

# Chapter 7

## Relearning after damage

Cognitive neuropsychology aims to extend our understanding of normal cognitive mechanisms on the basis of their pattern of breakdown due to brain damage in neurological patients. A major motivation for many researchers is that a more detailed analysis of the normal mechanism, and the way it is impaired in particular patients, may lead to the design of more effective therapy to remediate these impairments (Howard & Hatfield, 1987). Significant progress has been made in analyzing cognitive mechanisms and their impairments in terms of information-processing models, particularly in the domain of written language (Coltheart et al., 1980; Coltheart et al., 1987b; Patterson et al., 1985). However, relatively few remediation studies have been based directly on cognitive analyses, and while these few have been relatively successful, the specific contribution of the cognitive model is often unclear (for examples and general discussion, see Behrmann, 1987; Byng, 1988; Caramazza, 1989; de Partz, 1986; Mitchum & Berndt, 1990; Riddoch & Humphreys, 1991; Seron & Deloche, 1989; Wilson & Patterson, 1990).

The previous chapters have demonstrated how principles central to connectionist modeling— attractors and distributed representations—are useful in understanding and reproducing the detailed behavior of one class of neurological patients, deep dyslexics. Here we attempt to extend the relevance and usefulness of connectionist modeling in neuropsychology to address issues in the rehabilitation of cognitive deficits following brain damage. The main issues we will address concern the degree and speed with which behavior can be reestablished as a result of therapy, the extent that recovery due to treatment of particular items generalizes to other materials, and the possible bases on which to select items for therapy so as to maximize this generalization. Ultimately, the hope is that an analysis of the correspondence between the nature of recovery in patients and in damaged networks can inform theories of normal functions as well as lead to improved therapy for patients. At this stage the current research merely points in interesting directions.

We begin by summarizing some of the studies on remediation in acquired dyslexia based on cognitive models of normal reading, focusing on those that attempt to reestablish aspects of the mapping between orthography and semantics. We then describe previous preliminary work on the effects of relearning after damage in connectionist networks. Following this, a set of simulation

experiments are presented in which networks previously used to model impairment in mapping orthography to semantics in deep dyslexia are retrained after different types of damage. The degree of variability in the amount of recovery and generalization for different lesion locations in the networks has interesting implications for understanding the effects seen in patients. The chapter concludes by testing a particular hypothesis on how to select materials for therapy that maximize recovery and generalization.

## 7.1 Cognitive remediation of acquired dyslexia

Cognitive rehabilitation is based on the notion that, by ascribing a patient's deficits to the selective impairment of one (or a few) of a set of functionally separable subsystems involved in carrying out a task, therapy can more effectively focus on the remediation of particular types of representations and processes. In essence, the cognitive analysis enables a more detailed diagnosis of the impairment in an information-processing framework that can often suggest relevant therapeutic procedures. Since our focus up to this point has been on deep dyslexia, it would be most natural to consider remediations studies involving this type of patient. Recall that the standard characterization of these patients in terms of a dual-route model of reading is that the phonological route (i.e. the direct mapping from orthography to phonology) is severely impaired, while the semantic route is (relatively) intact. Unfortunately (for us but perhaps not for the patients), the few published remediations studies of deep dyslexics (de Partz, 1986; Hatfield, 1983) have attempted to reestablish the phonological route rather than the semantic route, and thus are of little direct relevance to the current research. In fact, most work in reestablishing the mapping between orthography and semantics has involved the complementary type of patient—surface dyslexics (Patterson et al., 1985). These patients rely primarily on the phonological route(s) due to impairment in the semantic route.

Coltheart & Byng (1989) undertook a remediation study with surface dyslexic E.E. It is commonly acknowledged that the main characteristics of surface dyslexia—regularization errors (e.g. reading PINT to rhyme with MINT)—can arise from a number of different underlying functional impairments (Coltheart & Funnell, 1987). On the basis of a number of preliminary tests, Coltheart & Byng identified E.E.'s specific deficit as being in accessing semantics from orthography. The most important indication was the occurrence of homophone confusions (i.e. misunderstanding TALE as TAIL). These errors demonstrate that the patient is mapping orthography to semantics via phonology (where TALE and TAIL are indistinguishable) rather than directly. To improve the patient's ability to associate the written form of words directly with their meanings, Coltheart & Byng designed a study involving words containing the spelling pattern OUGH (e.g. THROUGH, COUGH, BOUGH), whose pronunciations are highly irregular (and thus pose problems for patients like surface dyslexic who can only sound-out words). E.E. was retrained on 12 of 24 such words, in which

he studied the written words augmented with mnemonics for their meaning. Prior to retraining, four of the trained words were read correctly; after retraining, all 12 were read correctly. Thus the therapy was effective for the retrained items, and this was not due to “spontaneous recovery” or other non-specific effects because performance on the words did not improve in the period before or after therapy. Surprisingly, the 12 unretrained words also improved as a result of the therapy, from one correct prior to retraining, to seven correct after retraining. Thus the improvement due to generalization is 75% as large as the improvement caused by direct treatment. Similar effects were obtained in two additional studies. Thus Coltheart & Byng found that, under some conditions, retraining the mapping from orthography to semantics for some words can generalize to other words.

In a similar study, Scott & Byng (1988) attempted to remediate homophone confusions in another surface dyslexic, J.B. The treatment involved selecting the correct homophonic word from six alternatives in order to meaningfully complete each of 136 sentences. Over the course of 29 sessions, performance improved from about 70% to nearly perfect. A second task was administered pre- and post-therapy, in which each of 270 homophonic words (135 pairs, half of which were treated during therapy) was embedded in both an appropriate and inappropriate sentence, and J.B. had to sort the resulting 540 sentences accordingly. Similar to the Coltheart & Byng study, J.B. showed therapy-specific improvement for sentences containing both the retrained words and, to a lesser but still significant extent, the unretrained words. However, Scott & Byng failed to find improvement in J.B.’s *writing* of either set of homophonic words in sentence contexts, suggesting that generalization occurred within but not between orthographic domains.

Behrmann (1987) carried out an analogous study on C.C.M., a surface dysgraphic (i.e. a patient who *writes* with impaired lexical/semantic mediation). Behrmann used picture matching and sentence completion tasks to train C.C.M. to produce the appropriate spelling of 25 of 69 homophonic word pairs that she initially spelled inappropriately. Therapy improved overall performance from 49% to 67%, but no improvement on the unretrained homophone pairs was observed. However, the writing of 75 words with irregular spellings (e.g. COMB) did significantly improve as a result of the therapy. Thus, Behrmann’s study produced some type of generalization, but not specifically to other homophonic pairs.

Overall, the therapy studies that attempt to reestablish associations between orthography and semantics have succeeded in improving performance for retrained items, and often also for unretrained but related items. The degree of recovery and generalization can vary considerably across patients, although the comparison is difficult because different remediation techniques are employed. Connectionist networks should show similar effects if they are to be considered relevant for understanding and improving on patient therapy. We now describe preliminary work on relearning in these networks after damage.

Figure 7.1: The recovery of performance of the Hinton & Sejnowski network after various types of damage. The heavy line is a section of the original learning curve after a considerable number of learning sweeps. All the other lines show recovery after damaging the network once learning was complete (99.3% correct). The lines with open circles show the rapid recovery after 20% or 50% of the weights to the hidden units have been set to zero (but allowed to relearn). The dashed line shows recovery after five of the 20 hidden units have been permanently ablated. The remaining solid line is the case when uniform random noise between  $\pm 22$  is added to all the connections to the hidden units. In all cases, a successful trial was defined as one in which the network produced *exactly* the correct semantic features when given the grapheme input (from Hinton & Sejnowski, 1986, p. 311).

## 7.2 Previous studies of relearning in networks

As described in Section 2.5, H&S's efforts in modeling deep dyslexia were motivated in large part by the earlier work of Hinton & Sejnowski (1986) on the effects of damage in a simple Boltzmann Machine that mapped orthography to semantics. In addition to demonstrating visual and semantic influences in the errors produced by the damaged network, Hinton & Sejnowski investigated the behavior of the network in relearning associations after damage. Specifically, after the network had learned all 20 associations of three-letter strings to (arbitrary) semantic patterns, it was damaged in a variety of ways, either by zeroing or adding noise to the weights, or by removing hidden units. For each of these, when the damaged network was retrained on the associations, its performance improved much more quickly than when it was initially learning them and had reached the equivalent level of performance (see Figure 7.1). Sejnowski & Rosenberg (1987) later replicated the effect in NETtalk, a feed-forward back-propagation network for mapping

orthography to phonology that predated the Seidenberg & McClelland (1989) model.

Hinton & Sejnowski explain the rapid relearning after damage in terms of the shape of the error surface in weight space. During original learning, most combinations of weight changes (i.e. directions in weight space) would make overall performance on the task much worse; there are only a relatively small set of directions that would improve performance, even if only slightly. In geometrical terms, this corresponds to being at the bottom of a *ravine* in the error surface. Most directions slope steeply up a wall of the ravine—learning requires moving gradually down the more gently sloping floor. This situation still holds at the end of learning. Damage to the network, such as adding random noise to weights, moves the network in a random direction in weight space which has only a small component along the floor of the ravine—most of the movement is in a direction which takes the network up a steep wall.<sup>1</sup> As a result, there is a large gradient back towards the bottom of the ravine that causes rapid improvement in performance when the network is retrained on the associations. In fact, as the figure shows, almost all of the impairment due to damage is eliminated in the first few retraining sweeps—only the component of the noise along the floor of the ravine requires extended retraining to produce full recovery.

An even more interesting effect found by Hinton & Sejnowski, analogous to the generalization found with patients, was that rapid recovery of all associations occurred even if the network was retrained on only a subset of them. If only 18 of the 20 associations were used in retraining after random noise was added to the weights, correct performance on the remaining two improved from 30% to 90% for one pair, and from 17% to 98% in a second experiment involving a pair with higher error rates. However, the effect was not very robust—when 15 associations were retrained, performance on the remaining five got slightly worse. This transfer to unretrained associations is paradoxical because there is no intrinsic relationship between any of the associations—they were all generated randomly. However, during the original learning the weights capture whatever chance regularities there happen to be among the entire set of associations. Most of these regularities still hold for the retrained subset, and so the weights tend to move back towards values that capture these regularities during relearning. Since most of these apply to the unretrained associations as well, they also improve. In other words, since knowledge of the associations is distributed across all of the connections, when the damaged network is retrained on some of the associations, all of the weights are pushed back towards their original values. In geometrical terms, the direction of gradient to the bottom of the ravine in weight space for the *entire* set of associations is well-approximated by the gradient for only a subset of them.

Hinton & Plaut (1987) further investigated the transfer effect in a network in which each connection has both a slow, plastic weight and a fast, elastic one. The slow weights are like those normally used in connectionist networks—they change slowly and encode all of the long-term

---

<sup>1</sup>Zeroing weights or removing units does not actually move the network in a random direction in weight space. Nonetheless, the fact that these types of damage produce qualitatively similar relearning behavior suggests a similar explanation applies.

knowledge of the network. The fast weights change much more rapidly but continually regress towards zero, so that their values are determined solely by the recent past. The effective weight on a connection when computing unit activities is the sum of its slow and fast weight. Thus, at any instant we can think of the network's knowledge as consisting of long-term knowledge (in the slow weights) that captures the inherent regularities in the environment, with a temporary overlay of short-term knowledge (in fast weights) that compensates for particular characteristics of the current context.

One benefit of using fast weights is that they can learn to cancel out the interference in a set of old associations caused by more recent learning. To demonstrate this, Hinton & Plaut built a fully-connected feed-forward network with slow and fast weights, which had 10 input units, 100 hidden units, and 10 output units. The network was trained with back-propagation on 100 associations of random binary vectors of length 10, in which each component of each vector had probability 0.5 of being a 1. Although both the slow and fast weight on each connection experience the same error gradient, most initial learning occurs in the fast weights because they can change more quickly. However, their strong tendency to remain small prevents them from completely solving the task by themselves, and the slow weights gradually learn under the pressure of the residual error gradient. As the slow weights take over more of the task, the fast weights can decay further. Thus knowledge is gradually transferred from the fast to slow weights (from the short-term context to long-term knowledge) until, at the end of learning, the network performs the task perfectly, the fast weights are near zero, and all of the knowledge is in the slow weights.

Once the network had learned the 100 associations in this way, Hinton & Plaut trained it on five new random associations without further rehearsal of the original 100, continuing training until all of the new knowledge was in the slow weights. Because these associations are unrelated to the original ones, the weight changes they induce move the network in a random direction in weight space relative to the original task. In this way, the interference caused by training on five new associations is analogous to the addition of random noise used by Hinton & Sejnowski to corrupt their network's performance. In fact, performance on the original task was significantly impaired as a result of the interference training. The network was then retrained on only half of the original 100 associations. Not only did the retrained associations recover quickly, but performance on the remaining associations improved almost as much. The ratio of unretrained to retrained improvement was 0.83. In fact, there was considerable transfer (0.65) when only 10% of the original associations were retrained (see Figure 7.2).

Since the recovery of performance occurs during the first few retraining sweeps, it takes place almost entirely within the fast weights. Nothing in the interaction between fast and slow weights is required for the transfer effect—a network with only the slow weights (like the Hinton & Sejnowski network) would also exhibit it. The advantage of using both fast and slow weights is that the relearning in the fast weights need not permanently interfere with the new associations—if

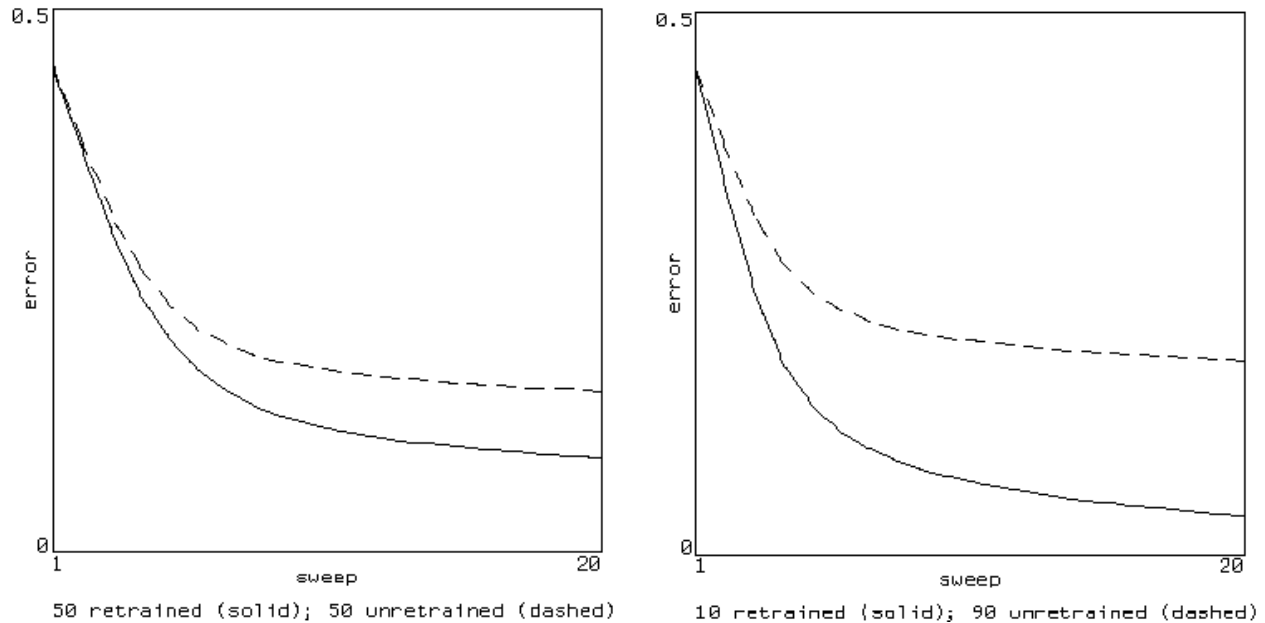


Figure 7.2: Average error per output unit for the retrained (solid) and unretrained (dashed) subsets when retraining on 50 (left) or 10 (right) of the original 100 associations (from Hinton & Plaut, 1987, p. 181). Notice that error rather than correct performance is being plotted, so improved performance is reflected by *decreasing* curves.

the fast weights are allowed to decay back to zero, the new knowledge is restored. However, if retraining is continued the knowledge will be gradually transferred into the slow weights with minimal interference to the new associations.<sup>2</sup>

The transfer effect depends on the degree to which the decision surfaces in weight space that accomplish unrelated associations are not completely orthogonal—movement towards some of them must tend to move towards the others. In analyzing the expected amount of transfer as the size of the network is increased, Hinton & Plaut showed it is important to distinguish two cases. If the input vector components have zero mean (e.g. units have states between  $-1$  and  $1$  as in the DBM in Chapter 5), the expected degree of orthogonality between randomly related

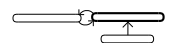

<sup>2</sup>In a set of unpublished experiments, Hinton & Plaut compared learning with fast and slow weights vs. learning with only slow weights in terms of the amount of interference to old knowledge caused by additional learning. A network with the same 10–100–10 architecture just described was trained on 50 random binary associations until performance was near-perfect and all of the knowledge was in the slow weights. It was then trained on an additional 50 associations without rehearsal of the original associations, either using both fast and slow weights, or just slow weights. Training again continued in each condition until performance on the new associations was near-perfect and all of the knowledge was in the slow weights. The interference of this new learning on the original knowledge was measured both in terms of the average error rate on the original associations, and the absolute distance moved in weight space. Learning the new associations with both fast and slow weights, compared with learning with slow weights alone, caused less error on the original associations (average sum-squared error per association: 0.93 vs. 1.23) and less movement in weight space (euclidean distance: 20.2 vs. 33.3). In essence, the fast weights perform a “look-ahead” search for a good set of weights close to the current slow weights. The residual error for the combination of fast and slow weights pulls the slow weights directly towards these good weights, in a kind of “shortest” descent, rather than simply in the direction of steepest descent.

vectors increases with their dimensionality and the expected transfer correspondingly decreases. However, if the vector components range between 0 and 1, randomly related vectors do not become increasingly orthogonal and the expected amount of transfer is independent of the dimensionality of the associations.

The rapid relearning after damage and the transfer to unretrained associations suggest that relearning after damage in connectionist networks may provide a useful framework for understanding the effects seen in the remediation of acquired dyslexia. However, the generalization experiments of Hinton & Sejnowski and Hinton & Plaut involved relearning after unrelated weight changes, but not after permanent damage. The explanation they offer, in terms of the gradient back towards the original set of weights, does not formally apply to a lesioned network. A lesion collapses weight space along the dimensions corresponding to removed connections, so the original set of weights is no longer possible. Rather, relearning must adjust the weights on the remaining connections to new values in order to compensate for the missing connections. The next section investigates the nature of the transfer effects after lesions to networks like those previously used to model deep dyslexia.

## 7.3 Experiments in relearning after damage

### 7.3.1 The training procedure

In order to investigate relearning after damage in networks that map orthography to semantics, we developed versions of two of the back-propagation networks used to model deep dyslexia: the  and  networks (see Figure 4.1, p. 73). The original versions of these networks could not be used because they were trained with momentum (see Equation 10.5, p. 304). We would like to compare the rate of relearning after damage with the rate of original learning at the same level of performance—this comparison would not be fair if the original learning had the benefit of momentum at this point but relearning did not.

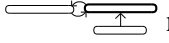
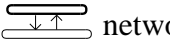
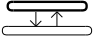
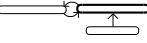

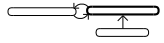
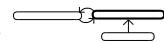
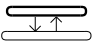
Accordingly, we trained new versions of the two networks on the H&S word set, using no input noise, a higher learning rate, and no momentum ( $\epsilon = 0.01$  and  $\alpha = 0.0$  in Equation 10.5, p. 304). In conventional learning, momentum is particularly important near the end of learning when all output units are easily on the correct side of 0.5 for every case, but do not quite achieve the accuracy required to stop training (within 0.1 of their correct states in previous simulations). In order to avoid prolonging this final stage of training unnecessarily, we lowered the required accuracy to 0.2. Each sememe unit achieved this level of accuracy over the last three of eight iterations for each of the 40 words after 4740 sweeps through the training set for the  network, and 6210 sweeps for the  network. Even with the reduced accuracy criterion, both networks easily satisfied the H&S proximity/gap response criteria for explicit naming of every word.

Figure 7.3 shows the learning curves of the two networks, in terms of the average proximity



of the generated and correct semantics, and the average percentage of words read correctly using the response criteria. Both networks show a rapid improvement in correct performance after about 500 sweeps, when performance has reached around 20%. We will compare the relatively “rapid” original learning at this point with the speed of relearning in the simulations to follow.

### 7.3.2 The effects of lesions

The main sets of connections in each of the two networks were then subjected to 20 instances of lesions of each standard severity, ranging from 0.05 to 0.7. For each lesion location, Figure 7.4 shows the correct performance of each network using the response criteria, as a function of lesion severity. The networks are about equally sensitive to  $0 \Rightarrow I$  lesions, but the  network is much more severely impaired by  $I \Rightarrow S$  lesions than the  network. Presumably this is because these connections are involved in implementing attractors in the former network but not the latter.  $S \Rightarrow I$  lesions in the  network behave quite similarly to  $S \Rightarrow C$  lesions in the  network. These effects were also true of their previously developed counterparts (see Figure 4.2, p. 78, for the  network, and Figure 4.10, p. 92, for the  network). Both of the current networks are also slightly less robust overall than the previous networks, but this most likely reflects the fact that the latter networks develop stronger attractors as a result of being trained with noisy input.

### 7.3.3 The relearning procedure

We used the following procedure to study the effects of relearning. For a given instance of a lesion, the responses to the 40 words were categorized as correct or incorrect—for this purpose errors and omissions were both considered incorrect and not distinguished. Half of the correct words and half of the incorrect words were randomly selected and placed in the “retrained” set; the remaining words were placed in the “unretrained” set. If an odd number of words were correct, the extra correct word was placed in the unretrained set and the extra incorrect word was placed in the retrained set. Thus both the retrained and unretrained sets always contained 20 words. No attempt was made to balance the average proximity of the retrained and unretrained sets, although these tended to be fairly similar.

The damaged network was then retrained for 50 sweeps on the retrained words only. Performance was measured at each sweep during relearning separately for the retrained and unretrained word sets, in terms of the average proximity of the generated and correct semantics, and the average percentage of words read correctly using the response criteria. In order to ensure that any relearning effects were not simply due to an imbalance in initial performance between the retrained and unretrained sets, the two sets were exchanged and the retraining was repeated, starting from the same initial set of weights. Finally, the weights were again reinitialized and the damaged network

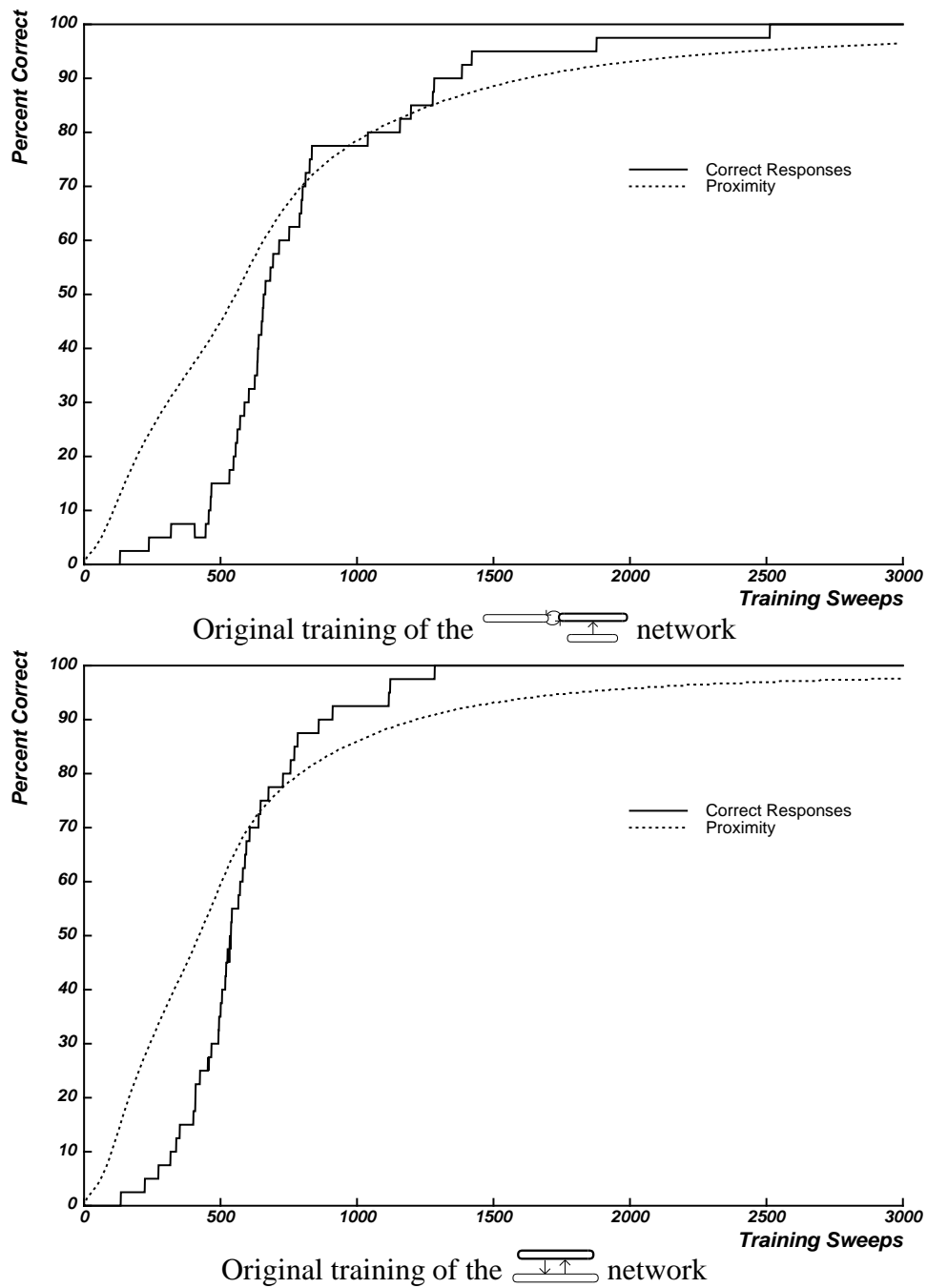
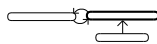



Figure 7.3: Improvement in correct performance during original learning in the  network (top) and  network (bottom). Proximity has been scaled relative to the initial proximity (0.455 for both networks).

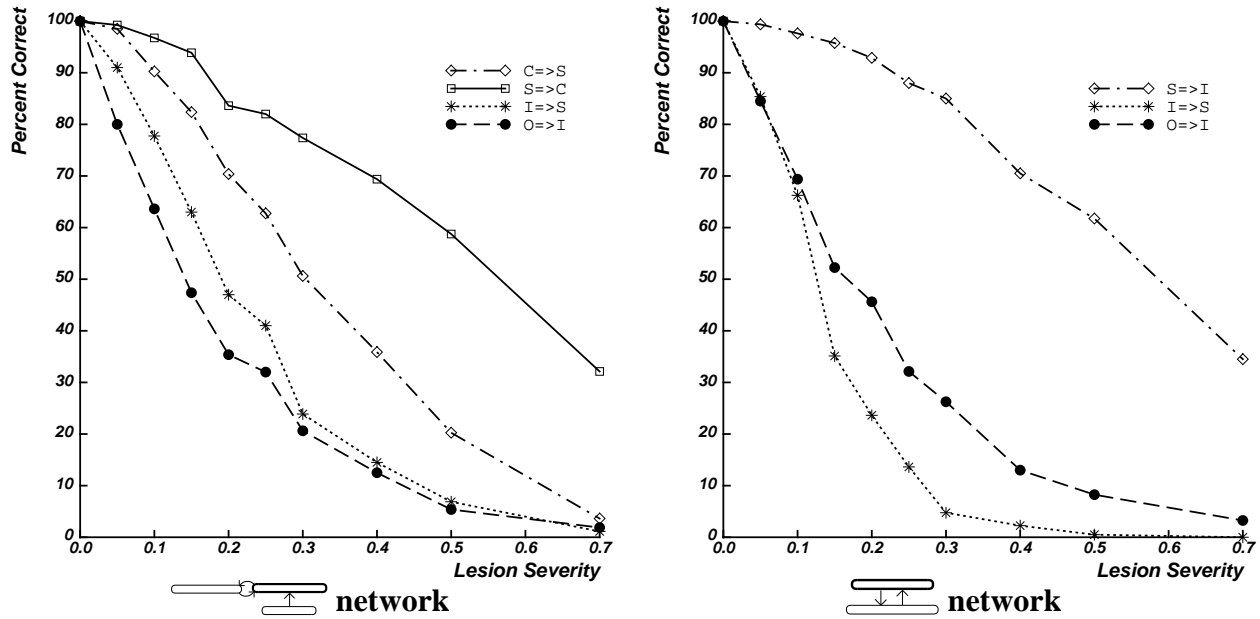







Figure 7.4: Correct performance rates using the response criteria, after lesions to each main set of connections as a function of lesion severity, for the  network (left) and  network (right).

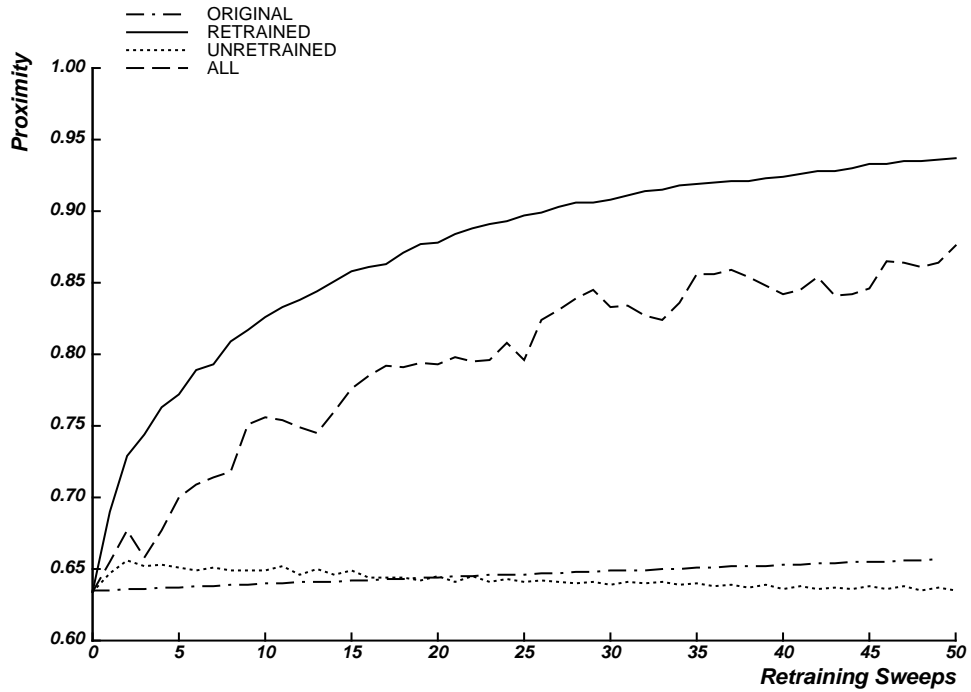
was retrained on all 40 words.

The retraining procedure requires about 250 times more computation than the procedure for gathering error data, and thus could not be applied to every instance of lesion at every location and severity in both networks.<sup>3</sup> Accordingly, we only investigated the particular lesion locations and severities that produced correct performance just above 20%. This level of performance was selected because it (just) falls within the range of performance included in our error analysis in previous chapters, but it is sufficiently poor to provide room for significant improvement over the course of retraining. We first consider the effect of relearning after lesions located prior to where attractors operate:  $O \Rightarrow I$  and  $I \Rightarrow S$  in the  network;  $O \Rightarrow I$  in the  network. Following this, we consider the effects of within-attractor lesions.

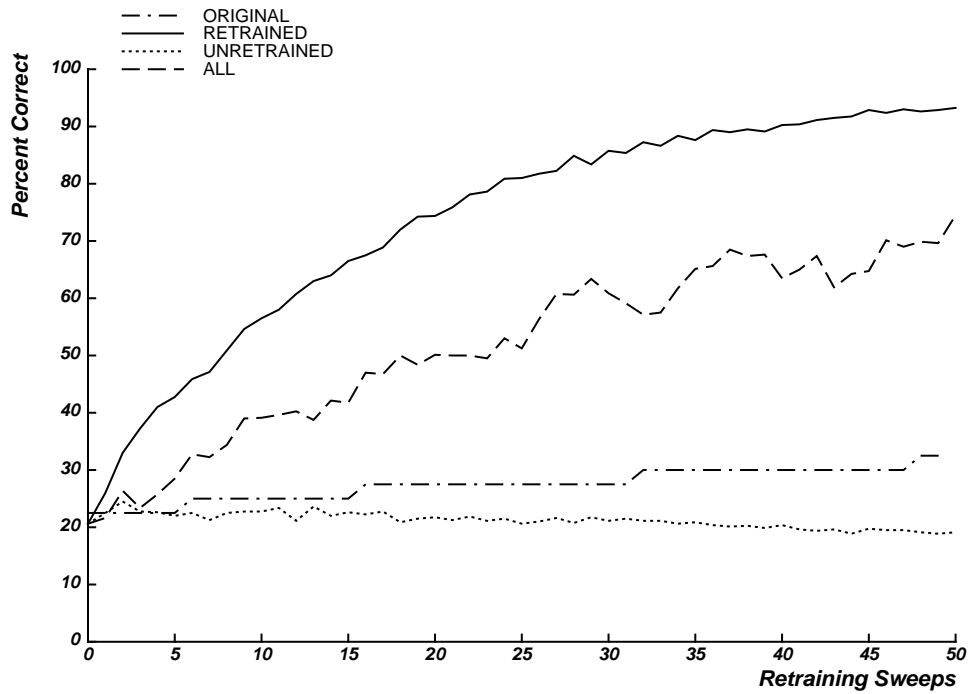
### 7.3.4 Pre-attractor lesions

Figure 7.5 presents the retraining results for  $O \Rightarrow I(0.3)$  lesions of the  network, averaged over all 20 lesion instances and over exchanges of the retrained and unretrained word sets. This lesion reduces overall correct performance to 20.6% on average. Data on the average proximity and correct performance are plotted separately. Included in each graph is the improvement over 50

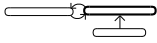
<sup>3</sup>For a given lesion, error collection requires a single (forward) pass through the network for each of the 40 words. Retraining requires, for each of 50 sweeps, 2 passes (forward and backward) for each of the 20 retrained words, 1 (forward) pass for each of the 20 unretrained words, run 2 times with the word sets exchanged, plus 2 (forward and backward) passes for each of the 40 words when retrained together, for a total of 10,000 passes.



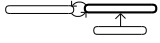

Retraining after  $O \Rightarrow I(0.3)$  Lesions

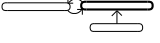


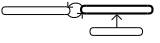
Retraining after  $O \Rightarrow I(0.3)$  Lesions

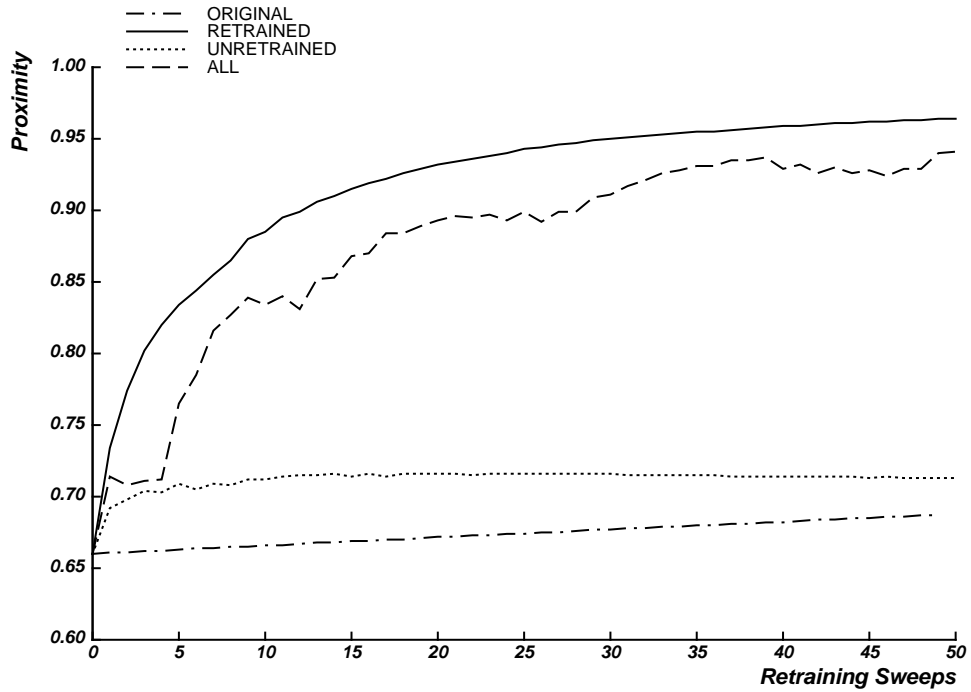
Figure 7.5: Retraining performance, in terms of average proximity (top) and correct performance (bottom), after  $O \Rightarrow I(0.3)$  lesions of the  network. Results are averaged over 20 instances of lesions and across exchanges of the retrained and unretrained word sets. Also included in each figure is the improvement in performance over 50 sweeps during the original learning when it had reached the same level.

sweeps during the original learning when it had reached the same level of performance. Comparing this curve with the relearning curve when all of the words are retrained after damage, relearning is significantly faster than original learning, both in terms of average proximity and correct performance. Retraining on only 20 of the words is faster still. The rapid recovery of performance after lesions replicates the effect Hinton & Sejnowski found when relearning with weights corrupted by noise.

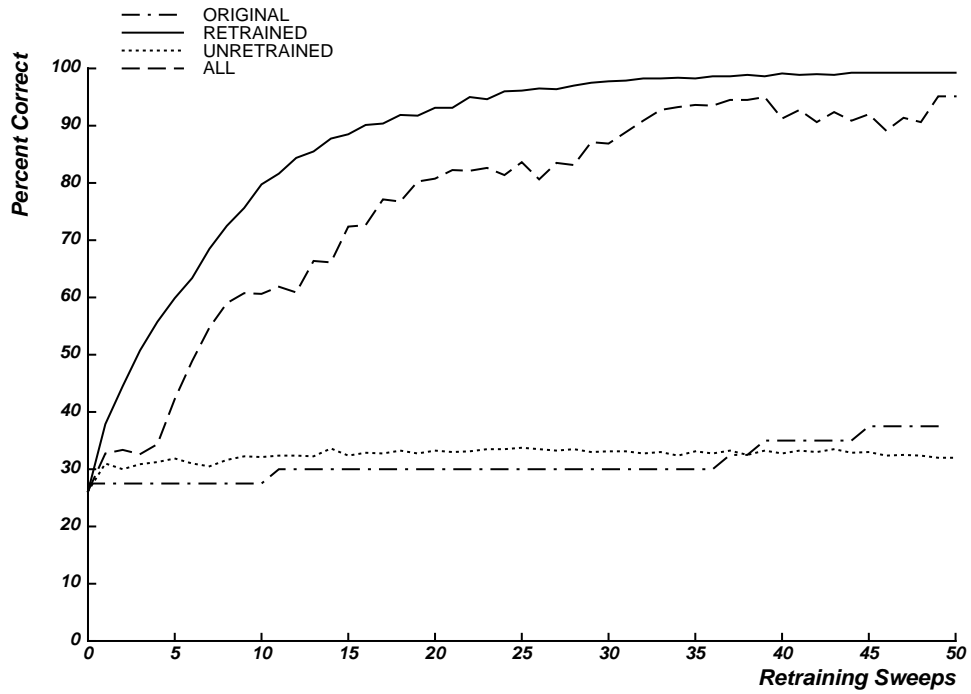
For a given word set, explicit retraining should produce the largest possible improvement in correct performance. Performance on the word set may also improve as a result of generalization when other words are retrained. As a measure of the degree of generalization caused by retraining, we can use the ratio of the improvement in performance when a word set is unretrained vs. when it is explicitly retrained. A generalization ratio of 1.0 would mean that performance improved on the words as much when retraining on the other words as when retraining on the words themselves. Using this measure, there is no evidence of generalization to the unretrained words when relearning after  $0 \Rightarrow I(0.3)$  in the  network—if anything, average correct performance on these words shows a trend towards getting slightly worse (mean generalization:  $-0.024$ ,  $t(39) = 1.17$ ,  $p = .25$ ).  $0 \Rightarrow I(0.3)$  lesions of the  network (26.3% correct), shown in Figure 7.6, produce similar results. Thus, the transfer effects previously found by Hinton & Sejnowski and Hinton & Plaut after adding noise to weights in networks have not been replicated in two of the deep dyslexia networks after lesions to the  $0 \Rightarrow I$  connections.

The data presented above is averaged over 20 instances of each lesion. It may still be the case that relearning after some of these lesions does show some degree of transfer. To check this possibility, as well as to convey a sense of the variability in recovery after particular lesions of the same severity, Figures 7.7 to 7.10 present the retraining results for correct performance for each of the 20 individual lesions to the  network. As can be seen from the figures, there is no significant transfer from the retrained to unretrained words for any of the lesions, although the degree of relearning when retraining on the full word set varies considerably from lesion to lesion. The learning curves are quite noisy because the magnitude of the initial gradient after damage is very large, causing the network to “overshoot” the bottom of the ravine in weight space. Repeated overshooting is reflected in oscillations across the ravine. However, the general trends shown in the averaged data—rapid relearning after damage but no generalization to unretrained words—are reflected in the graphs for individual lesions.


To demonstrate that the lack of generalization is not simply due to unstable relearning, the retraining experiment with  $0 \Rightarrow I(0.3)$  lesions of the  network was re-run using a reduced learning rate (half of the original rate). The smaller rate makes the network more conservative in the steps it takes in weight space, reducing the degree of overshooting when the gradient is large. As Figure 7.11 shows, more conservative relearning smoothes the learning curves considerably but continues to yield no transfer.



Retraining after  $O \Rightarrow I(0.3)$  Lesions



Retraining after  $O \Rightarrow I(0.3)$  Lesions

Figure 7.6: Retraining performance, in terms of average proximity (top) and correct performance (bottom), after  $O \Rightarrow I(0.3)$  lesions of the  network. Results are averaged over 20 instances of lesions and across exchanges of the retrained and unretrained word sets.

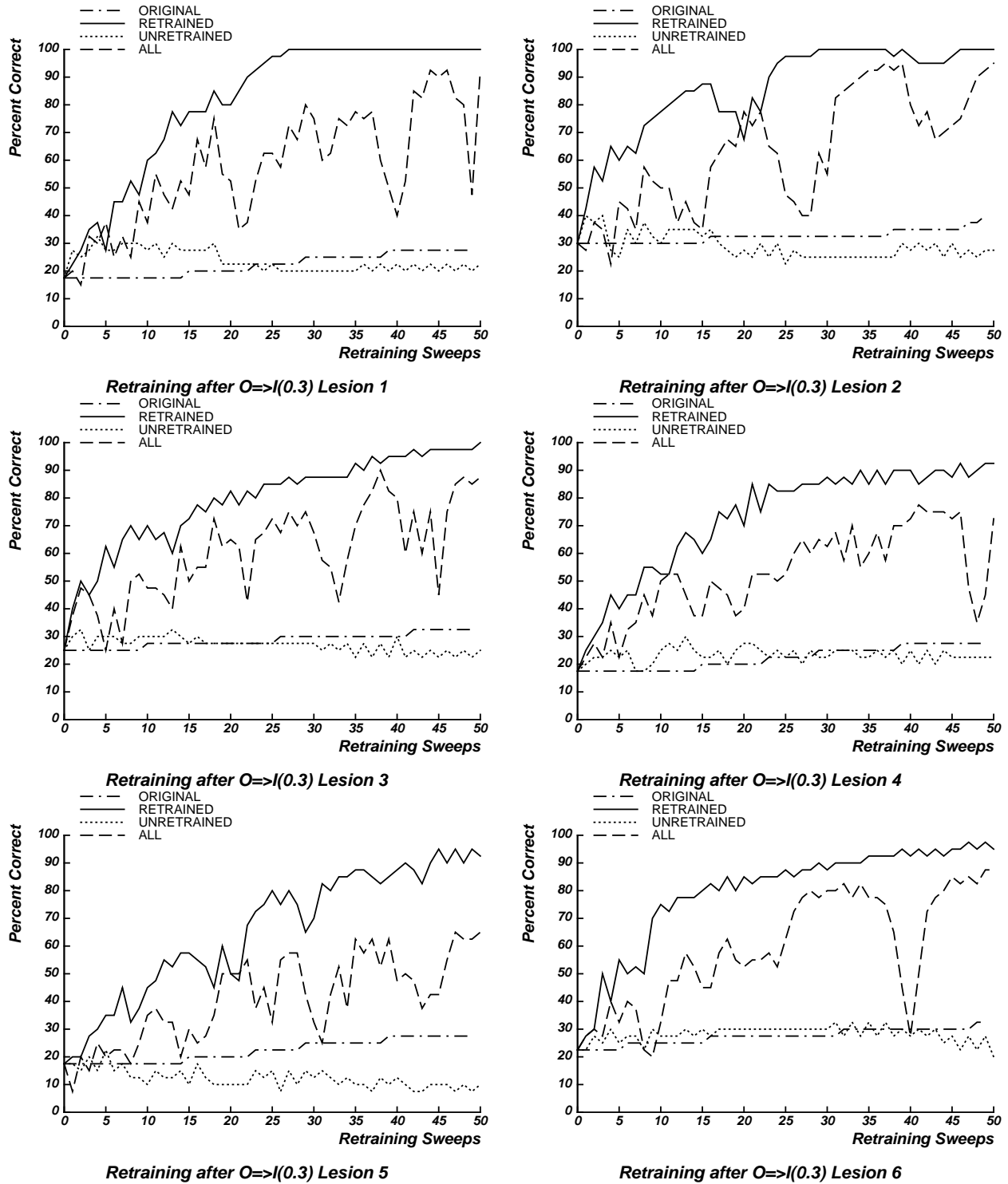
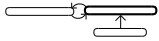


Figure 7.7: Retraining performance after  $O \Rightarrow I(0.3)$  lesions 1–6 of the  network.

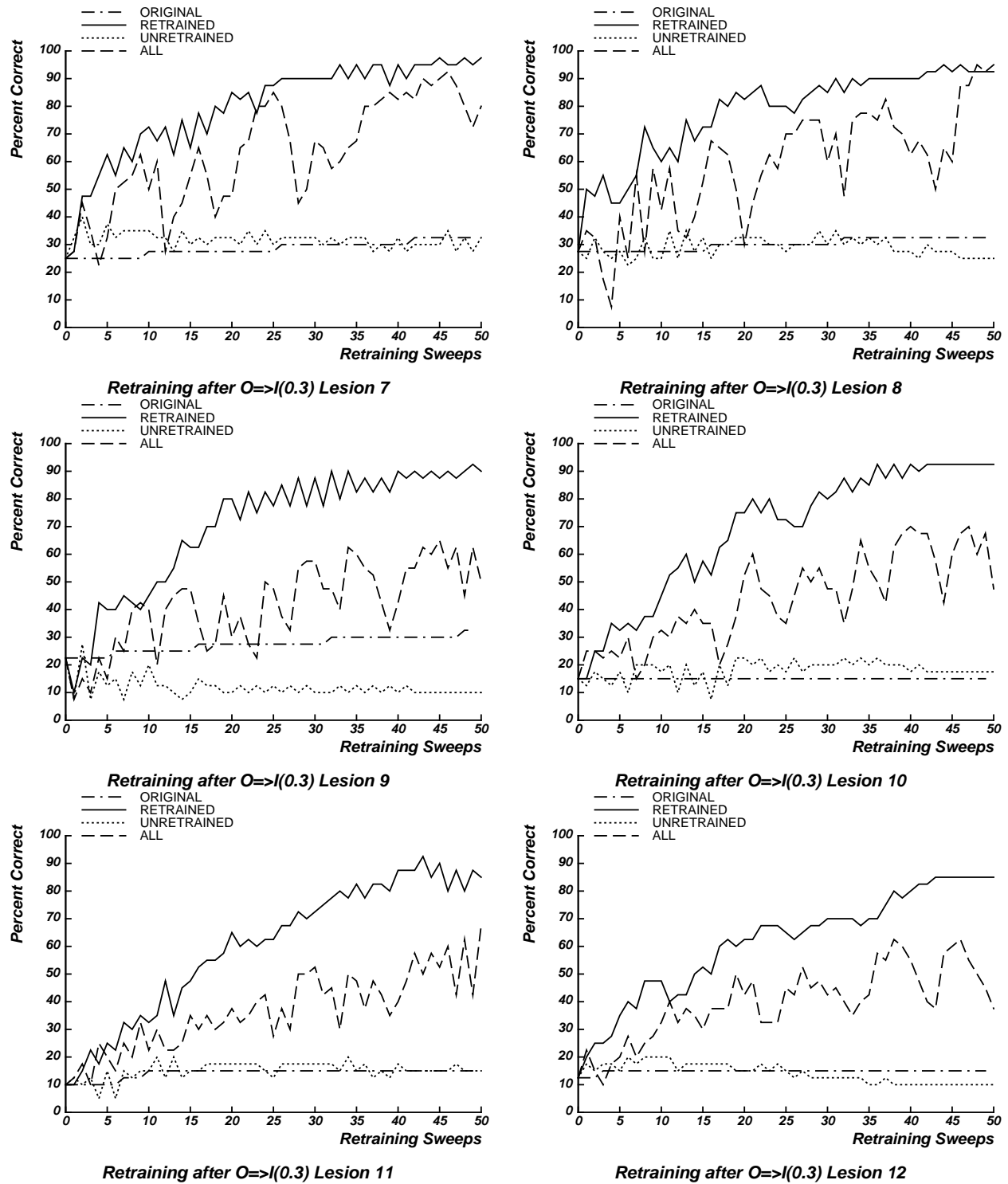
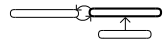


Figure 7.8: Retraining performance after  $O \Rightarrow I(0.3)$  lesions 7–12 of the  network.



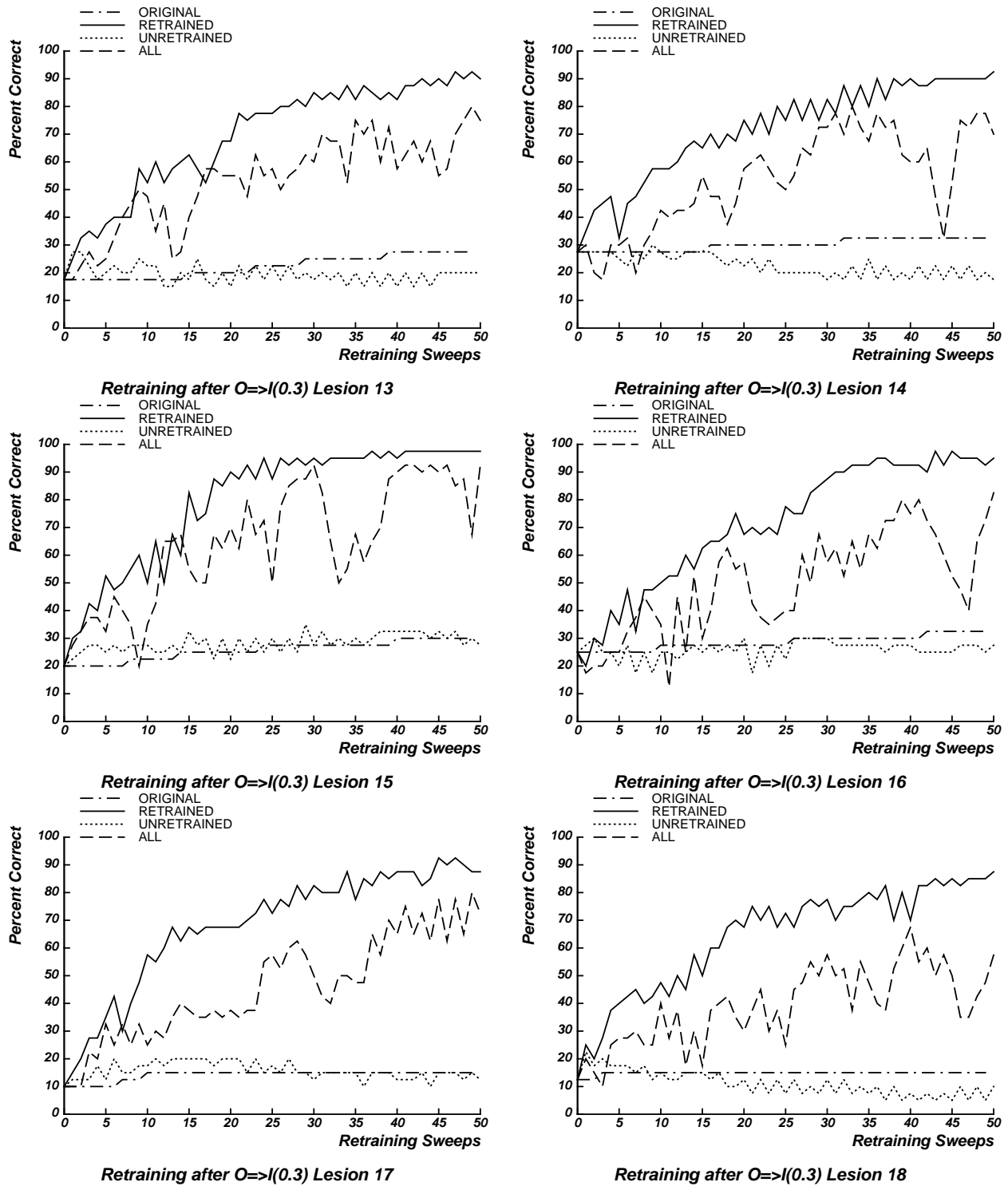
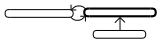


Figure 7.9: Retraining performance after  $O \Rightarrow I(0.3)$  lesions 13–18 of the  network.

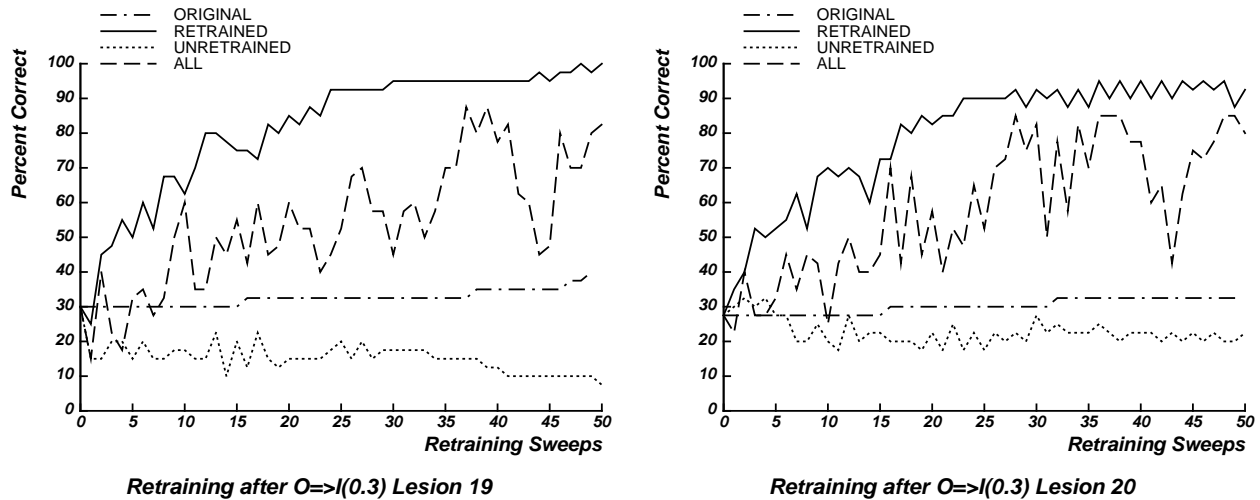
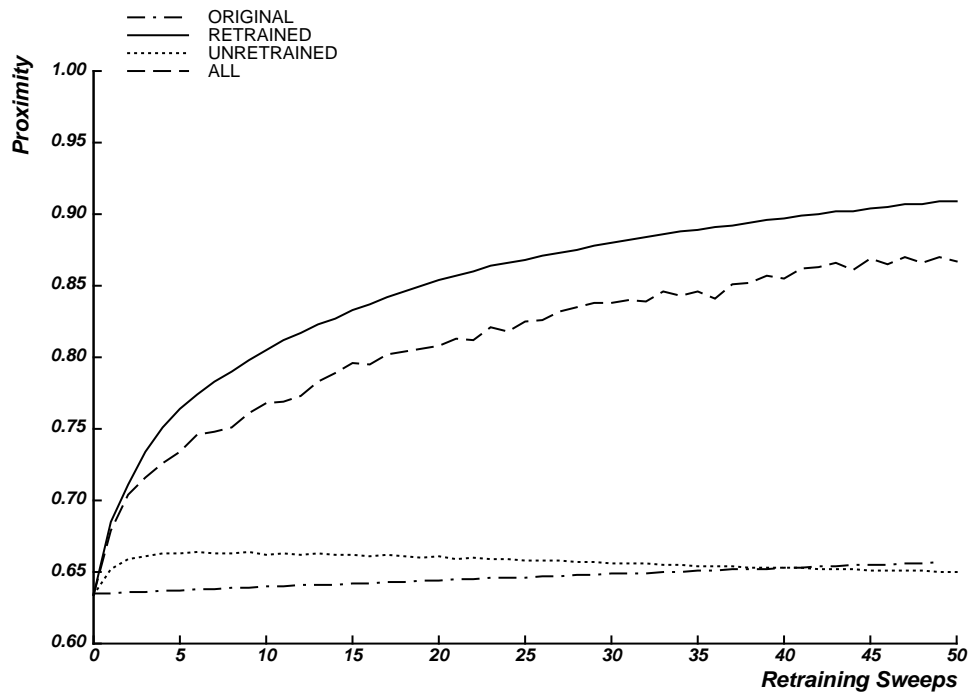


Figure 7.10: Retraining performance after  $O \Rightarrow I(0.3)$  lesions 19 and 20 of the network.

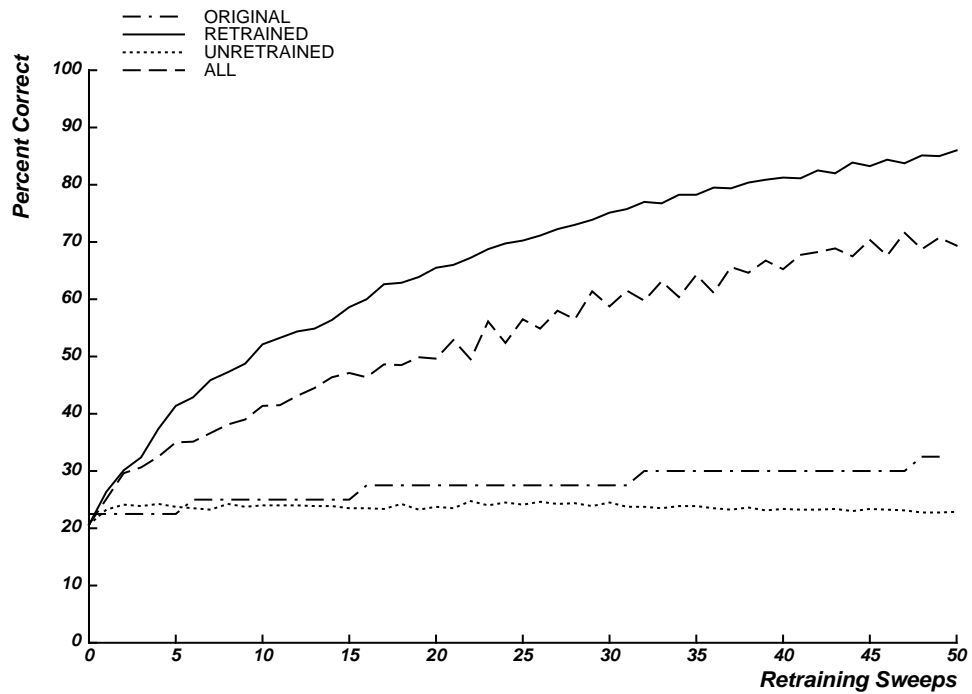
In the investigations of various network architectures in Chapter 4,  $I \Rightarrow S$  lesions generally produced similar patterns of errors as  $O \Rightarrow I$  lesions, although error rates were lower. One difference that was found is that networks in which the intermediate units are not involved in developing attractors, such as the network, tend to be more sensitive to  $I \Rightarrow S$  than  $O \Rightarrow I$  lesions. Another difference is that  $I \Rightarrow S$  lesions tend to produce stronger semantic influences in explicit error responses than do  $O \Rightarrow I$  lesions. These influences might also effect the degree of relearning and transfer after damage. To investigate this possibility, the relearning procedure was carried out after  $I \Rightarrow S$  lesions in the ; the corresponding data for  $I \Rightarrow S$  lesions in the network will be presented in the next section when within-attractor lesions are considered.

Figure 7.12 presents the retraining results for  $I \Rightarrow S(0.3)$  lesions of the network (23.9% correct performance), averaged over all 20 lesion instances and over exchanges of the retrained and unretrained word sets. Comparing with the corresponding results for  $O \Rightarrow I$  lesions (Figure 7.5, p. 190), correct performance improves more quickly and to a greater degree when relearning after  $I \Rightarrow S$  lesions when retraining on all 40 words. More importantly, as performance improves on the retrained set of words there is modest but significant generalization to the unretrained set (mean generalization in correct performance: 0.26,  $t(39) = 10.0, p < .001$ ). The results for individual lesions again show considerable variability in the rate and degree of relearning and transfer (see Figures 7.13 to 7.16). An average improvement in correct performance of around 20% on the unretrained words is not nearly as impressive as the amount of generalization found by Hinton & Plaut, but it is encouraging.

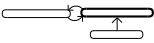
Previous simulations have demonstrated a difference in both qualitative and quantitative behavior between lesions prior to the operation of attractors, and lesions to connections directly involved in implementing the attractors. The most notable of these was a double dissociation of correct performance on concrete vs. abstract words after direct vs. clean-up lesions (see Chapter 6).



Slow Retraining after  $O \Rightarrow I(0.3)$  Lesions



Slow Retraining after  $O \Rightarrow I(0.3)$  Lesions

Figure 7.11: Retraining performance, in terms of average proximity (top) and correct performance (bottom), after  $O \Rightarrow I(0.3)$  lesions of the  network, when using a learning rate during relearning that is half the original rate. Results are averaged over 20 instances of lesions and across exchanges of the retrained and unretrained word sets.

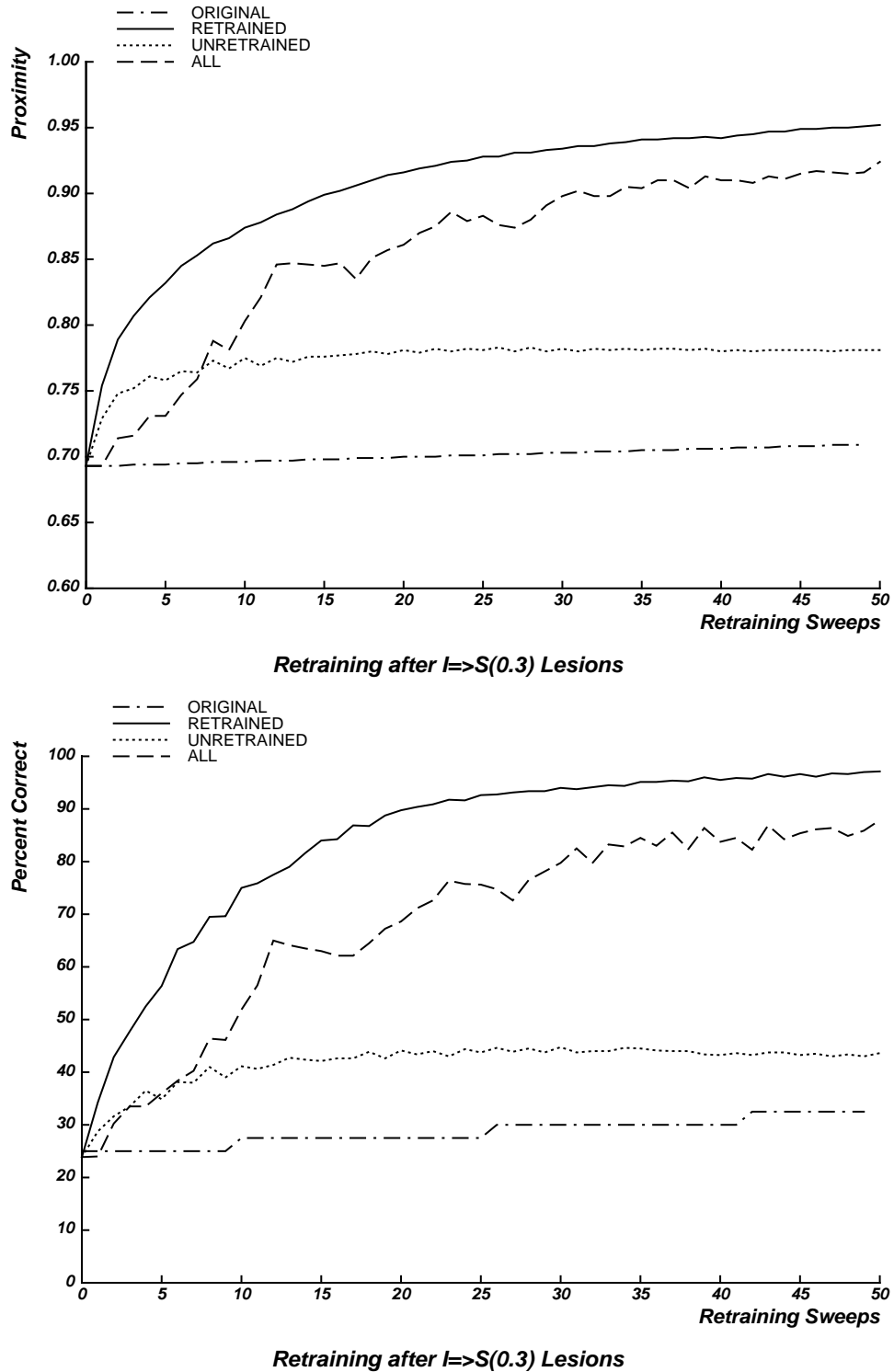
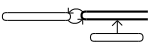


Figure 7.12: Retraining performance, in terms of average proximity (top) and correct performance (bottom), after  $I \Rightarrow S(0.3)$  lesions of the  network. Results are averaged over 20 instances of lesions and across exchanges of the retrained and unretrained word sets.

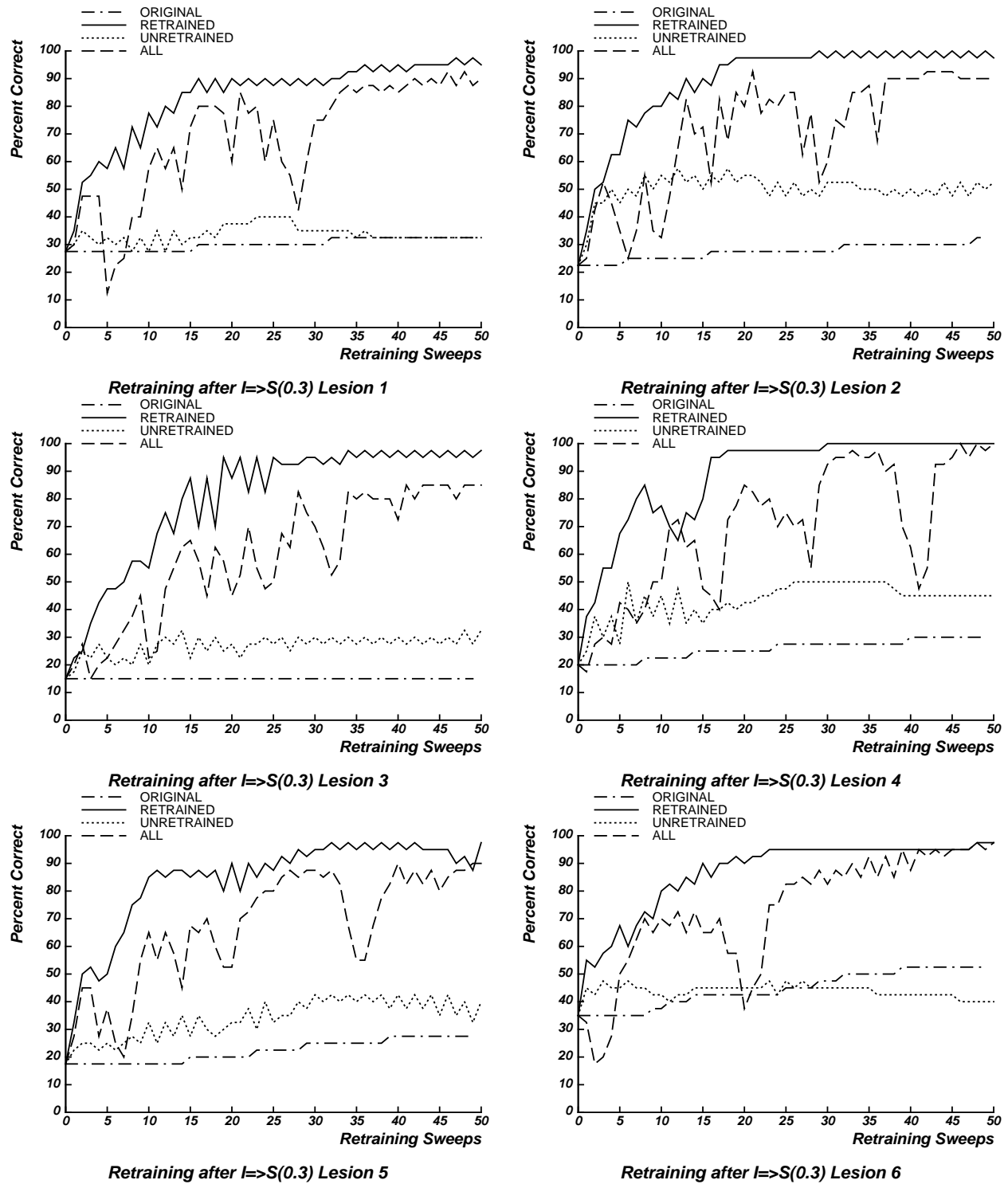
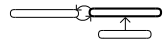


Figure 7.13: Retraining performance after  $I \Rightarrow S(0.3)$  lesions 1–6 of the  network.

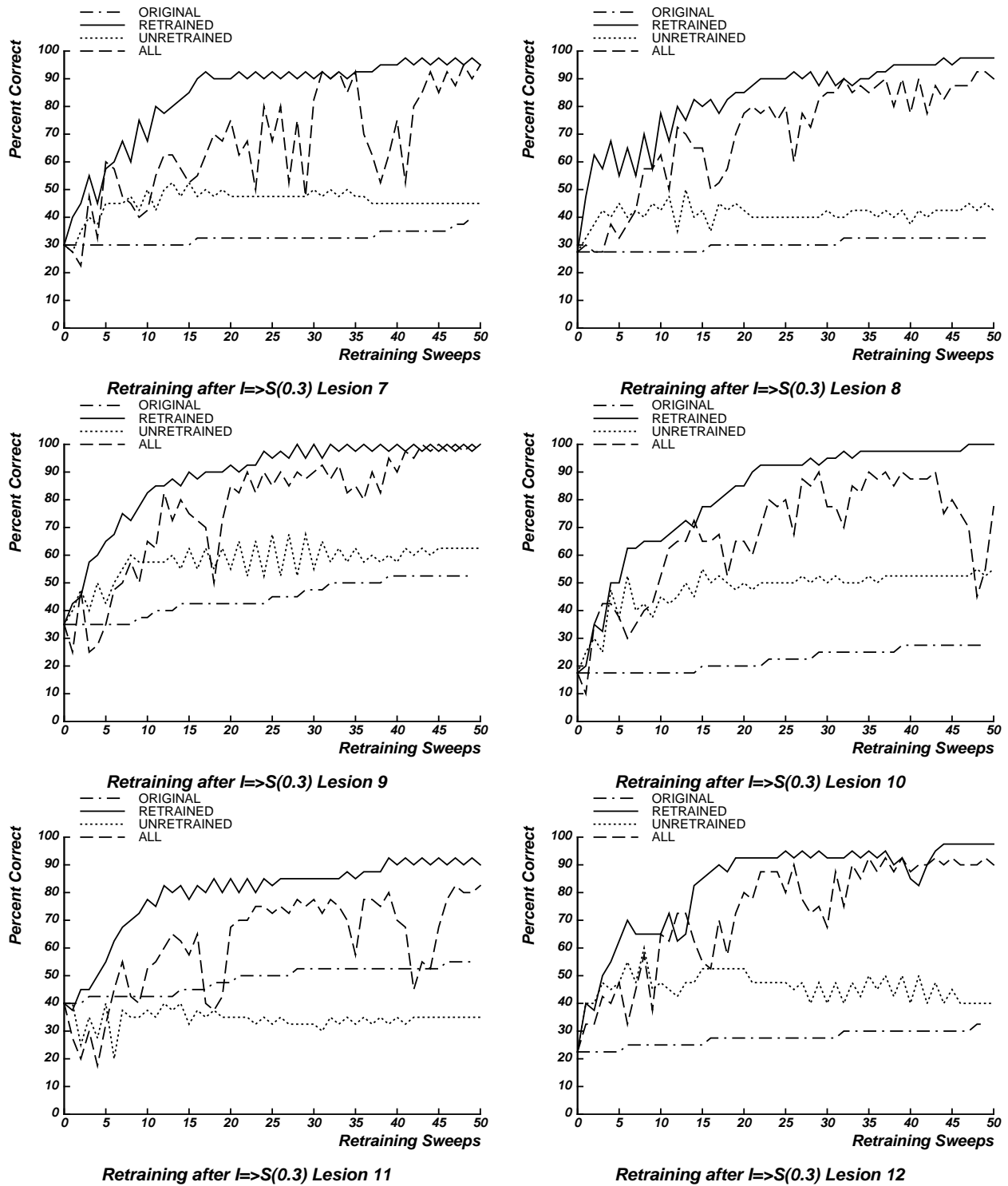
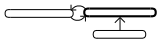


Figure 7.14: Retraining performance after  $I \Rightarrow S(0.3)$  lesions 7–12 of the  network.

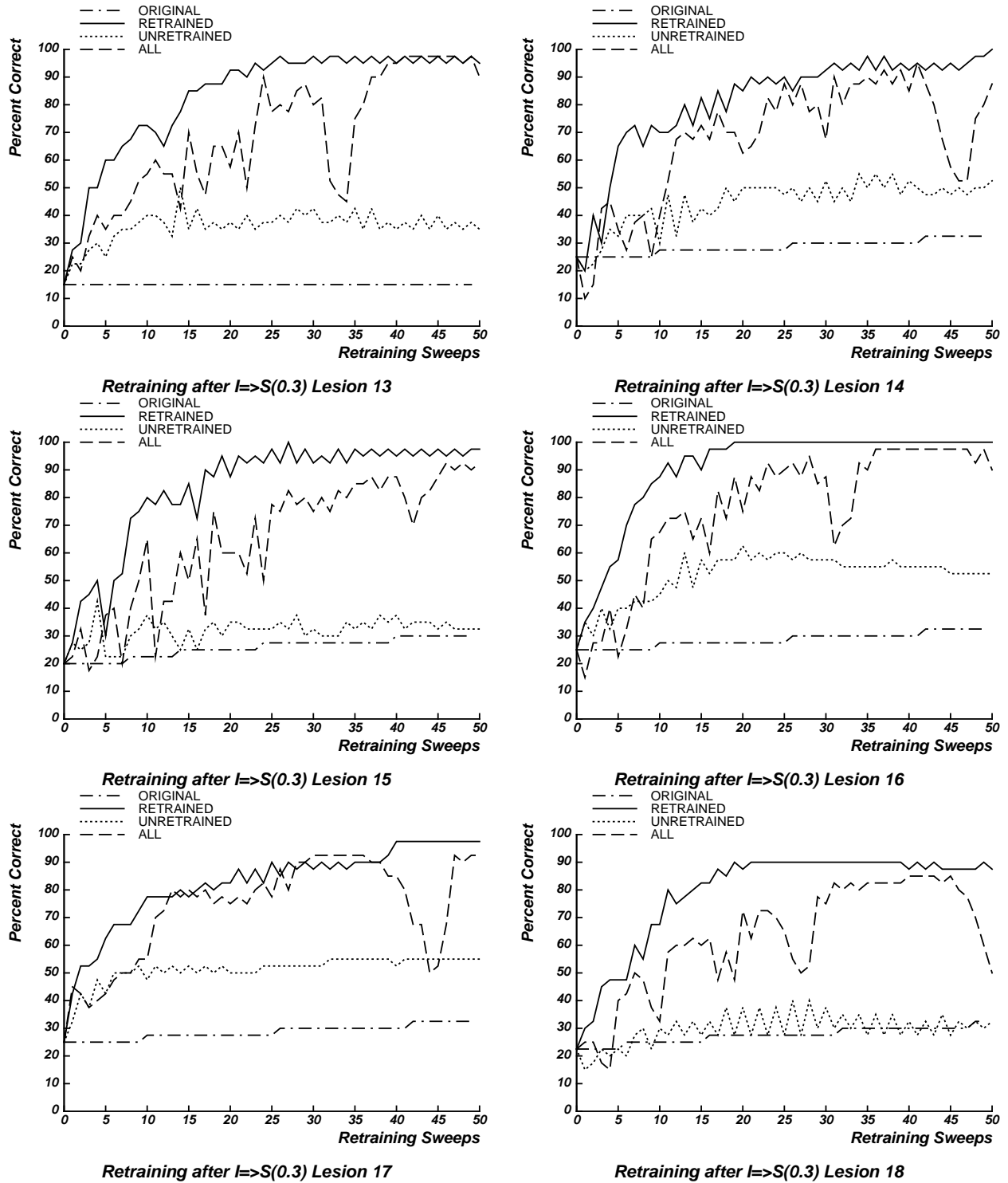
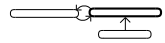


Figure 7.15: Retraining performance after I⇒S(0.3) lesions 13–18 of the  network.

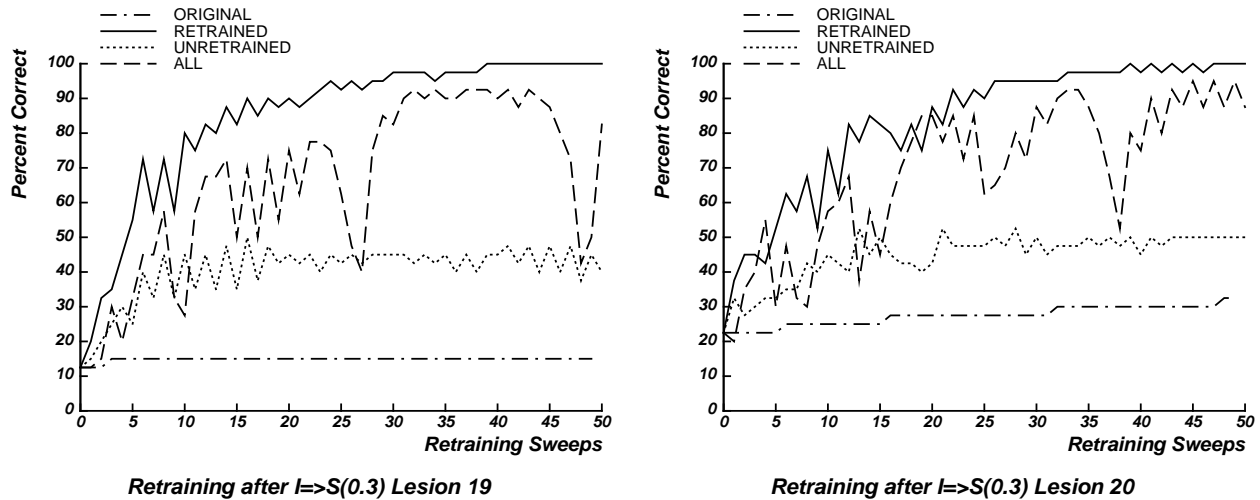
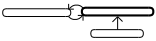
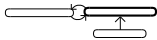


Figure 7.16: Retraining performance after  $I \Rightarrow S(0.3)$  lesions 19 and 20 of the  network.

These observed differences suggest that lesions to these two pathways may cause the network to behave qualitatively different in other respects as well. For this reason, it is important to investigate relearning after within-attractor lesions before making any general conclusions.

### 7.3.5 Within-attractor lesions

Considering the  network first, the most severe  $S \Rightarrow C$  lesions administered (0.7) only lowered correct performance to 32%. Thus we restrict our consideration to  $C \Rightarrow S$  lesions in this network. Figure 7.17 presents the retraining results averaged over all 20  $C \Rightarrow S(0.5)$  lesions, which reduced average correct performance to 20.3%. In contrast to  $O \Rightarrow I$  lesions, relearning after  $C \Rightarrow S$  lesions shows considerable transfer from the retrained to unretrained word sets, both in average proximity (mean generalization 0.56,  $t(39) = 55.2, p < .001$ ) and in correct performance (mean generalization 0.61,  $t(39) = 28.1, p < .001$ ). Correct performance on the unretrained words is more than doubled even though these words are never presented to the damaged network. In fact, relearning on all of the words is quite dramatic, with performance becoming essentially perfect after 50 sweeps.

It is also interesting that, on average, relearning on all 40 words is slightly faster than learning on only 20 words. This was not true after  $O \Rightarrow I$  lesions where no generalization was observed. This makes sense if the error gradients for words are more consistent after  $C \Rightarrow S$  lesions than after  $O \Rightarrow I$  lesions. The actual weight change made by the network after a retraining sweep is the vector sum of the weight changes dictated by each individual retrained word (scaled by the learning rate, which was the same when retraining on 20 or 40 words). The cosine of the angle (proximity) between this actual weight change vector and that for a particular word is an approximate measure of the degree to which the weight change will improve performance on that word. If the vectors for the set of retrained words point in different directions, their average proximity with the vector sum



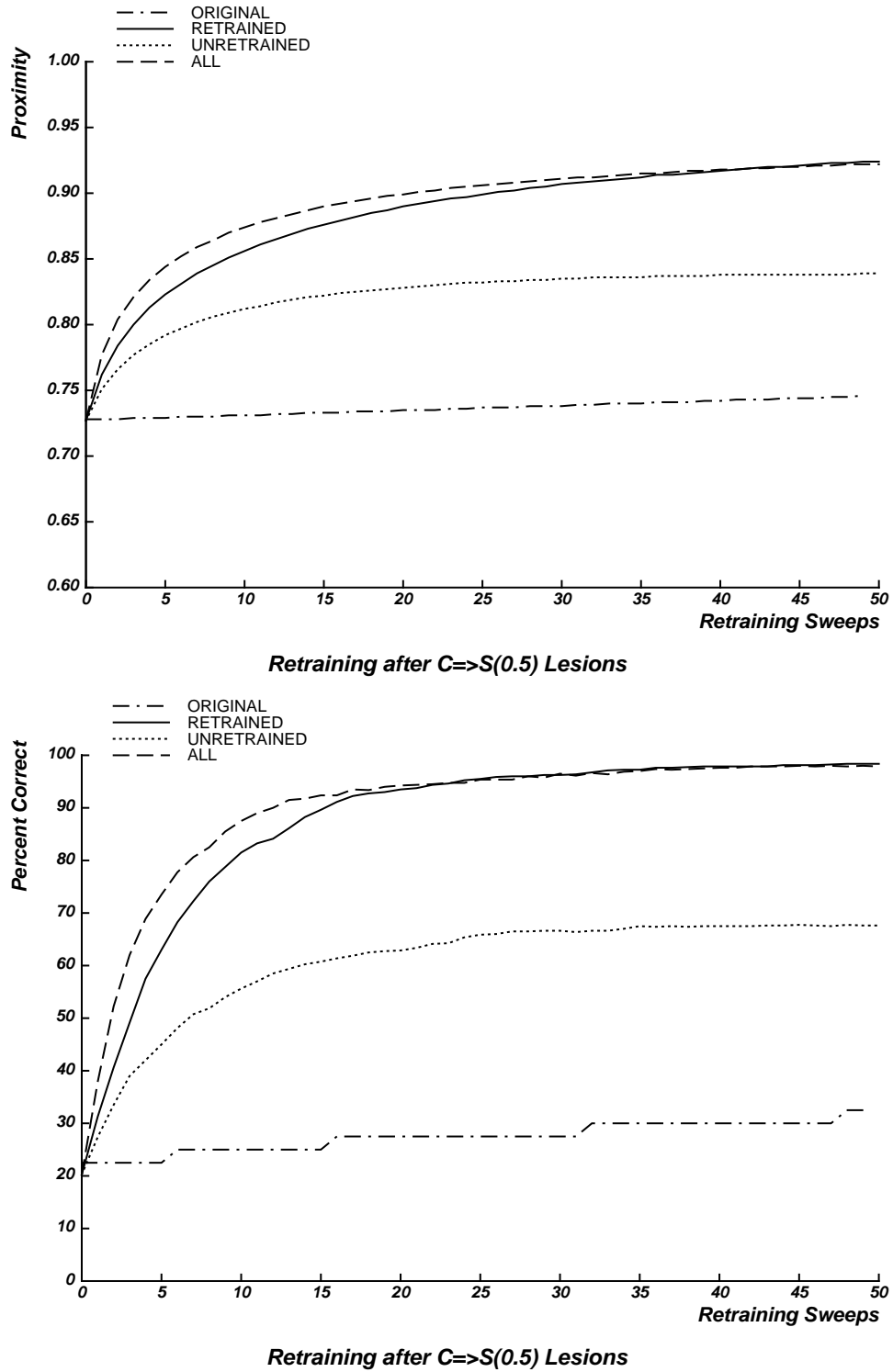
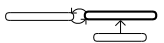
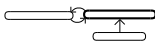


Figure 7.17: Retraining performance, in terms of average proximity (top) and correct performance (bottom), after  $C \Rightarrow S(0.5)$  lesions of the  network, averaged over 20 instances of lesions and across exchanges of the retrained and unretrained word sets.

will be low, and will decrease as the number of vectors (retrained words) increases. Thus average performance when relearning 20 words will be faster than when relearning 40 words. However, if the vectors for words are consistent, their vector sum will have a high proximity with each of them, as well as a large magnitude. In this case, relearning on more words will be faster. Greater consistency of the weight change vectors for words after  $C \Rightarrow S$  lesions than after  $O \Rightarrow I$  lesions also explains why generalization to unretrained words occurs after the former but not the latter: the direction of the vector sum for some of the words will better approximate the one for all of the words if the vectors are consistent.

The finding of considerable transfer for  $C \Rightarrow S$  lesions makes it even more important than for  $O \Rightarrow I$  or  $I \Rightarrow S$  lesions to determine the degree of variability in this effect across individual lesions. Perhaps the transfer is very strong in only a few of the lesions, but absent in most. Figures 7.18 to 7.21 present the retraining results for correct performance for each of the 20 individual  $C \Rightarrow S(0.5)$  lesions to the  network. In fact, quite the opposite is true—the transfer effect is substantial for all 20 lesions, although there is still considerable variability in its magnitude. For instance, correct performance on the unretrained set rises from 20% to near 80% after Lesion 2 (0.78 generalization), while after Lesion 9 it only rises from 40% to just over 60% (0.48 generalization).

Also notice that the learning curves are much more well-behaved than those after  $O \Rightarrow I$  lesions. This is not because the error gradients after clean-up lesions are smaller than those after direct pathway lesions, because relearning is actually *faster* after clean-up lesions. Rather, the dimensions of weight space that correspond to the direct pathway are more highly constrained—the ravines are narrower and turn more quickly. Thus the same-sized gradient will cause overshooting during relearning after a direct pathway lesion, but will produce smooth relearning after a clean-up lesions. This makes sense if we remember that the direct pathway must mediate between two *arbitrarily-related* domains—orthography and semantics—while the clean-up pathway operates within a single domain. The constraints on the direct pathway are harder for connectionist mechanisms to satisfy than those on the clean-up pathway.

The curves for the unretrained set for a few of the lesions (e.g. 8, 13) show the effects of *overlearning*—once performance on the retrained set becomes near perfect, continued retraining on this set slightly degrades the previous recovery of performance on the unretrained set. Overlearning is a standard problem when the information in the available training examples underconstrains the parameters of an optimization procedure (such as learning in a connectionist network). An operational approach for preventing overlearning, known as “cross-validation” (Morgan & Bourlard, 1990), is to observe performance on a set of examples drawn from the same distribution as the trained examples but not actually used directly in learning. Training is halted when performance on the *untrained* set peaks. Cross-validation may be a useful technique in patient therapy if the goal is to maximize overall performance within an entire domain rather than just on the specific treated items.

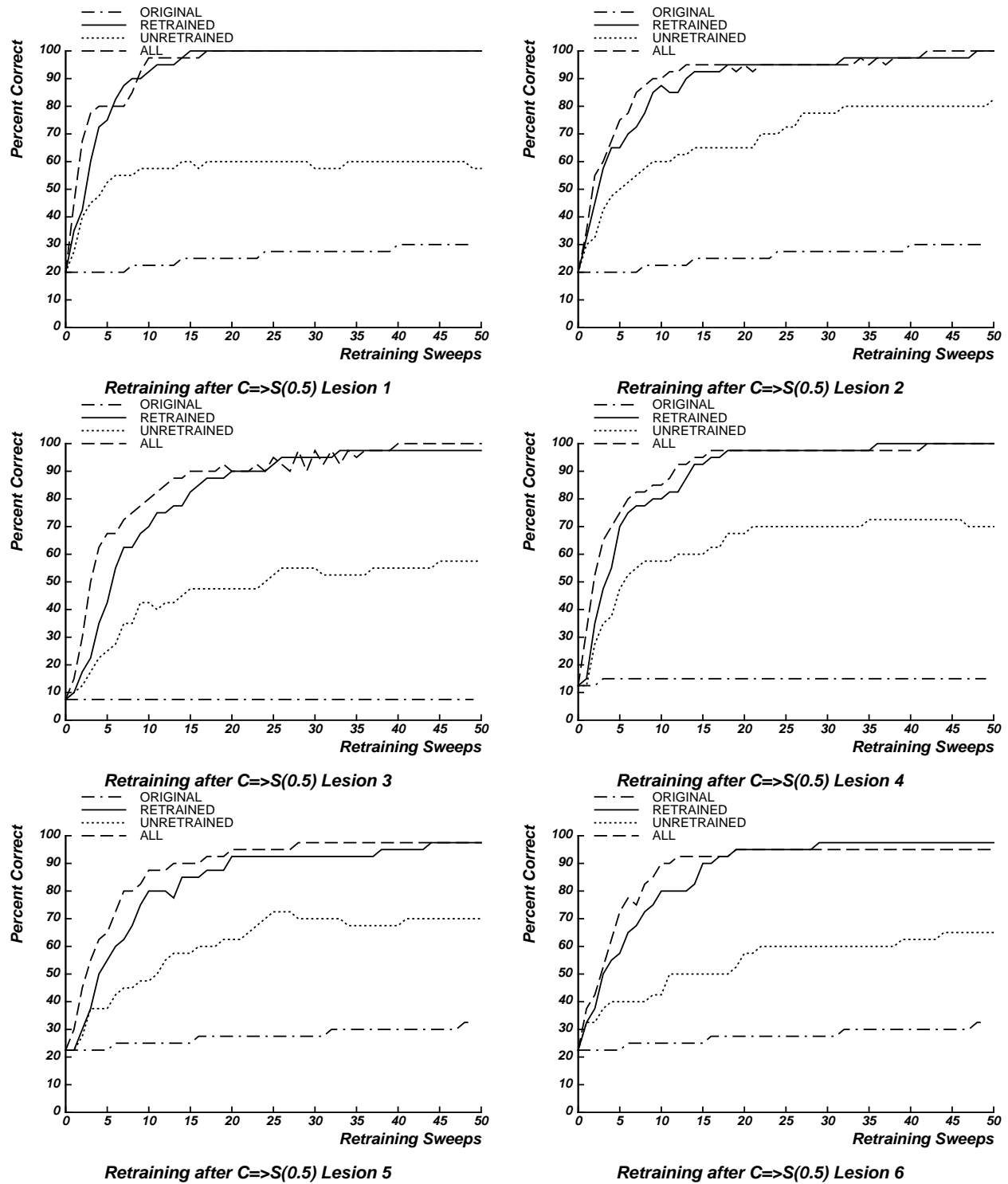
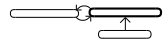


Figure 7.18: Retraining performance after  $C \Rightarrow S(0.5)$  lesions 1–6 of the  network.

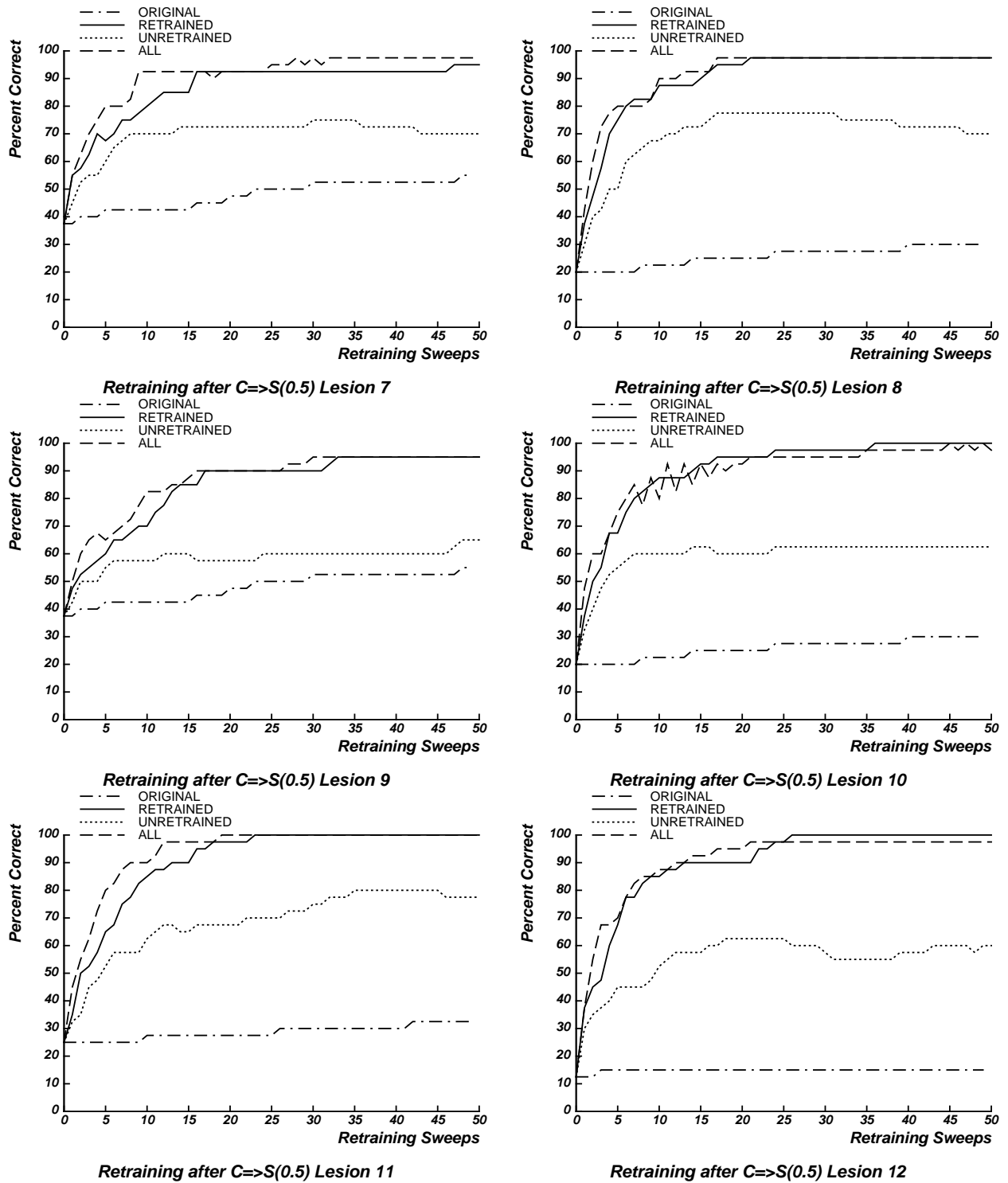
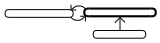


Figure 7.19: Retraining performance after C=>S(0.5) lesions 7–12 of the  network.

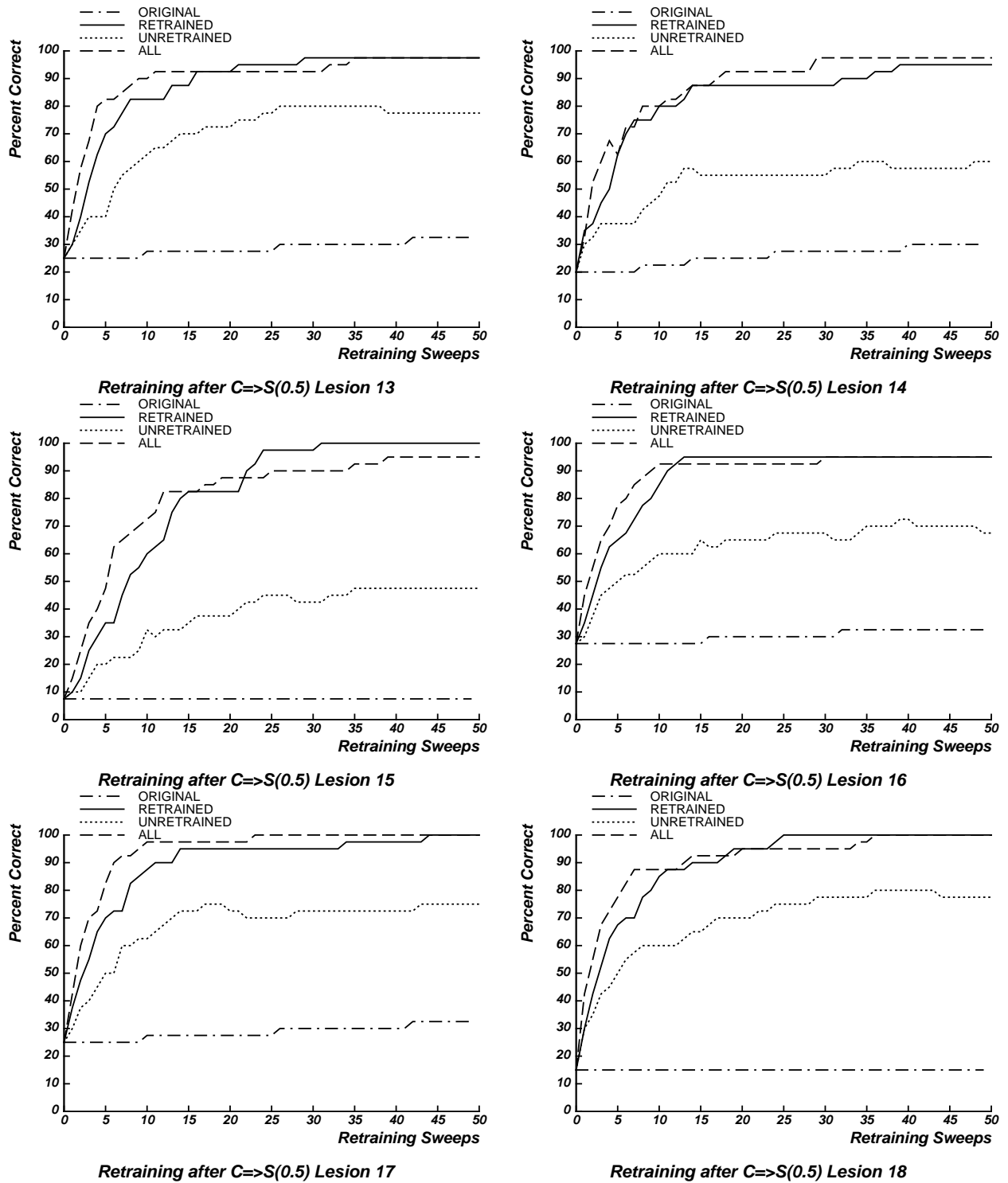
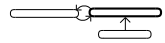


Figure 7.20: Retraining performance after  $C \Rightarrow S(0.5)$  lesions 13–18 of the  network.

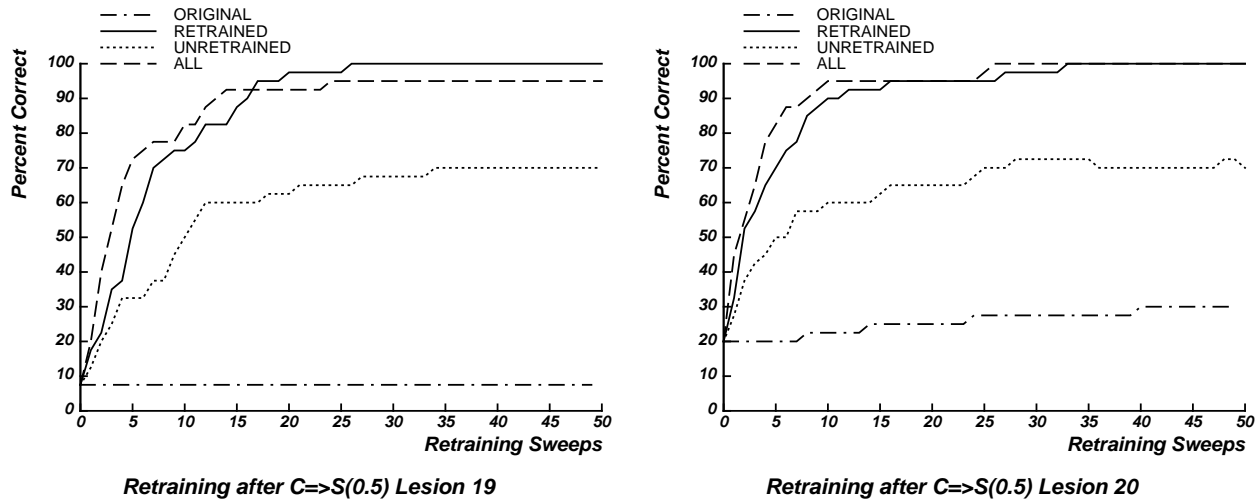
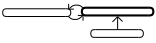


Figure 7.21: Retraining performance after  $C \Rightarrow S(0.5)$  lesions 19 and 20 of the  network.


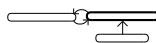

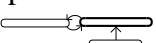
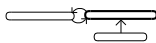

Turning to the  network, within-attractor lesions in this network involve  $I \Rightarrow S$  and  $S \Rightarrow I$  connections. The most appropriate levels of severity for lesions of these connections produce slightly higher levels of correct performance than the lesions used for relearning in the  network: 23.6% correct after  $I \Rightarrow S(0.2)$  lesions, and 34.5% correct after  $S \Rightarrow I(0.7)$  lesions. Nonetheless, the performance levels are close enough to warrant a comparison of relearning effects after within-attractor lesions in the two networks.

Figure 7.22 presents the averaged retraining results for  $I \Rightarrow S(0.2)$  lesions of the  network. Retraining after these lesions produces a considerable degree of generalization to unretrained words. The mean generalization in correct performance is 0.54. This is much greater than for the pre-attractor  $I \Rightarrow S(0.3)$  lesions in the  network (Figure 7.12, p. 198), but not quite as much as the within-attractor  $C \Rightarrow S(0.5)$  lesions (Figure 7.17, p. 203). There is much less generalization after  $S \Rightarrow I(0.7)$  lesions (see Figure 7.23) although it is still significant (mean proximity generalization 0.27,  $t(39) = 15.7, p < .001$ ; mean correct performance generalization 0.25,  $t(39) = 8.84, p < .001$ ). Interestingly, these lesions also show a considerable degree of overlearning. Generalization would be greater (0.39 for correct performance) if retraining were halted after only 20 sweeps. In fact, the degree of overlearning is quite marked for some individual lesions (e.g. 9, 14, see Figures 7.24 to 7.27). The variability in the amount of transfer is also interesting. Some lesions show no transfer during relearning (e.g. 1, 6). Retraining after others shows moderate initial transfer that is lost through overlearning (e.g. 10, 12). Still others show sustained transfer (e.g. 15, 17). It seems that the most consistent characteristic of relearning after lesions to these connections is the occurrence of overlearning, rather than the pattern of generalization *per se*.

In general, relearning after within-attractor lesions in the  and  networks is much faster than original learning and produces significant generalization to unretrained words. In

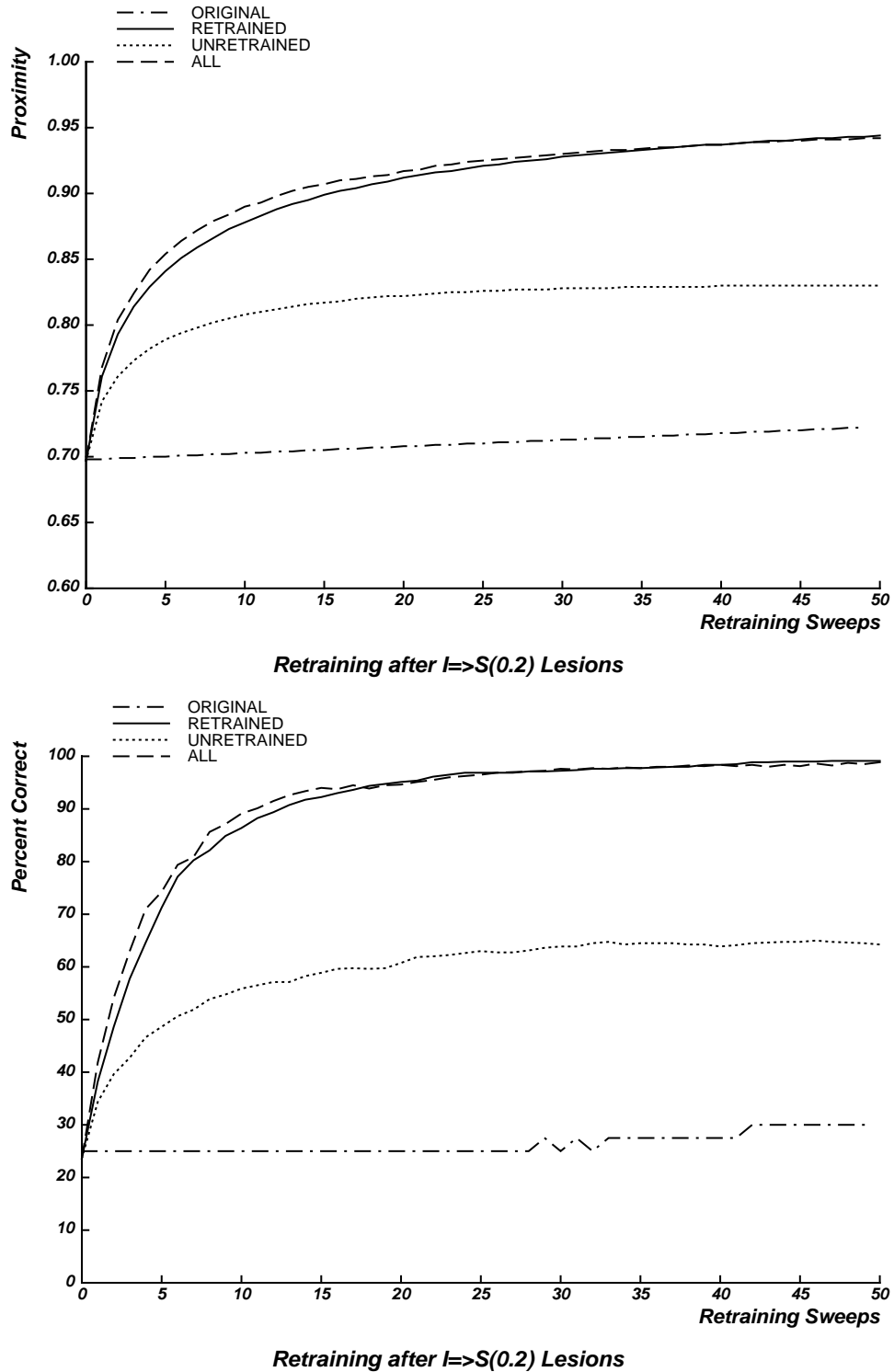



Figure 7.22: Retraining performance, in terms of average proximity (top) and correct performance (bottom), after  $I \Rightarrow S(0.2)$  lesions of the  network, averaged over 20 instances of lesions and across exchanges of the retrained and unretrained word sets.

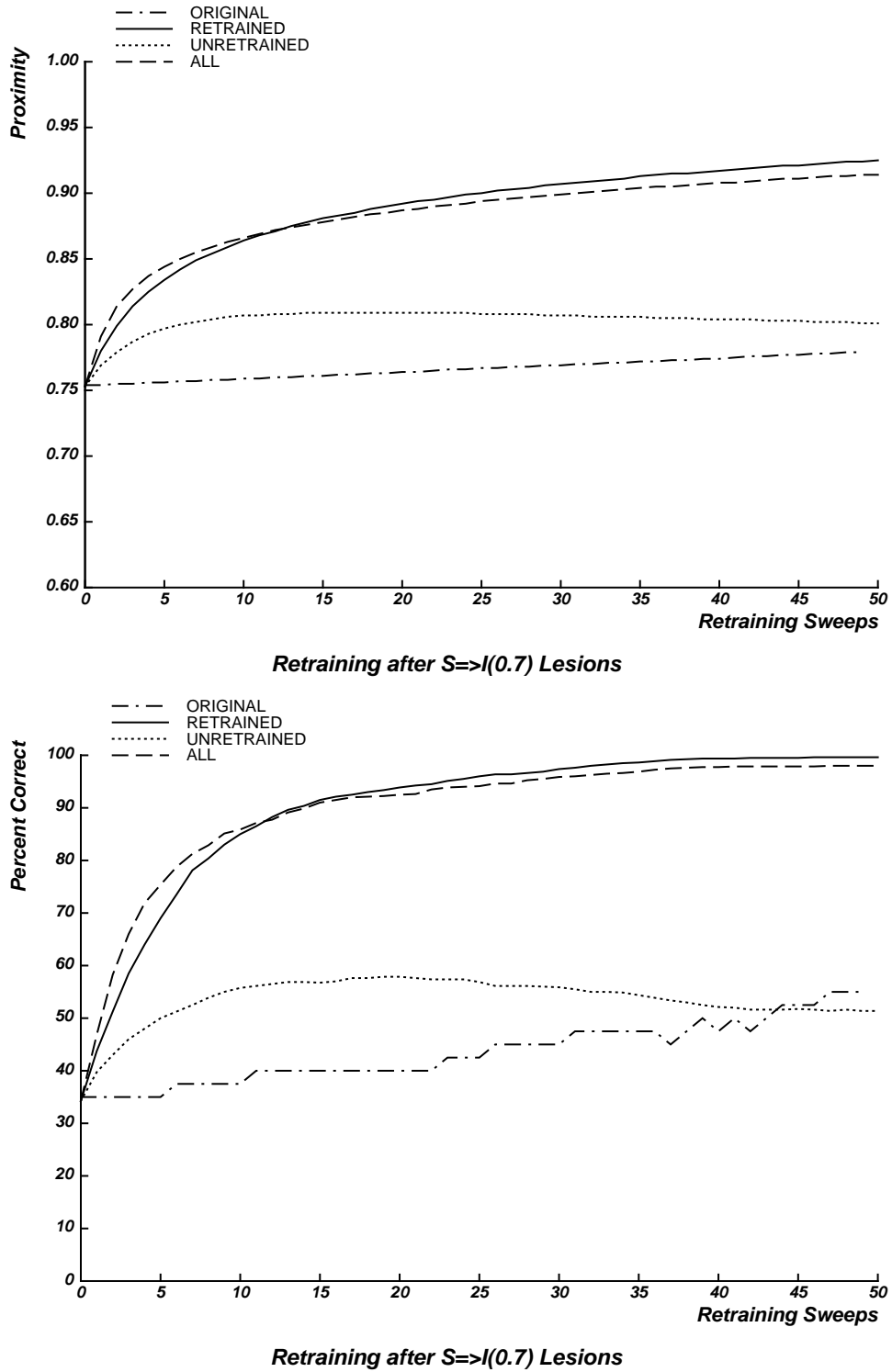



Figure 7.23: Retraining performance, in terms of average proximity (top) and correct performance (bottom), after  $S \Rightarrow I(0.7)$  lesions of the  network, averaged over 20 instances of lesions and across exchanges of the retrained and unretrained word sets.



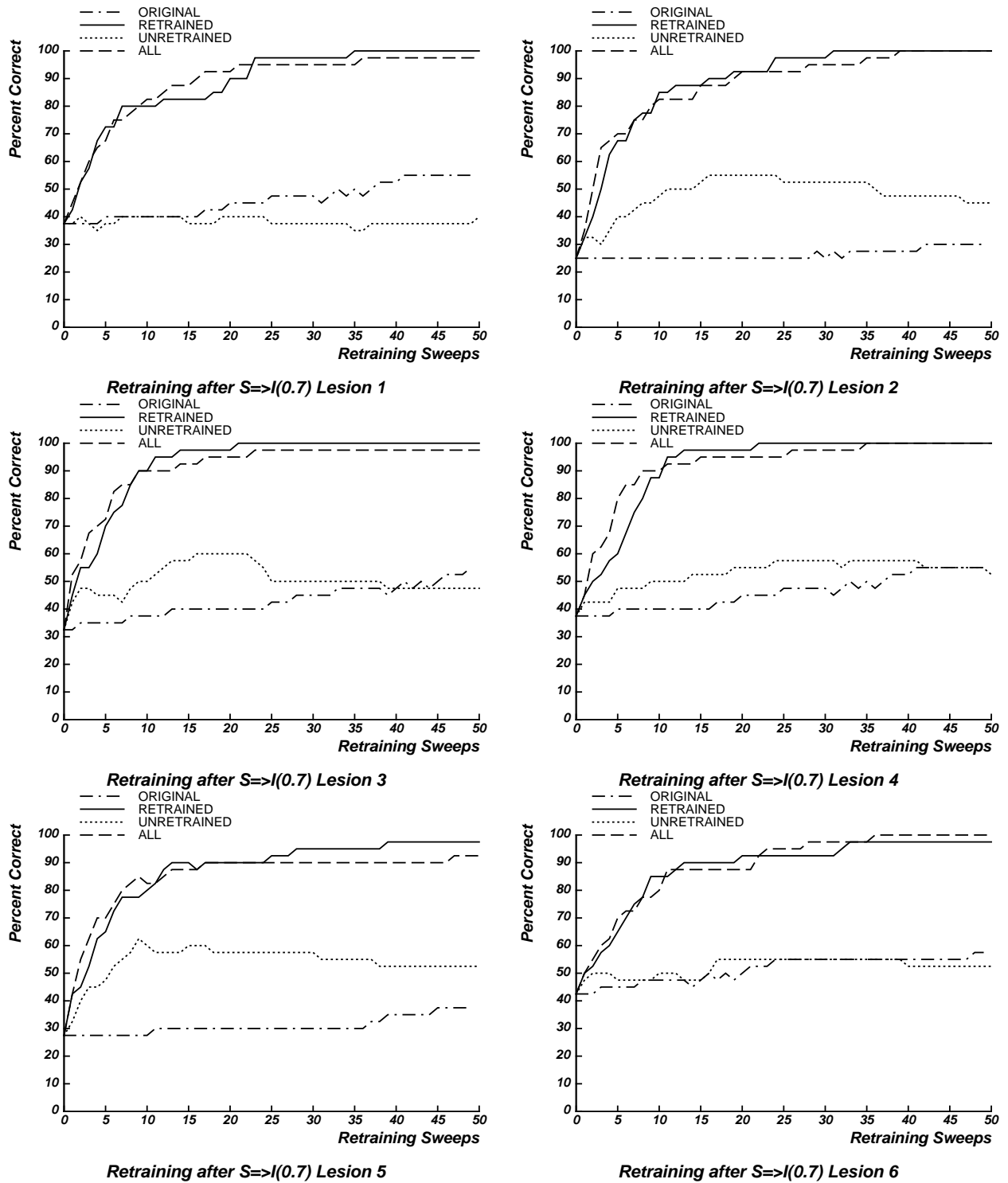



Figure 7.24: Retraining performance after  $S \Rightarrow I(0.7)$  lesions 1–6 of the  network.

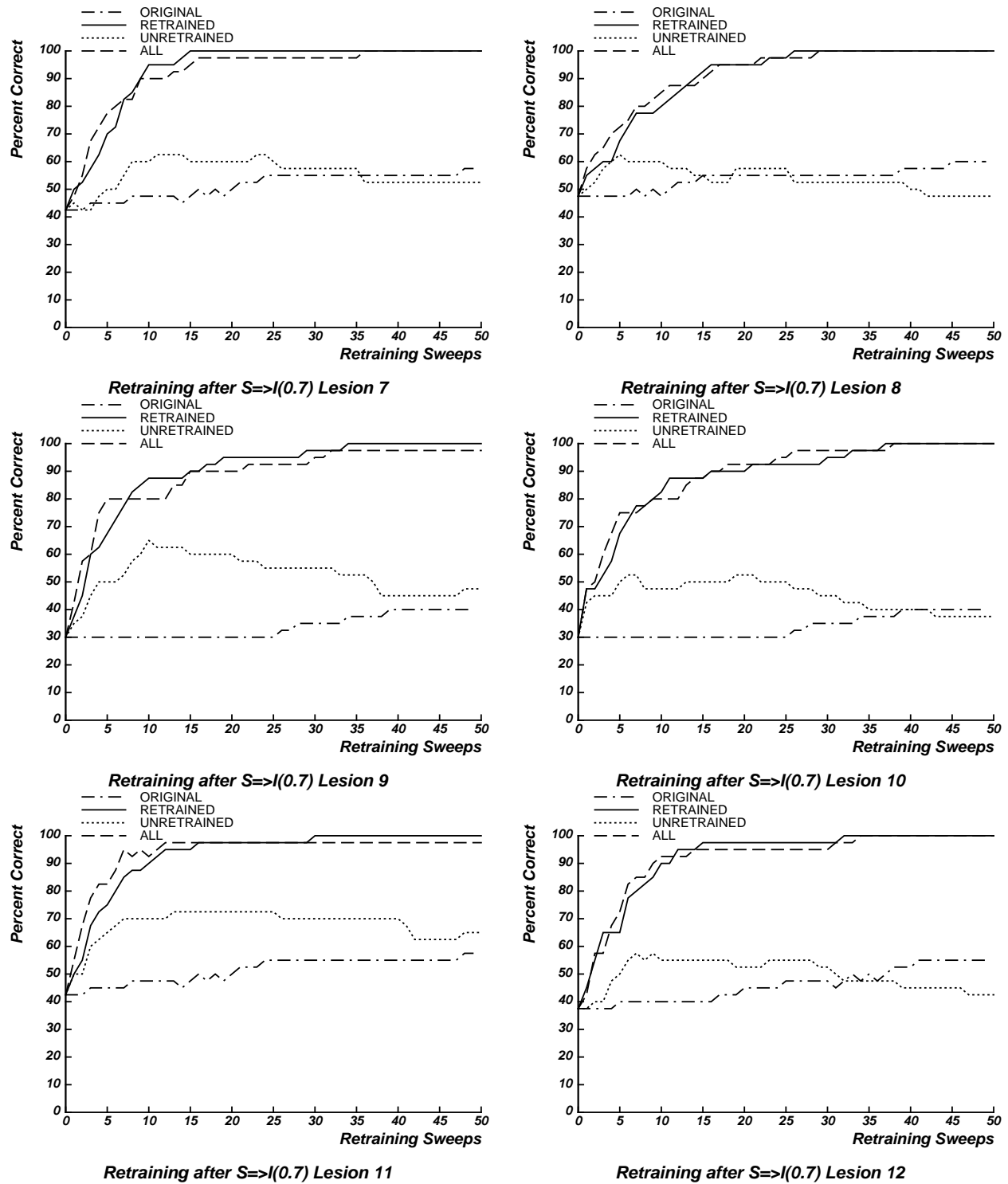



Figure 7.25: Retraining performance after  $S \Rightarrow I(0.7)$  lesions 7–12 of the  network.

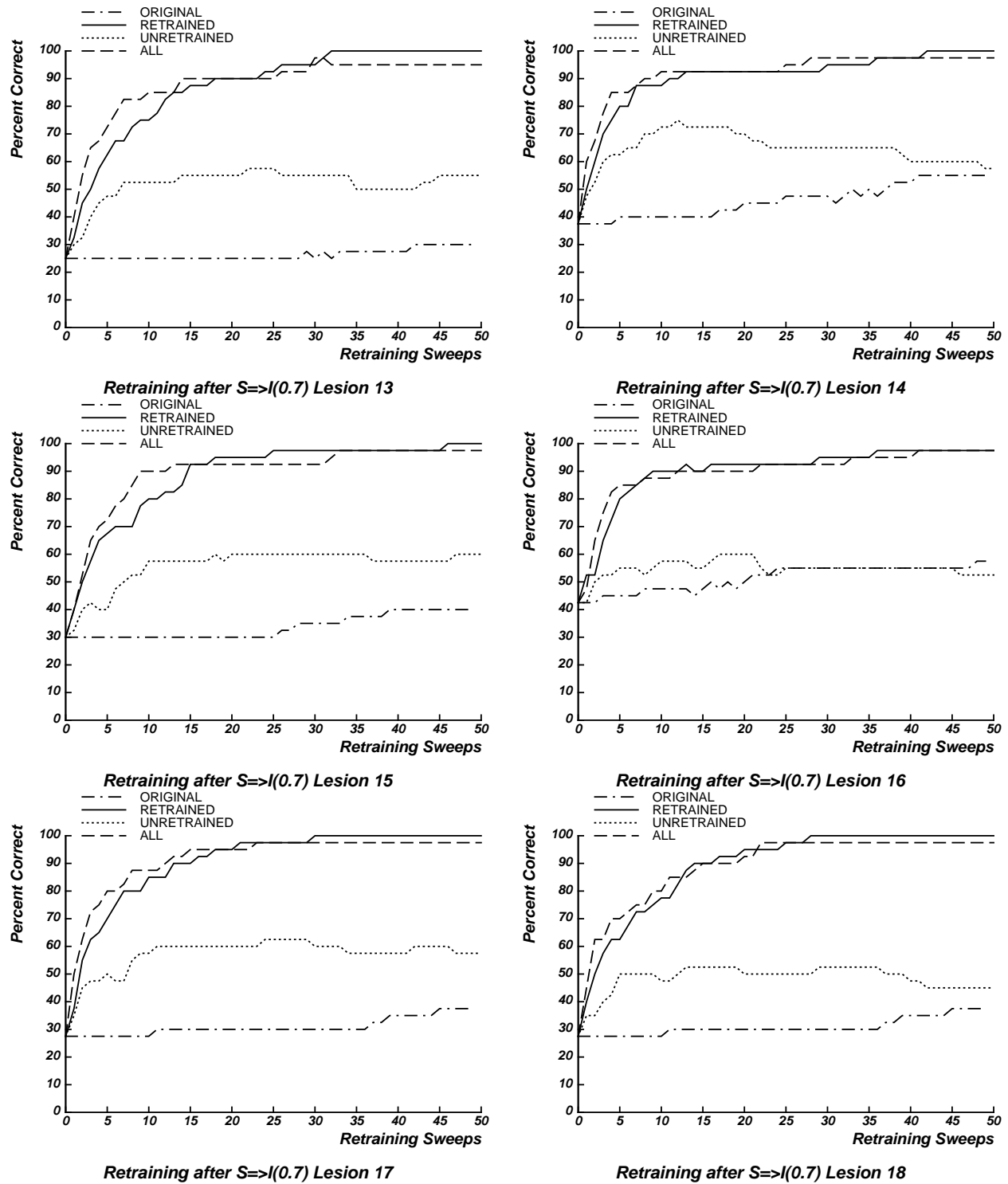



Figure 7.26: Retraining performance after  $S \Rightarrow I(0.7)$  lesions 13–18 of the  network.

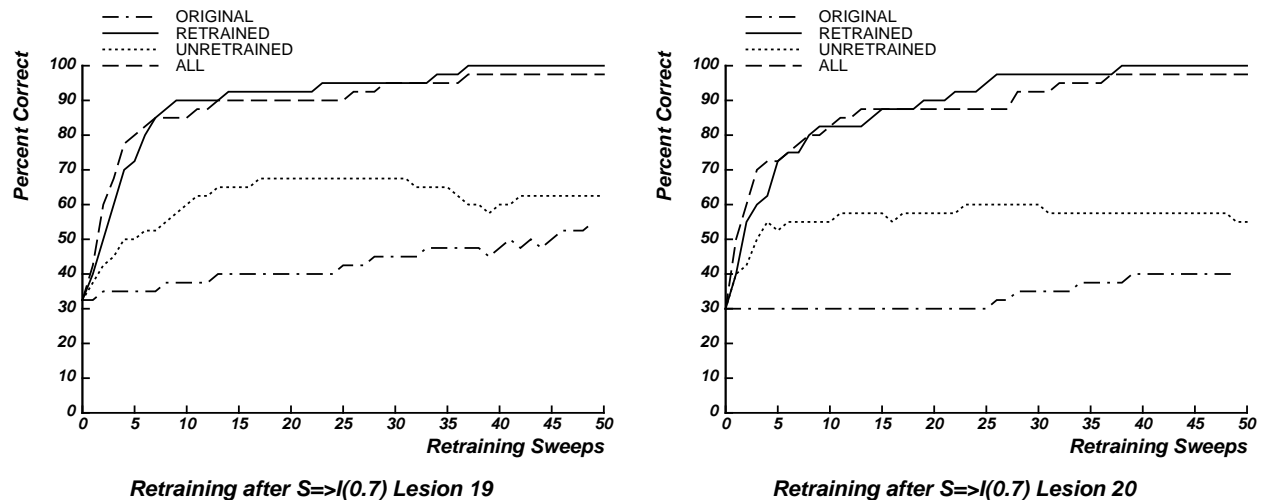



Figure 7.27: Retraining performance after  $S \Rightarrow I(0.7)$  lesions 19 and 20 of the  network.

in addition,  $S \Rightarrow I$  lesions in the  network are particularly prone to overlearning.

### 7.3.6 An explanation for differences in relearning and generalization

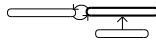
The work of Hinton & Sejnowski and Hinton & Plaut produced two main effects: (1) relearning in networks with weights corrupted by noise is significantly faster than original learning, and (2) retraining on some associations can also improve performance on others. These effects were believed to be general to all connectionist networks that employ distributed representations. The current simulations investigate relearning after lesions in networks that map orthography to semantics using attractors. They may be more likely to be directly relevant to patient therapy because the type of damage used—permanent lesions rather than corruption by noise—better approximates the type of brain damage suffered by patients.<sup>4</sup>

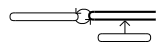
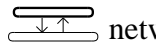
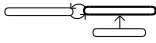
The lesion simulations demonstrate considerable differences in the degree of relearning and generalization as a function of lesion location. Specifically, lesions to connections that are involved in implementing semantic attractors result in almost complete relearning after damage and considerable transfer to unretrained associations. In contrast, relearning after lesions prior to the level where attractors operate is much less effective, and produces no generalization to unretrained associations. Although there is considerable variability across individual lesions, by and large they all show the same basic trends. Thus the relearning and generalization effects appear to be less general than originally believed, and the differences require explanation.

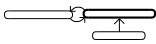
As described above, the degree of rapid relearning and generalization depends on the consistency of the directions of movement in weight space that would be optimal for individual words. While this is typically described in terms of the degree of overlap in the distributed representations of

<sup>4</sup>Hinton & Sejnowski demonstrated rapid relearning after permanent removal of hidden units, but did not investigate generalization to unretrained items under these conditions.

words, it depends more precisely on the consistency or structure in the mapping from input to output. Viewed as an abstract task, there is no systematic structure in mapping orthographic strings onto their semantics. However, when instantiated in a network, the task is broken down by the learning procedure into a number of separate transformations involving intermediate representations carried out by different parts of the network. These transformations constitute “subtasks” that may differ considerably in their degree of structure.

For example, compare  $O \Rightarrow I$  connections with  $I \Rightarrow S$  connections in the  network. The subtask of the  $O \Rightarrow I$  connections is to generate intermediate layer representations that are as semantically organized as possible from visually organized inputs. Since semantic similarity is unrelated to visual similarity, there is no structure in this subtask. However, to the degree to which the  $O \Rightarrow I$  connections succeed in inducing semantic organization at the intermediate layer, the subtask of the  $I \Rightarrow S$  connections of generating actual semantic representations from the intermediate ones *does* have structure (i.e. similar “input” (intermediate) patterns predict similar “output” (semantic) patterns).

This explains the relative differences in the degree of rapid relearning and generalization observed after lesions to different locations in the  and  networks. In essence, the effectiveness of relearning after a lesion to a set of connections simply reflects to degree to which the mapping those connections carry out is structured. Lesions to  $O \Rightarrow I$  connections in both networks produce relatively slower relearning and no generalization because their mapping is unstructured.  $I \Rightarrow S$  lesions in the  network yield somewhat faster relearning and slight generalization because their task is somewhat more (semantically) structured. Finally, lesions to connections that implement semantic attractors produce the fastest relearning and greatest generalization because the mappings they carry out are highly semantically structured. Thus the important distinction is not so much between pre- vs. within-attractor connections, but rather the degree to which the mappings that sets of connections implement are semantically structured. This distinction was not made by Hinton & Sejnowski or Hinton & Plaut because, in their simulations, corrupting weight changes affected input-to-hidden and hidden-to-output connections equally. The current work predicts that relearning after corrupting only hidden-to-output weights should yield more generalization than after corrupting input-to-hidden weights.

This explanation implicitly assumes that all of the relearning after a set of connections is lesioned takes place only among the remaining connections at that location. While this is not precisely true, there is a strong bias towards making weight changes near the lesion in relearning. To illustrate this, Figure 7.28 shows the difference of the average weight change per connection for each set of connections in the  network from the average for the entire network when relearning after  $O \Rightarrow I(0.3)$  vs.  $C \Rightarrow S(0.5)$  lesions. When relearning after  $O \Rightarrow I$  lesions, connections at this location (and also the biases of the intermediate) change much more dramatically than any other connections in the network. In fact, weights changes within the clean-up pathway are all

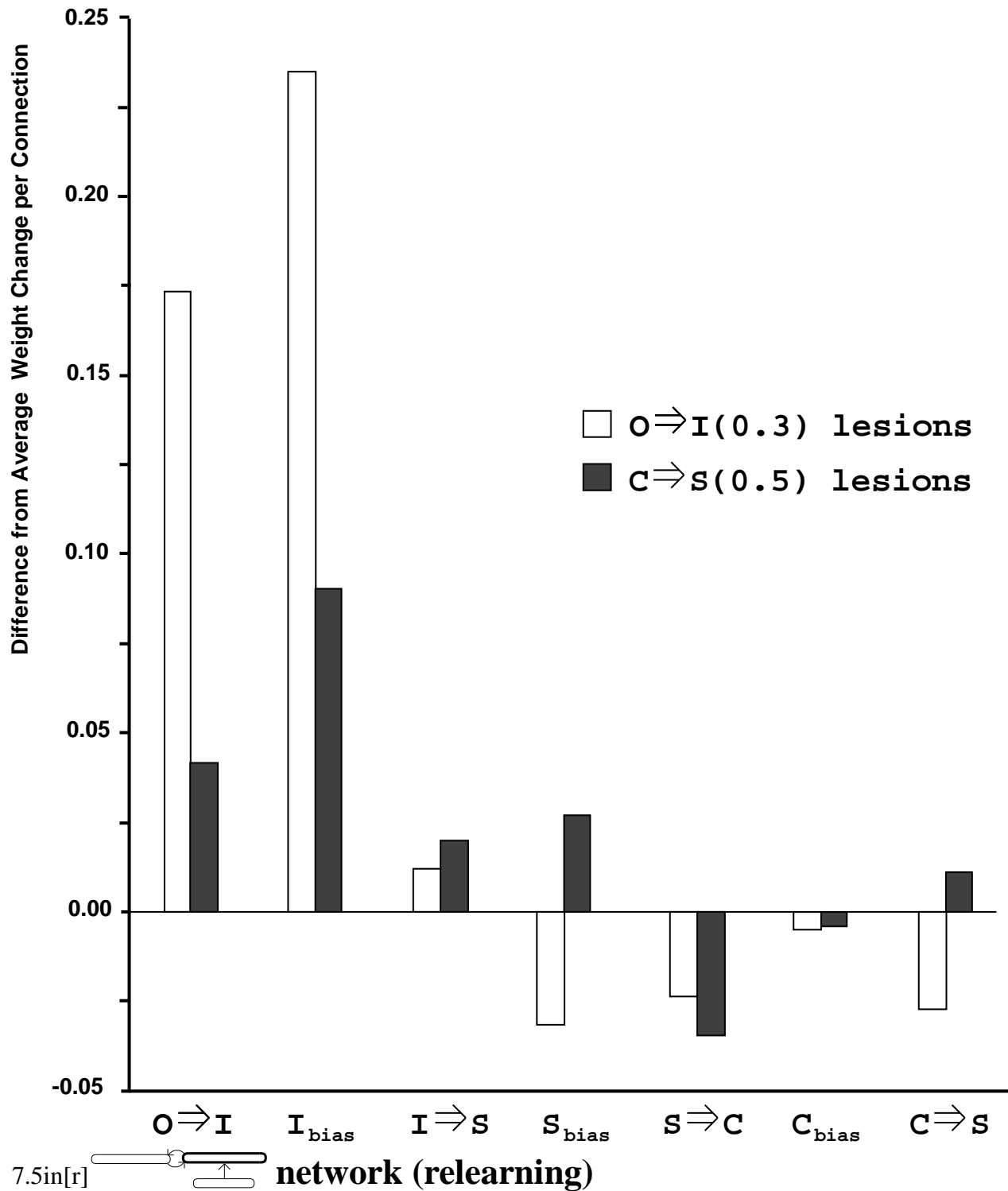



Figure 7.28: Differences in the amount of weight change per connection for different sets of connections of the network when relearning after  $O \Rightarrow I(0.3)$  vs.  $C \Rightarrow S(0.5)$  lesions. Each value is the mean weight change per connection at each location subtracted from the mean for the entire network, when relearning half of the words.

below-average under these conditions. Relearning after  $C \Rightarrow S$  lesions also results in considerable learning within the direct pathway, but it is much reduced relative to  $O \Rightarrow I$  lesions. Furthermore,  $C \Rightarrow S$  connections themselves, and the biases of the semantic units, show above-average changes, in contrast to relearning after  $O \Rightarrow I$  lesions. Thus the relative amount of relearning in different parts of the network appears to be strongly biased towards the location of lesion.

Finally, why do  $S \Rightarrow I$  lesions in the  network result in overlearning? Notice that  $S \Rightarrow I$  connections are different from other sets of connections that produce significant transfer ( $I \Rightarrow S$  and  $C \Rightarrow S$ ) in that relearning with  $S \Rightarrow I$  connections improves performance by altering the intermediate layer representations of words. This enables the network to become increasingly specialized at representing the retrained words (at the expense of the unretrained words) as retraining is prolonged. Relearning in  $I \Rightarrow S$  or  $C \Rightarrow S$  connections can only provide improved access to fixed semantic representations. Thus retraining on the retrained words can only affect the unretrained words to the degree that their pre-existing representations overlap (at the intermediate or clean-up layer). This overlap is the basis of generalization but is not affected by prolonged retraining.

### 7.3.7 Comparisons with patient rehabilitation studies

Studies of cognitive rehabilitation of acquired dyslexics in the domain of reading for (or writing from) meaning have demonstrated considerable relearning of treated items and (often) improvement on untreated but related items. Relearning after lesions to a network that operates in the same domain results in similar qualitative effects, although the magnitude of the effects depends on the particular location of damage. Thus at a general level, the cause of rapid relearning and generalization in the network—distributed representations and structure in subtasks—may provide an explanation for the nature of recovery in these patients. Unfortunately, the current simulations are too limited, and sufficiently different from the patient studies, that a more detailed comparison would be premature.

One specific hypothesis that does come out of the relearning simulations relates to the systematic differences observed in the degree of relearning and generalization as a function of lesion location. The simulations predict that a patient with a functional impairment close to or within semantics should show considerable generalization, while one with an impairment close to orthography should show little or none. Conversely, the degree of generalization observed in a patient can be used to predict the fine-grained location of their functional impairment *within* the semantic route. However, the variability in the degree of relearning and generalization observed across instances of the same lesion weaken these predictions somewhat.

## 7.4 Designing therapy to maximize generalization

Ideally, we would like to use our understanding of the impairment in a particular patient to lead to the design a rehabilitation strategy that maximizes recovery. A potential benefit of connectionist



Figure 7.29: The patterns of semantic activity corresponding to the prototype of each semantic category.

modeling in neuropsychological rehabilitation is that it provides a framework for investigating the relative effectiveness of alternative rehabilitation strategies. To this point, we have only investigated variables, such as lesion location, that (unfortunately) are not under the control of the therapist. The only type of decision that is available within the current framework is what items are selected for retraining. Even here, the limited size and complexity of the training set severely constrains the alternative strategies that can be investigated. Below we explore the implications of perhaps the simplest distinction that can be made in the domain of reading for meaning—semantic prototypicality.

#### 7.4.1 Semantic prototypicality

In general, word meanings differ with respect to how representative they are of their semantic category. One measure of representativeness is the distance of the semantics of the word to the prototype of its category. In a feature-based semantic representation like the one we employ, the prototype of a category can be defined simply as the pattern of semantic activity that is the *average* of the patterns for the words in the category (i.e. the centroid of the patterns in semantic space).<sup>5</sup> Figure 7.29 shows the patterns of semantic activity corresponding to the prototypes of the semantic categories in the H&S word set. In the 68-dimensional space (hypercube) of all possible semantic representations, the point for the prototype of a category defined in this way is the point at the “center of mass” of all the points for words in the category. Although the semantics of each word is located at a corner of semantic space, the prototypes fall somewhere within the hypercube. This is not a problem since the prototype need not correspond to the semantics of any particular word.

For each word in a category, we can compute its proximity to the category prototype and use this as a measure of how “representative” the word is of its category. Figure 7.30 shows the proximity of each word to its category’s prototype. First notice that the average “prototypicality” varies considerably across categories. This average value reflects how tightly a category is clustered, and corresponds to the visual impression of within-category similarity seen in the similarity matrix for the semantic representations (see Figure 2.7, p. 36). Thus “body parts” are most tightly clustered, while “indoor objects” and “outdoor objects” are each relatively spread out in semantic space.

<sup>5</sup>These are the same patterns as were used by H&S as the basis for performing between-category discrimination tasks (see Section 2.6.4).



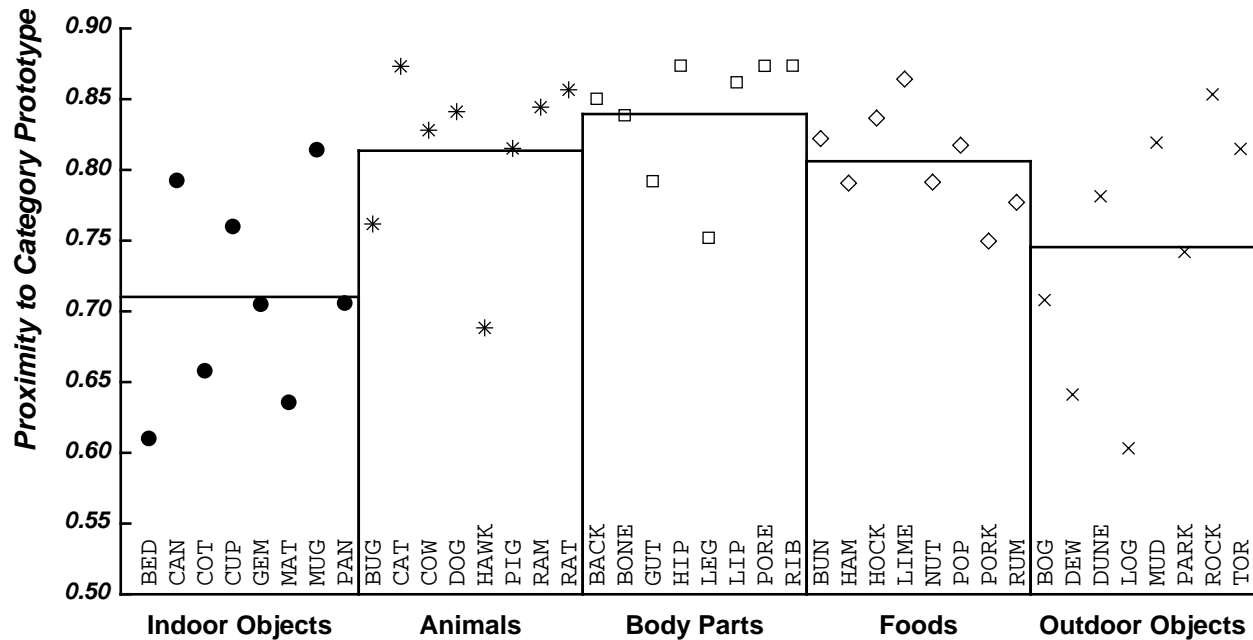


Figure 7.30: The proximity of each word to the prototype of its category. The height of the bar for each category is the average proximity to the prototype for words in that category—not a measure of the location of the prototype itself.

The variation of prototypicality within each category is also quite different across categories. For instance, “foods” are all fairly equally close to their prototype. In contrast, a few “outdoor objects” (e.g. MUD, ROCK, and TOR) are quite close to the category prototype, while others (e.g. DEW and LOG) are quite atypical of the category.

#### 7.4.2 Selecting retraining items based on prototypicality

The straightforward way of testing the effect of prototypicality on generalization during relearning would be to retrain the network after every lesion on the four words in each category with the highest proximity to the prototype. However, it is important to balance the retrained and unretrained word sets for correct performance, otherwise any observed differences might simply reflect the fact that there is more room for improvement on a word set with poorer initial performance.

Accordingly, in selecting the retrained words for relearning, words in each category were first divided into correct and incorrect sets. Among each set, the half with the highest prototypicality were added to the retrained set and the rest to the unretrained set. If there was an odd number of incorrect words in the first category, “indoor objects,” the extra word was assigned to the retrained or unretrained set randomly—extra errors in subsequent categories were assigned in such a way as to maintain the balance of correct performance between the retrained and unretrained sets. In this way, both the retrained and unretrained sets contained 20 words (4 from each category) and differed by at most one error in performance (randomly over lesions). Within this constraint, the retrained set was

biased as strongly as possible towards including words with the highest prototypicality. We will refer to the retrained and unretrained sets thus defined as the “prototypical” and “non-prototypical” sets, respectively.

It should be noted that the relearning procedure involves a comparison in which the type of the unretrained set (prototypical or non-prototypical) varies as well as the type of the retrained set. Hence, any differences in generalization from retrained to unretrained words may arise from characteristics of the *unretrained* “test” set rather than the retrained set. In fact, further simulations support this conclusion. However, we start by simply demonstrating that separating words into retrained and unretrained sets based on semantic prototypicality influences the degree of generalization in relearning after damage.

### 7.4.3 Effects of prototypicality on generalization

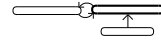
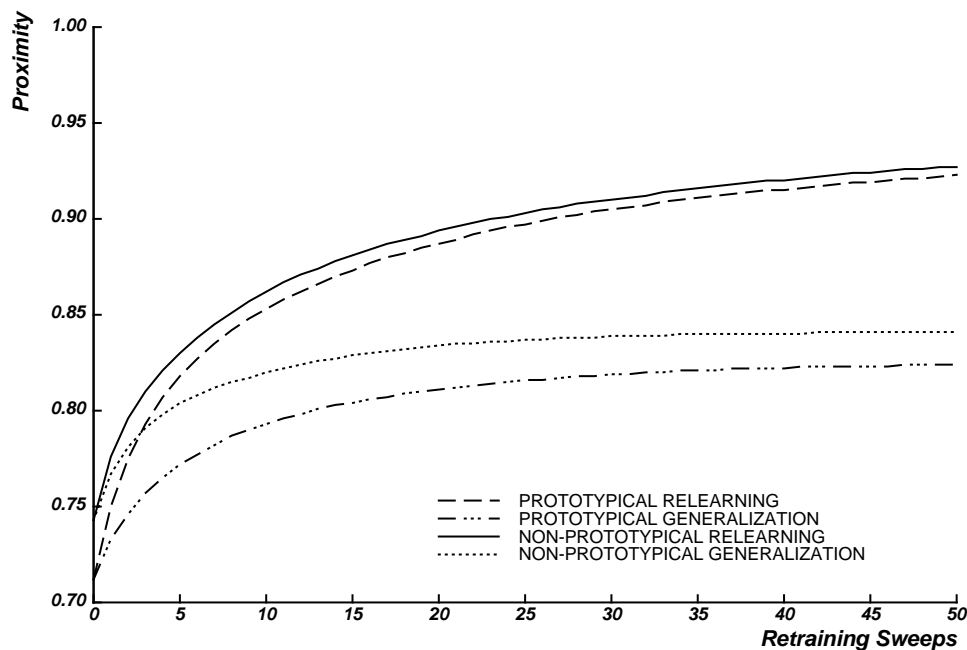
We investigated the effects of prototypicality on generalization after  $C \Rightarrow S(0.5)$  lesions in the  network. This lesion was chosen because the generalization during relearning is significant but allows room for increases or decreases. The identical 20 instances of lesion studied above were administered using the same random number generator seeds. Thus, the curves for the original learning, and for retraining on all 40 words, are identical to those found in Figures 7.17 to 7.21.

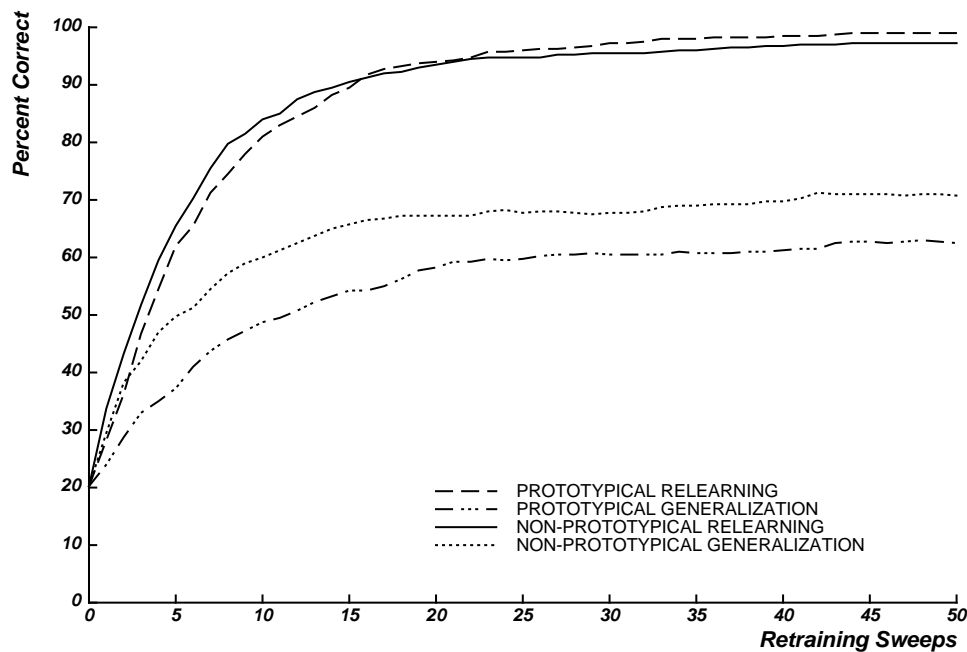
Figure 7.31 presents graphs of the data for average proximity and explicit correct performance. In the graphs, “relearning” denotes the performance on the retrained word set, while “generalization” denotes the performance of the unretrained word set. There is no significant difference in the amount of relearning between the two word sets (mean correct performance improvement: 77.0% for prototypical words, 78.8% for non-prototypical words, paired  $t(19) = 1.58, p = .13$ ). However, performance on the prototypical words improves to a greater extent when retraining on the non-prototypical words than *vice versa* (50.5% prototypical vs. 42.3% non-prototypical, paired  $t(19) = 3.42, p < .005$ ). Thus the generalization ratio of unretrained to retrained improvement is higher for the prototypical than non-prototypical words (0.65 for prototypical vs. 0.53 for non-prototypical, paired  $t(19) = 3.50, p < .005$ ). Although the effect is not dramatic, retraining on non-prototypical words results in greater generalization (and greater overall improvement) than retraining on more prototypical words.

### 7.4.4 An explanation for the prototypicality effects

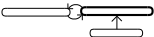
Why does retraining on non-prototypical words generalize more to prototypical words than retraining on prototypical words generalizes to non-prototypical words? Earlier it was argued that the amount of generalization observed in relearning after a lesion to a set of connections depends on the structure in the mapping that those connections carry out. Words in a category that are highly



Retraining after  $C \Rightarrow S(0.5)$  Lesions



Retraining after  $C \Rightarrow S(0.5)$  Lesions

Figure 7.31: Relearning and generalization in average proximity (top) and correct performance (bottom) on prototypical and non-prototypical words from retraining after  $C \Rightarrow S(0.5)$  lesions of the  network.

prototypical also tend to be highly similar to each other. Thus prototypical words are precisely those for which the structure in the clean-up mapping is highest—words with similar semantics require similar clean-up influences. In contrast, words with atypical semantics require more idiosyncratic support from the clean-up pathway.

This contrast can be quantified in the following way. Each word generates a particular final pattern of activity over the clean-up units. We can compute the average or “prototypical” clean-up for a category in the same way as for semantic representations. Words vary in the degree to which the clean-up they require differs from the average clean-up for the category—the proximity of a word’s clean-up representation to its category’s prototype is a measure of its clean-up prototypicality. In fact, semantic and clean-up prototypicality is highly correlated ( $0.86$ ,  $t(38) = 10.3$ ,  $p < .001$ ). Thus semantically typical words require similar clean-up, while semantically atypical words require peculiar clean-up.

Let us assume that the effect of retraining on a word is to adjust the clean-up representations, or their influence on semantics, towards the “desired” clean-up of that word, and that the effect on unretrained words is a function of the average effects on the clean-up due to the retrained words.<sup>6</sup> Relearning on the retrained words will cause generalization to the extent that the clean-up moves towards that needed by the unretrained words. The average effect of retraining on the non-prototypical words will be to move the clean-up towards the average for the category, which is near to the clean-up needed by each of the prototypical words. Retraining on the prototypical words also moves the average clean-up towards the average for the category, but this does not move closer to the clean-up needed by each of the non-prototypical words. In other words, the average clean-up of the non-prototypical words is closer to the desired clean-up of each of the prototypical words than *vice versa* (average proximity of the prototypical words to the non-prototypical average is  $0.66$ , average proximity of the non-prototypical words to the prototypical average is  $0.60$ , within-category comparison  $F(1, 30) = 6.90$ ,  $p < .02$ ). By analogy, in a set of randomly distributed points, the average of the outliers may be quite near the central points, but the average of the central points is still quite far from the outliers. (see Figure 7.32). This effect is diminished as the dimensionality of the representations is increased. However, it is enhanced to the degree that the distribution of non-prototypical words is uniformly distributed around the category mean. Individual words may be quite deviant from the mean along some dimensions, but as long as there are others that collectively are equally deviant in the opposite direction, the average for the non-prototypical words will still be close to the prototypical words. This explains why retraining on non-prototypical words yields more generalization to prototypical words than *vice versa*.

---

<sup>6</sup>This is not completely true because significant relearning also takes place in the direct pathway. In fact, the largest difference in the amount of weight change per connection is in the biases of the intermediate units ( $0.16$  for prototypical retraining,  $0.14$  for non-prototypical retraining, paired  $t(19) = 3.4544$ ,  $p < .005$ ) and in the  $I \Rightarrow S$  connections ( $0.09$  prototypical vs.  $0.10$  non-prototypical, paired  $t(19) = 9.44$ ,  $p < .001$ ). However, the same argument applies to the extent that the intermediate representations, and their influence on semantics, are semantically organized.

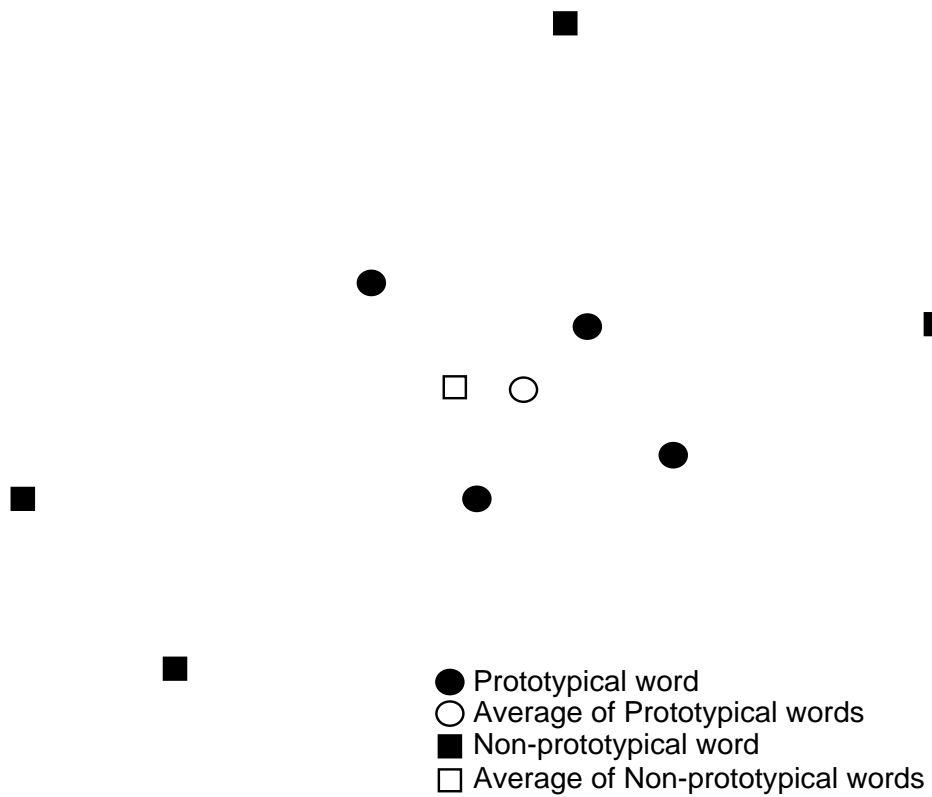


Figure 7.32: Prototypical words are closer to the average of the non-prototypical words than the non-prototypical words are to the average of the prototypical words.

However, this explanation leaves open the possibility that difference is due to greater generalization *to* prototypical words rather than *from* non-prototypical words. That is, the current experiment confounds retraining on non-prototypical words with testing generalization on prototypical words. Perhaps prototypical words recover to a greater extent regardless of the nature of the retrained set. A further experiment was run to test this possibility.

#### 7.4.5 A more detailed test of prototypicality effects

We would like to compare generalization to prototypical vs. non-prototypical words when retraining on either type of word. In order to do this, another relearning experiment was run, in which words were divided into prototypical and non-prototypical groups as described above, and then one group was further divided in half. One of these halves formed the retrained set, while the other formed one unretrained set, and the words of the opposite type formed a second unretrained set. For example, the network might be retrained on 10 prototypical words (2 from each category), with generalization tested both on the remaining 10 prototypical words, and on the 20 non-prototypical words. For both prototypical and non-prototypical words, all possible choices of pairs of retrained words from each category were tested for each of 20 lesions.

For each condition, we computed the generalization ratio (i.e. ratio of unretrained to retrained improvement in correct performance) for both the prototypical and non-prototypical unretrained sets. Figure 7.33 presents a bar graph of the averages of these values, separated by the type of the retrained word set. Replicating the previous results, retraining on non-prototypical words produces more generalization to prototypical words (0.55, left black bar) than *vice versa* (right white bar, 0.42;  $F(1, 238) = 38.5, p < .001$ ). However, retraining on other prototypical words produces even more generalization (0.60 vs. 0.55,  $F(1, 238) = 5.03, p < .05$ ). Overall, prototypical words recover better than non-prototypical words, regardless of the nature of the retraining set (0.58 vs. 0.47,  $F(1, 476) = 33.4, p < .001$ ). The main effect of the type of retrained set favors non-prototypical words (0.53 vs. 0.51) but it is not significant ( $F(1, 476) = 1.49, p = .22$ ). However, the interaction between retrained and unretrained type is significant ( $F(1, 476) = 16.8, p < .001$ ), with prototypical words generalizing particularly poorly to non-prototypical words. Thus the effects of prototypicality in relearning appear to arise from the greater improvement of prototypical words, rather than a specific advantage in retraining on non-prototypical words.

An important question in evaluating the implications of these results for patient therapy is how they would be affected by using much larger, and more realistic, retrained and unretrained sets. In the current experiment, only two words in each category are retrained. The preceding explanation of how non-prototypical words can better approximate the clean-up required by prototypical words than *vice versa* relies on there being sufficiently many non-prototypical words to approximate the average for the category. Clearly, only two words cannot accomplish this. In contrast, even a few prototypical words can provide a reasonable estimate. For this reason, it is likely that retraining

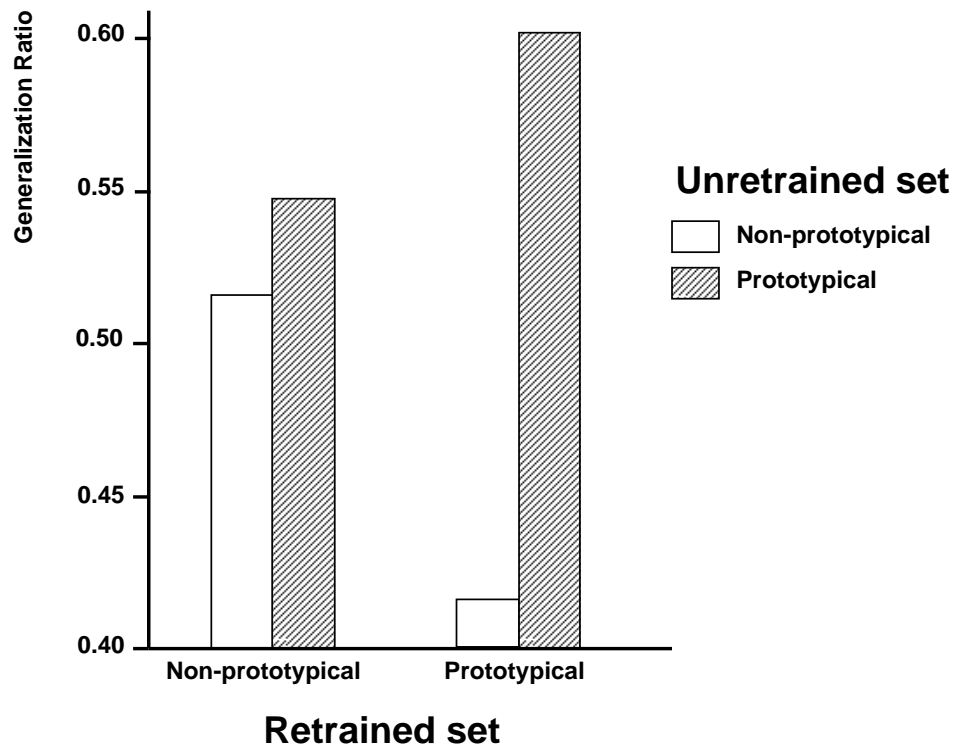


Figure 7.33: Generalization from prototypical or non-prototypical retrained sets, to prototypical or non-prototypical unretrained sets.

on many more non-prototypical words would increase the generalization they produce to a much greater extent than for prototypical words. If this were true, there would also be a main effect of the retrained type, with non-prototypical better than prototypical. However, this speculation must be tested in a larger-scale simulation.

## 7.5 Summary

Theoretical analyses of cognitive impairments following brain damage should lead to the design of more effective strategies for rehabilitation. Results presented in previous chapters have demonstrated the effectiveness of connectionist networks at reproducing the detailed characteristics of some patients with impairments in mapping orthography to semantics. Simulations in this chapter extend the relevance of connectionist modeling in neuropsychology to address issues in rehabilitation, concerning the degree and speed with which behavior can be reestablished through retraining, the extent that recovery due to treatment of particular items generalizes to other materials, and possible bases on which to select items for treatment so as to maximize this generalization.

Attempts at cognitive rehabilitation of the mapping between orthography and semantics in patients have resulted in considerable improvement in performance on treated words, as well as significant generalization to untreated but related words, although the degree of recovery can vary considerably. The degree of relearning and generalization after damage in a network that performs the analogous task depends considerably on the location of damage. These differences can be understood in terms of the amount of structure in the subtasks performed by parts of the network. The differences also provide a possible explanation for the variability in recovery observed in patients, and generate hypotheses about the specific location of their underlying functional impairment.

A potential benefit of connectionist modeling in neuropsychological rehabilitation is that it provides a framework for investigating the relative effectiveness of alternative rehabilitation strategies. Further simulations found an effect of semantic prototypicality on the degree of generalization produced by relearning, although the results are too limited to support definitive implications for patient therapy.

Overall, while the results of the current research are modest, they demonstrate that investigations of relearning after damage in connectionist networks can provide an account of the general nature of relearning and generalization in patients, as well as generate interesting hypotheses about the design of effective patient therapy.