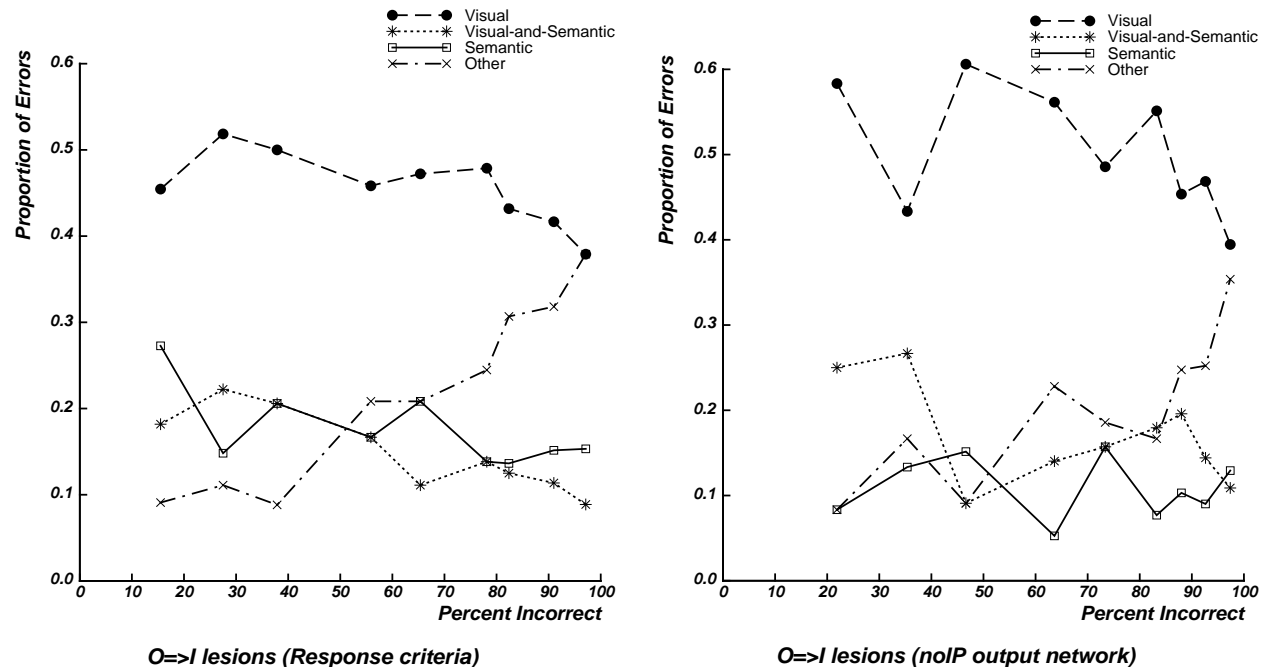
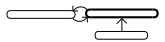


Figure 4.25: Proportions of each error type as a function of incorrect performance for $O \Rightarrow I$ lesions in the  network, using the response criteria and the noIP output network.

stimulus increases steadily as performance gets worse. The proportions of the remaining error types all decrease at about the same rate, both when using the response criteria and the noIP output network. Thus for the moderate lesions we consider the relative proportions of the various error types do not change drastically with lesion severity, and so our decision to average over lesions producing moderate correct performance appears warranted.

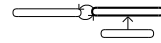
4.10 Error patterns for individual lesions

Our procedure for lesioning a set of connections involves randomly selecting some proportion of the connections and removing them from the network. In order to ensure that the ensuing effects are not peculiar to the particular connections removed, we carry out 20 instances of each type of lesion and average the results across them. On the other hand, it must be kept in mind that the model is compared with individual patients, each of whom have a particular lesion. In a sense, for a given simulation experiment with four locations of nine severities of lesion, we are creating 720 simulated patients, with a relatively high proportion of them displaying the characteristics of deep dyslexia. However, there are some issues in deep dyslexia, involving the relationship of performance on individual words for the same lesion, that to this point we have been unable to address.

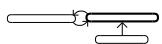

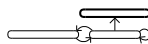
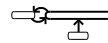
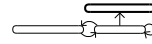
good performance on the left and poor performance on the right.

One issue concerns the correct performance on words that are given as responses in errors. Some theories of reading errors in deep dyslexia (e.g. Morton & Patterson, 1980) assume that a word produces an error when its lexical entry is missing from some lexicon, with a closely-matching word whose lexical entry is present being given as the response. If we also assume that words are read correctly when their entries are present in the lexicon, such a theory predicts that words given as responses in errors should always be read correctly.

In fact, patients usually, but not always, adhere to this pattern. For example, D.E. read SWEAR as “curse” but then gave the response “I don’t know” to CURSE as stimulus (K. Patterson, personal communication). G.R. gave no response to SHORT or GOOD, but produced the errors LITTLE \Rightarrow “short” and BRIGHT \Rightarrow “good”, as well as the errors BLUE \Rightarrow “green” and GREEN \Rightarrow “peas” (Barry & Richardson, 1988). In fact, at another time G.R. read correctly only 54% of words he had previously given as responses in semantic errors—just slightly better than his original correct performance of 45% (Marshall & Newcombe, 1966).

If we examine the pattern of correct and incorrect performance for individual lesions of the  network when using the response criteria, we find that only 64.1% of the words given as the response in an error are read correctly. 31.2% of error responses produce an omission while 4.6% lead to another error. Separating errors by type, the responses in mixed visual-and-semantic errors are most likely to be read correctly (74.2%), followed by visual errors (70.9%) and semantic errors (63.0%), while responses in “other” errors are least likely (48.1%). The high rate of omissions may simply be due to our stringent criteria for overt responses. However, the fact that 4.6% of error responses produce other errors when presented as stimuli clearly violates the prediction of a theory that explains errors in terms of missing lexical entries. In the damaged network, the attractor for a word is not either present or absent, but rather can effectively operate to produce a response given some inputs but not others.

It is possible for an even more perplexing relationship to hold among the words producing errors in a patient. It has been observed that a pair of words may produce each other as error responses. For example, G.R. produced THUNDER \Rightarrow “storm” and STORM \Rightarrow “thunder” (Marshall & Newcombe, 1966), while D.E. produced ANSWER \Rightarrow “ask” and ASKED \Rightarrow “answer” (K. Patterson, personal communication). It is hard to imagine how a mechanism that maps letter strings to pronunciations via meaning might possibly produce such behavior under damage.

Such response reversals occur in our simulations, but they are very rare. None are found in the corpus of errors produced by the  network. However, both the  and  networks produce a few of them when using the response criteria. For example, a $0 \Rightarrow I(0.1)$ lesion to the  network resulted in the visual errors MAT \Rightarrow “mud” and MUD \Rightarrow “mat”, while a $0 \Rightarrow I(0.7)$ lesion produced the visual errors MUG \Rightarrow “nut” and NUT \Rightarrow “mug”. Similarly in the  network, a $0 \Rightarrow I(0.3)$ produced the “other” errors MUG \Rightarrow “hock” and HOCK \Rightarrow “mug”, while a $0 \Rightarrow I(0.7)$ lesion produced the mixed visual-and-semantic errors HIP \Rightarrow “lip” and LIP \Rightarrow

“hip”.

How might a network produce such response reversals? Recalling Figure 2.10 (p. 43), we can interpret damage to the direct pathway as corrupting the initial pattern of semantic activity derived from orthography. One explanation for the existence of response reversals is that the attractors for words are sensitive to different aspects of this pattern. For example, suppose that the attractor for HIP depends on some particular set of initial semantic features to distinguish it from LIP, but the attractor for LIP depends on a *different* set to distinguish it from HIP (this cannot be represented in a two-dimensional rendition of semantic feature space like that in Figure 2.10). If both of these sets of features are lost due to a particular lesion, the errors HIP \Rightarrow “lip” and LIP \Rightarrow “hip” are both possible. In essence, an explanation for response reversals must allow a more complicated interaction between orthographic and semantic information than is typically provided in theories based on discrete lexical entries for words.

4.11 Summary

An examination of the effects of lesions on five alternative architectures for mapping orthography to semantics has served both to demonstrate the generality of the basic H&S results as well as to clarify the influences of aspects of network architecture on the detailed pattern of errors. A consideration of more specific effects at the level of individual lesions, error types, and words reinforced the correspondence of network and patient behavior.

Perhaps the most general principle to emerge from these experiments is the importance of the nature of the attractors developed by the network. Although network architecture can have a strong influence on this process, ultimately it is the learning procedure which derives the actual connection weights that implement the attractors. Thus it is important that we evaluate whether the nature of the attractors, and hence the behavior they exhibit under damage, are the result of specific characteristics of the back-propagation learning procedure, or whether the results would generalize to other types of attractor networks. The next chapter addresses this issue by attempting to replicate and extend the results obtained thus far using a deterministic Boltzmann Machine and a closely-related stochastic GRAIN network.

