To appear in: G. Houghton (Ed.), *Connectionist Models in Cognitive Psychology*, pp. 83-111.
Brighton, UK: Psychology Press, 2005.

Connectionist Neuropsychology

John A. Bullinaria

School of Computer Science, The University of Birmingham
Birmingham, B15 2TT, UK

1 Introduction

The field of cognitive neuropsychology employs the patterns of performance observed in brain damaged patients to constrain our models of normal cognitive function. This methodology was historically based upon simple "box and arrow" models, with particular cognitive deficits being taken as indicative of the selective breakdown of corresponding "boxes" or "arrows". In principle, within this framework, one should be able to piece together a complete model of mental structure by studying patients with complementary patterns of deficit (e.g., Caramazza, 1986; Shallice, 1988). The concept of *double dissociation* has been of particular importance for this enterprise, with its presence being taken to imply modularity across a whole range of systems. I shall review the technical details and provide specific examples later, but the basic inference is that if one patient can perform task 1 but not task 2, and a second patient can perform task 2 but not task 1, then a natural way to explain this is in terms of separate modules for the two tasks.

Cognitive modelling has now moved on, and the use of connectionist techniques to provide detailed models of the inner workings of these modules or "boxes" is becoming increasingly common. Typically, networks of simplified neurons loosely based on real brain structures are set up with general architectures based on known physiology, and trained to perform appropriately simplified versions of the human tasks. The models are iteratively refined by requiring their learning process to match children's development, their generalization performance to match human generalization performance, their reaction times to match human reaction times, and so on. These individual network models can then be wired together in the manner of the old box and arrow models, and all the old explanations of patient data can carry through. The obvious advantage this provides is that one can now take a more detailed look at the performance and degradation of the various components, and the removal of neurons or connections in these models constitute a more natural analogue of real brain damage (Farah, 1994). However, in addition to providing an elaboration of the previous models, one can also question in detail the validity of the old assumptions of neuropsychological inference. In particular, Bullinaria & Chater (1995) have considered the possibility that double dissociation does not really imply modularity, but may also be possible as a result of damage to fully distributed connectionist systems. They concluded that, assuming one successfully avoids small scale artefacts, only single dissociations are possible without modularity. Moreover, these single dissociations were seen to be rooted in natural regularity effects with regular mappings more robust than irregular mappings. These general arguments have since been extended from simple abstract mappings through to more realistic single route models of reading which show how surface dyslexia like effects can arise, but phonological dyslexia effects cannot (Bullinaria, 1994, 1997a,b).

Whilst finding a counter-example to the inference from double dissociation to modularity would clearly settle the matter, failing to find a counter example will always be less conclusive. There have also been some reports in the literature containing conflicting

conclusions concerning models from the class investigated by Bullinaria & Chater (1995). Marchman (1993), for example, has studied models of past tense production and seemingly found dissociations with the irregular items more robust than the regulars. Moreover, Plaut (1995) has apparently found a connectionist double dissociation without modularity. Naturally, these apparent contradictions have caused a certain amount of confusion, particularly amongst researchers unfamiliar with the detailed workings of connectionist models. In this chapter I shall review and extend the work of Bullinaria & Chater (1995) with view to minimising future confusion in this area.

The concrete illustrative models, on which the following discussion shall be based, all have the same simplified structure of a fully-connected feed-forward network with one hidden layer trained using some form of gradient descent error minimization algorithm on some combination of regular and irregular mappings. A set of regular items (defined as such because they follow consistent mappings in the training set) will naturally be easier to learn than irregular items, and consequently they get learnt more quickly and accurately. We shall see that this then results in them requiring more damage for them to be lost again. This sounds simple enough, but we have to be careful about the details of our definition of regularity. In terms of network learning, a very high frequency "irregular" item might be deemed more "regular" than a consistent set of regular items whose total frequency is still much less than the irregular item. Also, if an irregular item is very "close" in the input/output space to a regular set, then we might deem that item particularly irregular and the regular items less regular than usual. (Though talking about "consistency", rather than "regularity", is usually more useful in such cases.) In the following sections I shall present explicit neural network simulation results and argue that, as long as one controls for such confounding effects, the basic conclusion of Bullinaria & Chater (1995) holds. We shall see how the opposite "regularity" effect found by Marchman (1993) arises, and argue that it would be more correctly labeled a "frequency" effect. We shall also see how Plaut's (1995) double dissociation is consistent with our findings, it being more a different use of terminology than a different conclusion. Our discussion will also reveal how it is possible to obtain a valid strong double dissociation between high frequency irregulars and low frequency regulars due to global damage of a fully distributed connectionist system without modularity. Since regularity and frequency do tend to anti-correlate in natural language, such potential confounds are seen to require particular care in many language processing experiments and models.

In the remainder of this chapter, I shall begin by reviewing the important relevant ideas from cognitive neuropsychology: the traditional inference from double dissociation to modularity, the types of system that may exhibit double dissociation, and the problem of *resource artefacts*. This will include a varied selection of examples from the neuropsychology literature to provide an idea of what is required of connectionist models in this area. Next I will outline the basic properties of the connectionist models most commonly employed in cognitive psychology. The general properties of these models then lead naturally to a range of explicit neural network learning and lesion simulations that explore those issues of particular relevance to connectionist neuropsychology. First, I explain the general consequences of network damage as originally discussed by Bullinaria & Chater (1995), then I bring the apparently contradictory results of Plaut (1995) and Marchman (1993) into the same framework, and finally I present some more recent simulations that explore the frequency-regularity confound which is at the root of many of the recent confusions. I will end with a more general discussion of connectionist dissociations and some conclusions. Throughout I shall concentrate on the general principles that may be applied to any of the models described elsewhere in this book, rather than on presenting a series of specific case studies.

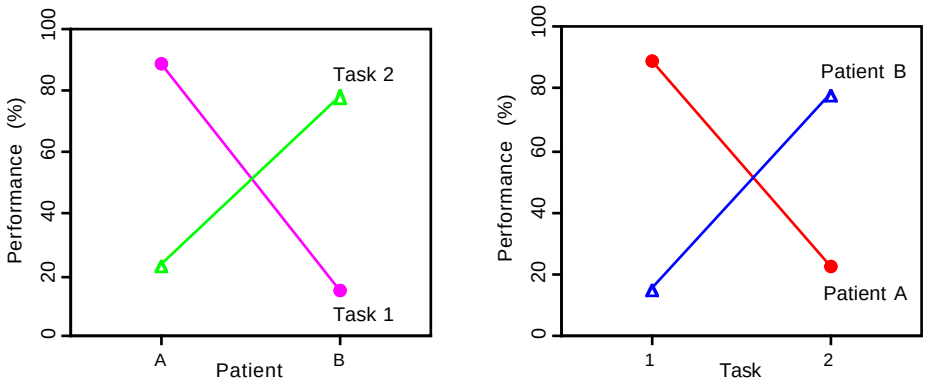

Figure 1: A strong cross-over double dissociation for Tasks 1 & 2, Patients A & B.

2 Cognitive Neuropsychology

Whilst data from individual normal subjects, or individual brain damaged patients, can undoubtedly constrain cognitive models, certain patterns of deficit across populations of patients can provide much stronger constraints. If a patient A performs very much better on Task 1 than on Task 2, then we say that we have a strong *single dissociation*. If two patients, A and B, have opposite single dissociations, then together they form a *double dissociation.* This pattern of performance can be conveniently plotted as in Figure 1. The various types of dissociation and their implications have been discussed in some detail by Shallice (1988) and Dunn & Kirsner (1988). Of particular relevance to us is the inference from double dissociation to modularity of function, which forms an important part of the foundations of cognitive neuropsychology (Teuber, 1955).

Any observed double dissociation (DD) of performance has a natural explanation in terms of the existence of separate modules associated with the two tasks, with the two patients suffering damage to a different one of them. A classic and well known example occurs in the field of acquired reading problems. Extreme cases of surface dyslexia include patient KT who could read 100% of non-words and regular words but could only manage 47% of irregular words (McCarthy & Warington, 1986). Conversely, the phonological dyslexic patient WB could read 90% of real words but was unable to read even the simplest of non-words (Funnell, 1983). This loss of exception words by surface dyslexics, together with the loss of non-words by phonological dyslexics, constitutes a DD which can be taken to imply separate Lexical and Rule-Based modules in a Dual Route Model of reading (e.g. by Coltheart, Curtis, Atkins & Haller, 1993).

However, the modules do not necessarily have to operate in parallel like this – the same data could be taken to imply modules that operate in series (e.g. by Patterson & Marcel, 1992; Bullinaria, 1997b). In fact there could be any number of different modular accounts for a particular DD, and the account that appears most natural from the point of view of boxes and arrows might not look so natural from the point of view of connectionist systems. This is another reason why it is important to explore the more detailed models offered by connectionism. For our reading example, a Rule Based box that can not deal with exception words seems less natural when it becomes clear that a neural network trained on all words will automatically learn to process the exception words as well as the regular words (Seidenberg & McClelland, 1989), and on damage result in surface dyslexia type deficits right down to the details of the regularization errors (Bullinaria, 1994, 1997a; Plaut, McClelland, Seidenberg & Patterson, 1996). Furthermore, proficient reading using

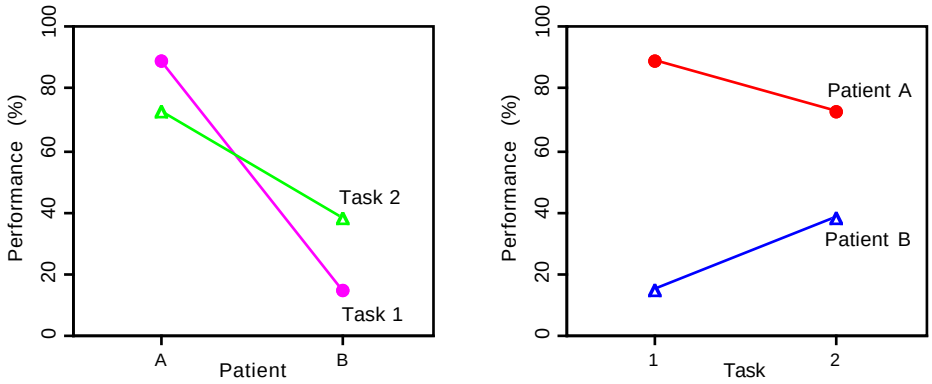

Figure 2: A weak double dissociation for Tasks 1 & 2, Patients A & B.

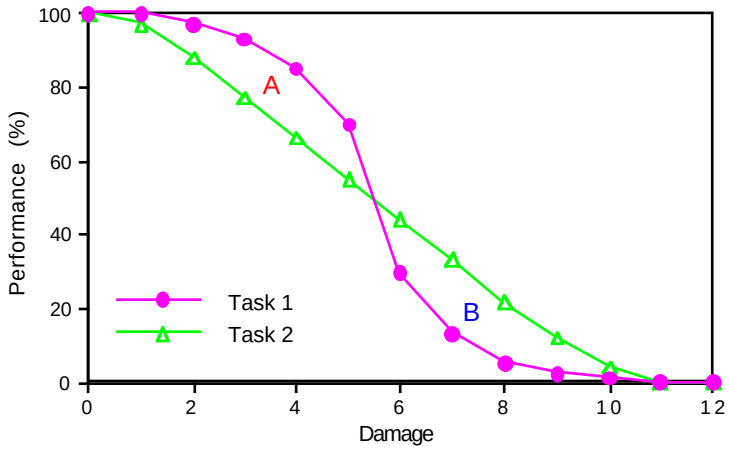

Figure 3: Tasks depending on the resources in different ways can lead to DDs.

only a Lexical/Semantic route begins to look increasingly unnatural when we find that a neural network, trained to map between orthography, phonology and semantics, prefers to access semantics from orthography via phonology rather than by direct activation (Bullinaria, 1997b). Even if the details of these particular neural network models turn out to be inaccurate, they do highlight the fact that the assignment of modules is not as clear cut as many would suggest, and show how connectionist models can be used to test the computational plausibility of the possible box and arrow frameworks.

We shall not delve into the details here, but other areas in which DD has been observed and taken to infer mental structure include regular versus irregular past tense production (e.g. Pinker, 1991, 1997; Lavric, Pizzagalli, Forstmeir & Rippon, 2001), lexical versus syntactic components of number processing (Caramazza & McCloskey, 1989), aspects of visual processing (Warrington, 1985; De Renzi, 1986; Farah, 1990), long term versus short term memory (Shallice, 1979), episodic versus semantic memory (Shoben, Wescourt & Smith, 1978), natural kinds versus artifacts in picture naming (Warrington & Shallice, 1984), to name but a few. Often it seems that DD is taken quite generally to imply modularity, and this is despite Dunn & Kirsner (1988) having shown that this inference cannot generally be justified, and Shallice (1988, p249) providing a whole list of non-modular systems that can produce dissociations (even double dissociations) when damaged (e.g. topographic maps, overlapping processing regions, coupled systems). Some early

neural network models (Wood, 1978; Sartori, 1988) also seemed to indicate that DD was even possible in distributed systems, but these were very small scale models and the effects have since been seen to be largely the consequence of individual neurons acting as "modules" in their own right. This led Shallice (1988, p257) to believe that "as yet there is no suggestion that a strong double dissociation can take place from two lesions within a properly distributed network".

Before moving on to test whether this claim has stood the test of time, we need to consider one further complication known as the problem of *resource artefacts*. As illustrated in Figures 2 and 3, a DD with a crossover in terms of patient performance, but not in task performance, can be explained as a resource artefact in a single system. All that is required is for the two tasks to depend on a single resource in different manners, such that which task is performed better depends on the amount of resource that has been spared by the damage. Clearly such a pattern of dissociation should NOT be taken to imply modularity (Shallice, 1998, p 234). As we shall see later, DDs of this type are actually rather easily obtainable in connectionist models. Devlin, Gonnerman, Anderson & Seidenberg (1998) have presented a particularly interesting example involving a connectionist account of category specific semantic deficits. The importance of connectionist modelling here is not that we can get a form of DD which is not possible in other types of model, but rather that it provides a finer grain of detail that allows us to demonstrate explicitly that, given appropriate assumptions about the information and representations being processed, the observed DD really can arise in this manner.

3 Neural Network Models

This section provides a review of the relevant features common to most neural network models used in cognitive psychology. This will prepare us for the later sections in which we discuss some explicit simulations that have been formulated to elucidate the properties that form the basis of *connectionist neuropsychology*.

Most neural network models of human performance on psychological tasks tend to be based on simple feed-forward networks that map between chosen simplified input and output representations via a single hidden layer, or have such a system as an identifiable sub-component. For this reason I shall concentrate my discussion on such systems. Extensions to more complicated systems will be readily apparent. Whilst it is obviously important that the chosen representations are appropriate for the task in hand, allow that task to be accomplished in a sufficiently human like manner, and have a reasonable degree of biological and psychological plausibility, the details will not affect the general discussion that follows.

Clearly, an important feature of connectionist models is that the networks *learn* to perform their given task by iteratively adjusting their connection weights (e.g. by some form of gradient descent algorithm) to minimise the output errors for an appropriate training set of input-output pairs. Generally, we simply assume that the quick and convenient learning algorithms we choose to use will generate similar results to those produced by more biologically plausible learning procedures. Comparisons between Back Propagation and Contrastive Hebbian Learning by Plaut & Shallice (1993) provide some justification for this assumption. We can then compare the development of the networks' performance during training and their final performance (e.g. their output errors, generalization ability, reaction times, priming effects, speed-accuracy trade-offs, robustness to damage, etc.) with human subjects to narrow down the correct architecture, representations, and so on, to generate increasingly accurate models. Here, of course, we are particularly interested in simulating neuropsychological effects by lesioning our trained networks.

To ease the subsequent discussion it is worth defining some notation. Our networks will be set up so that the output of each processing unit *i* for each training pattern *P* is the sigmoid (or "squashing function") of the sum of the bias/threshold of that unit plus the weighted activations flowing into it from the units *j* of the previous layer. We write

$$Out_i(P) = sigmoid(Sum_i(P)) \qquad Sum_i(P) = \sum_j w_{ij} Prev_j(P)$$

in which, for mathematical convenience, we define $Prev_0(P) = 1$ so that the bias w_{i0} can be treated in exactly the same way as the real connection weights. Then, to train the network, we specify a suitable error function E to minimise and iteratively update the weights w_{ij} (now including the biases) to reduce this error using gradient descent

$$\Delta w_{ij} = -\eta \frac{\partial E}{\partial w_{ij}}$$

In other words, we take repeated steps Δw_{ij} in weight space in the direction that best reduces the error E. Typically for this we use either the sum-squared output error measure

$$E = \frac{1}{2} \sum_P \sum_i \left| Target_i(P) - Out_i(P) \right|^2$$

or, for classification problems with binary outputs, the cross-entropy error measure

$$E = -\sum_P \sum_i \left[Target_i(P).\log\left(Out_i(P)\right) + \left(1 - Target_i(P)\right).\log\left(1 - Out_i(P)\right) \right]$$

(Hinton, 1989; Bishop, 1995). Often it is also appropriate to add some form of regularization term to the gradient descent cost function to smooth the outputs or improve the generalization. For example, adding a term to E that is quadratic in the weights will result in a linear weight decay during training and restrict over-fitting of the training data (Bishop, 1995). Whilst such extra factors will not usually affect the kinds of results we are concerned with here, one should always check to make sure, because there are situations in which they can have a significant effect on how well the different training patterns get learnt (e.g. Bullinaria, Riddell & Rushton, 1999).

A crucial feature of this learning approach, that underlies much of what follows, is that a network's performance on one pattern will be affected by its training on other patterns. It is helpful to begin by illustrating this with a concrete example originally presented by Seidenberg & McClelland (1989) for their reading model, but using data from my own reading model (Bullinaria, 1997a). In both cases we have a neural network model mapping from a simplified representation of orthography to a simplified representation of phonology via one hidden layer. Figure 4 shows how the output performance on the regular word "tint" varies as the result of further training of a partially trained network. First, training on the regular non-word "wint", that already has a very low error (0.000001), has no effect on the word "tint" because it generates very small weight changes. We have a ceiling effect. Compare this with three different word types with matched and relatively high error scores. Training on the regular word "dint" (0.00252) improves performance, i.e. reduces the error, because the weight changes generated for the "int" ending are also appropriate for "tint". In fact, because of its relatively low error (0.00022), even training on "tint" itself has less effect than training on "dint". Training on the irregular word "pint" (error 0.00296) worsens performance because the weight changes generated for the "int" ending here (with a long "i" sound rather than the regular short "i" sound) are inappropriate for "tint". Finally, training on the control word "comb" (0.00296) has little effect because the weight

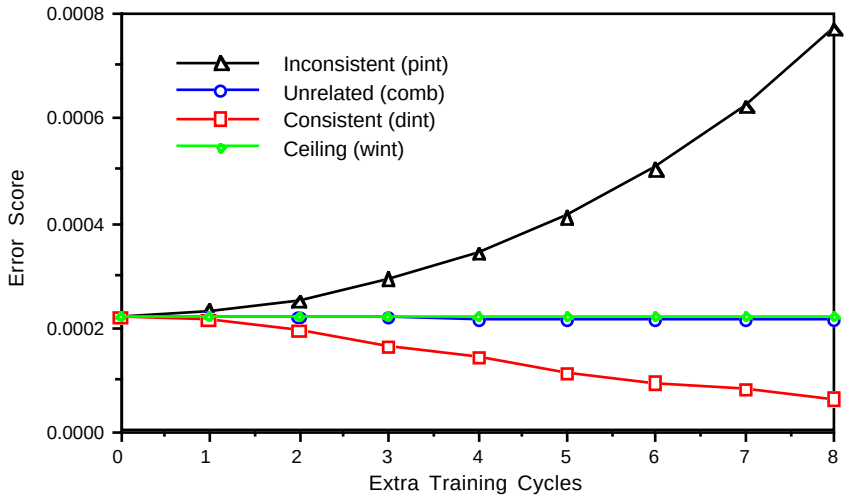

Figure 4: The effect on the word "tint" of repeated training on other words.

changes have little relevance for "tint".

By considering the implications of these performance changes for a full set of training data, it is easy to understand why the network tends to learn to read consistent sets of regular words before exceptional words, and why it generally ends up performing better (i.e. with lower output activation error scores) on regular words than on exception words. Similarly, we can understand why having inconsistent neighbours will be detrimental to learning and final reading performance. It also reveals why high frequency exception words should be learnt faster than low frequency exception words, and why we should expect ceiling effects whereby the performance on the higher frequency exception words eventually catches up that of the regular words. All this is consistent with empirical data from humans.

Given this simple example, we can easily see what will happen in the general case. Amongst other things, it follows straightforwardly from adding up the network weight change contributions due to individual training patterns that:

1. High frequency items get learnt more quickly than low frequency items, because the appropriate weight changes get applied more often.

2. Regular items will get learnt more quickly than irregular items, because consistent weight changes combine and inconsistent weight changes cancel.

3. Ceiling effects will arise as the training items are mastered, because the sigmoids saturate and the weight changes tend to zero.

These fundamental properties of neural network learning lead automatically to many of the interesting successes of connectionist models, such as human-like age of acquisition effects, patterns of reaction times, speed-accuracy trade-off effects, and so on (Bullinaria, 1997a, 1999).

Once we have trained our networks, and confirmed that they are performing in a sufficiently human-like manner, we can then set about inflicting simulated brain damage on them. Small (1991) has considered the various ways in which connectionist networks might be lesioned, and discussed their neurobiological and clinical neurological relevance. He identifies two broad classes of lesion: *diffuse* such as globally scaling or adding noise to

all the weights, and *focal* such as removing adjacent subsets of connections and/or hidden units. Which of these we choose will naturally depend on the type of patient we are modelling. Focal lesions would be appropriate for stroke patients, whereas diffuse lesions would be required for diseases such as Alzheimer's. Clearly, for our abstract models it will be appropriate to examine all these possibilities. Finally, we should be aware that relearning after damage may affect the observed pattern of deficits, and so we must check this also (Plaut, 1996; Harley, 1996).

4 Learning and Lesioning Simulations

In this section we shall explore in some detail the relation between the basic learning and lesioning effects that arise automatically in the class of neural networks outlined above. Fortunately, it proves feasible to do this by simulating some fairly small networks that are required to perform some rather simple sets of regular and irregular mappings of varying frequency.

Consider first a simple fully-connected feed-forward network with 10 input units, 100 hidden units and 10 output units with binary inputs and output targets trained on two sets of 100 regular items (permuted identity mappings) and two sets of 10 irregular items (random mappings). One of the input bits is used to signal which of the two regular mappings should be applied. The two sets of regular items used here are equivalent since the ordering of the network's input and output units is arbitrary, but we shall have one set appearing during training with a frequency of 20 times the other. Similarly for the two irregular sets. Such frequency differences can be implemented naturally over many epochs by manipulating the probability that a given pattern is used for training in a given epoch, but we can also implement them within each single epoch by scaling the weight change contributions in proportion to the frequencies. As long as the weight changes per epoch are kept small, it seems to make little difference which method we choose. Clearly, though, if the training set contains some very low frequency items and we use the many epochs approach, we need to be careful that the network is trained over enough epochs for all items to be used a reasonable number of times. The network was actually trained using the many epochs approach by gradient descent on a sum squared error measure with no regularization. The predicted regularity and frequency effects were found, as can be seen clearly in Figure 5 which shows how the mean output $Sum_i(P)$'s develop during training for each of the four item types (high frequency regular, low frequency regular, high frequency irregular, low frequency irregular) and two target activations (0, 1). If we set a particular correct response threshold for the $Sum_i(P)$'s, e.g. ±2.2 corresponding to output activations less than 0.1 for targets of 0 and greater than 0.9 for targets of 1, we see that the more regular and higher frequency items are the first to be learned during training and end up furthest from the thresholds when the training is stopped. If we add a regularization term to the gradient descent error function that leads to weight decay during training, the $Sum_i(P)$'s eventually level off rather than increasing indefinitely as in Figure 5, but we still get the same clear item type dependence. Bullinaria & Chater (1995) also found similar regularity effects for networks trained using the conjugate gradient learning algorithm on a training set of 224 regular items and 32 less regular items, and again for a set of 224 regular items and 16 random items that employed an error correcting coding. It seems that the pattern of results is quite robust with respect to the implementational details.

We can now turn to the consequences of lesioning these networks. Bullinaria & Chater (1995) found that damaging trained networks by removing random hidden units, removing random connections, globally scaling the weights, or adding random noise to the weights, all led to very similar patterns of results. Moreover, by plotting the $Sum_i(P)$'s against increasing degrees of damage, we could understand why. Figure 6 shows the effect of

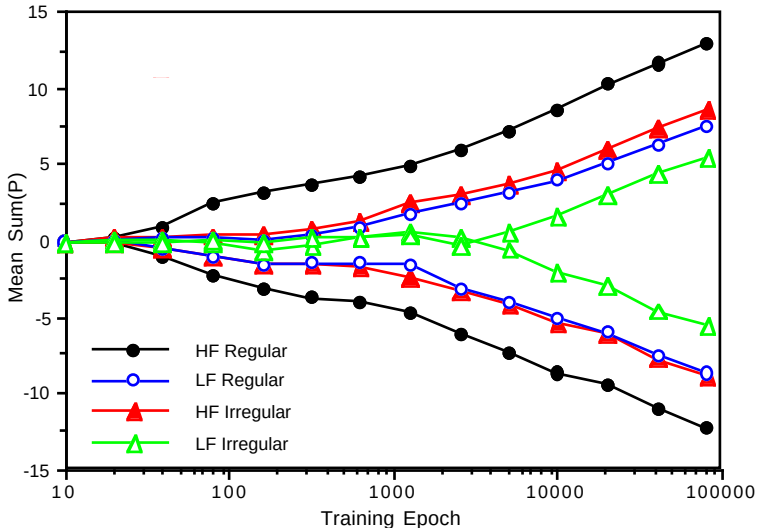

Figure 5: Learning curves for a simple network trained on quasi-regular mappings.

removing increasingly large numbers of connections from our network – we see that we get the reverse of the pattern of learning seen in Figure 5. If we set a particular correct response threshold for the $Sum_i(P)$'s as above, we see that the items that are first to be learnt during training and end up furthest from the thresholds when the training is stopped, tend to be the last to be lost during damage, and hence we get clear dissociations with the regulars more robust than the irregulars, and high frequency items more robust than low frequency items. Removing random sets of hidden units, or globally reducing all the weights by repeated application of constant scale factors, result in a similar pattern. Adding random noise to all the weights produces more of a general random walk rather than a drift to zero $Sum_i(P)$, but still it is the patterns that start nearest the thresholds that tend to cross it first, again resulting in the regulars more being robust than the irregulars. These basic effects extend easily to more realistic models, for example, surface dyslexia in the reading model of Bullinaria (1994, 1997a). Here we not only successfully simulate the relative error proportions for the various word categories, but also the types of errors that are produced. The closest threshold to an irregularly pronounced letter will be that of the regular pronunciation, and hence the errors will be predominantly regularizations, exactly as is observed in human surface dyslexics. The same basic considerations also allow us to understand various developmental and reaction time effects (Bullinaria, 1999).

After brain damage, patients often (but not always) show a rapid improvement in performance (Geshwind, 1985). This is important to connectionist modellers for two reasons. First, if relearning occurs automatically and quickly in patients, then we need to be sure that the same effects are observed in our models and that we are comparing patient and model data at equivalent stages of the relearning process. Secondly, our models may be of assistance in formulating appropriate remedial strategies for brain damaged patients (Wilson & Patterson, 1990; Plaut, 1996). It has been known for some time that the information remaining after damage does allow rapid relearning in neural networks ranging from standard back-propagation models (Sejnowski & Rosenberg, 1987) through to Boltzmann machines (Hinton & Sejnowski, 1986). It is also clear from the discussion above that, since both learning and damage result in the same regularity and frequency effects, it is unlikely that relearning using the original training data will reverse this pattern, indeed it is likely to enhance it (Bullinaria & Chater, 1995). Obviously, if some

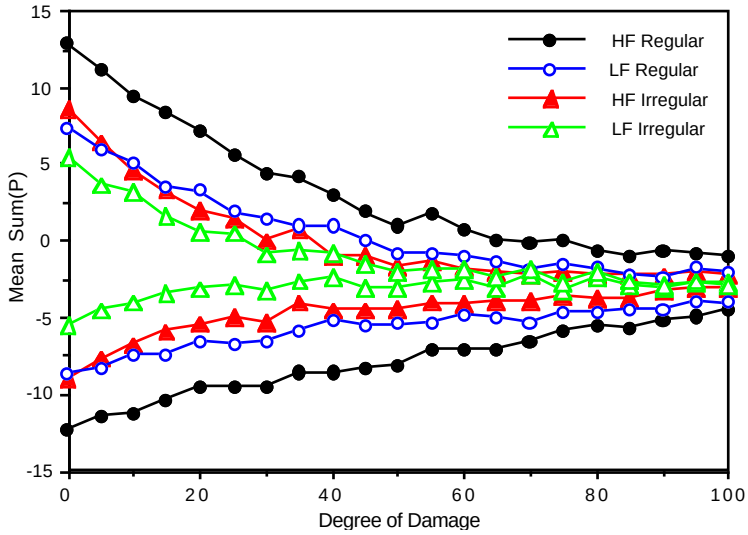

Figure 6: Damage curves corresponding to Figure 5 due to removal of connections.

rehabilitation regime is employed that involves a very different set of training examples to that of the original learning process, then it is possible for different results to arise (Plaut, 1996). In this case our models may be used to predict or refine appropriate relearning strategies and the patients' responses should be used to validate our models. In Section 7 we shall see that more complicated outcomes of relearning are possible if two or more factors, such as regularity and frequency, are confounded.

The general point to be made here is that some items are naturally learnt more quickly and more accurately than others, and the effects of subsequent network damage follow automatically from these patterns of learning. There are actually many other factors, in addition to regularity and frequency, that can cause the differing learning and damage rates and we can explore them all in a similar manner and use them in models of neuropsychological data in the same way. Consistency and neighbourhood density are the most closely related to regularity, and are commonly found in language models such as the reading and spelling models of Plaut et al. (1996) and Bullinaria (1997a). Representation sparseness or pattern strength are often used to distinguish between concrete and abstract semantics, such as in the models of Plaut & Shallice (1993) and Plaut (1995). Correlation, redundancy and dimensionality are commonly used in models to distinguish the semantics of natural things versus artefacts, such as in the model of Devlin et al. (1998). At some level of description, all these may be regarded as forms of regularity, and their effects can easily be confused. Which we use will depend on exactly what we are attempting to model, but clearly, if we want to make claims about neuropsychological deficits involving one of them, we need to be careful to control for all the others.

5 Avoiding Small Scale Artefacts

Modelling massively parallel brain processes by simulating neural networks on serial computers is only rendered feasible by abstracting out the essential details and scaling down the size of the networks. It is clearly important for all connectionist models to check that the abstraction and scaling process has not been taken so far that we miss some of the important fundamental properties of the system we are modelling, or introduce features that are nothing but small scale artefacts. Bullinaria & Chater (1995) showed that such

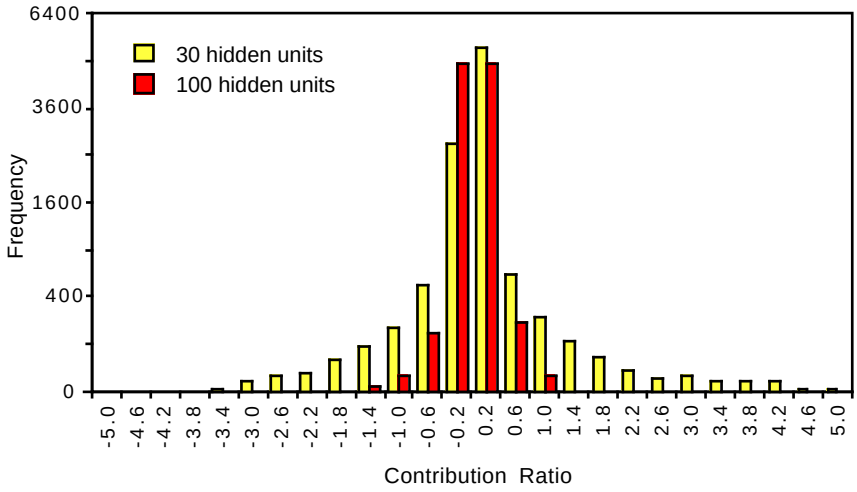

Figure 7: The distribution of output contribution ratios C for two typical networks.

artefacts can arise particularly easily in the field of connectionist neuropsychology. This complication is of such importance that it is worth discussing in more detail here.

The damage curves of Figure 6 are relatively smooth because we have averaged over many output units and many training items, and because our network has many more hidden units and connections than are actually required to perform the given mappings. For smaller networks, however, the effect of individual damage contributions can be large enough to produce wildly fluctuating performance on individual items, which in turn can result in dissociations in arbitrary directions. Often these small scale artefacts are sufficient to produce convincing looking double dissociations. The early models of Wood (1978) and Sartori (1988) are classic examples of this. As soon as we scale up to larger networks, in which the individual contributions each have a small effect on the outputs, the "regulars lost" dissociations disappear (Bullinaria & Chater, 1995). We always find that the apparent double dissociations dissolve into single dissociations as the network is made more distributed. We are then left with a simple "regularity effect" as discussed above.

It would clearly make our modelling endeavours much easier if we had some independent procedure for determining when our networks are sufficiently distributed to obtain reliable results. In effect, we need to make sure that our individual processing units are not acting as "modules" in their own right, and the obvious way to do this is by checking to see that all the individual contributions $c_{ij} = w_{ij}\, Prev_j(P)$ feeding into to an output unit i are small compared to the total $Sum_i(P) = \sum_j c_{ij}$. Clearly, if the network damage corresponds to the removal of unit j or the connection ij, then the contribution c_{ij} to the output i will be lost. If the lost contribution is small compared to the corresponding total, i.e. the ratio $C = c_{ij} / \sum_k c_{ik}$ is much less than one, then the output activation will not be changed much and it will take many such lost contributions to result in an output change large enough to be deemed an error. This is the brain-like resilience to damage often known as *graceful degradation*. Fortunately this distribution of information processing tends to occur automatically simply by supplying the network with a sufficiently large number of hidden units.

Figure 7 shows the distribution of 10000 typical individual contribution ratios C for the high frequency regular outputs in networks with 30 and 100 hidden units trained on the quasi-regular mapping discussed above. For 100 hidden units, there are very few contributions with ratios C larger than one, but with only 30 hidden units, many

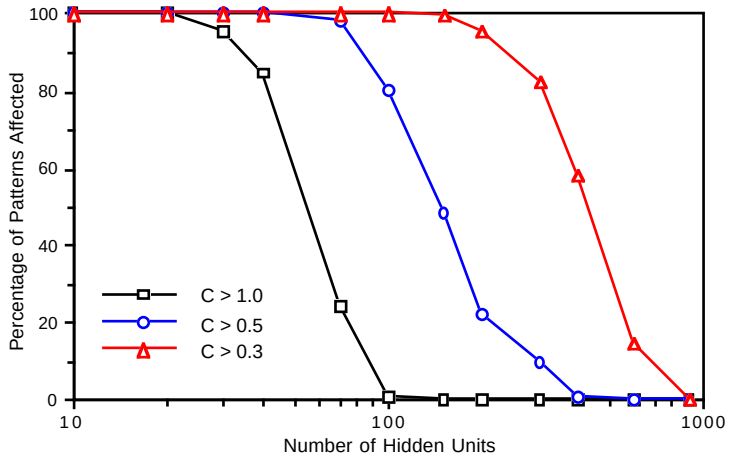

Figure 8: The fall off of large contributions C with number of hidden units.

contributions are much greater than their corresponding total and their removal will result in wild fluctuations in the outputs. The reduction in the number of large contribution ratios as we increase the number of hidden units is shown in Figure 8. Unfortunately, in general, it seems that we really do need to use a surprisingly large number of hidden units to avoid the small scale artefacts – tens, or even hundreds, of times the minimal number required to learn the given task.

It is natural to ask what can be done if limited computational resources render the use of sufficiently large numbers of hidden units impossible. Well, consider the effect of network damage on the histograms in Figure 7. Obviously, after removing a random subset of the hidden units or connections, the number of contributions will be reduced by some factor α. However, in large fully distributed networks, the mean contribution will not change much, and so the total contribution after damage is simply reduced to $\alpha\, Sum_i(P) = \alpha\, \Sigma w_{ij} Prev_j(P)$. Note that we can achieve exactly the same result by simply globally scaling all the weights w_{ij} by the same factor α. In smaller networks, of course, this equivalence breaks down because the means tend to suffer relatively large random fluctuations during damage. However, since global weight scaling does not suffer from such random fluctuations, it can be used to simulate a smoothed form of lesioning and give a reasonable approximation in small networks to what will happen in more realistic networks. Alternatively, if one wants to claim that each hidden unit in our model actually corresponds to a number of real neurons, then the weight scaling can be regarded as removing a fraction α of these corresponding real neurons. Either way, this procedure involves approximating a form of focal damage by a form of diffuse damage, and there are clear limits to the validity of the approximation. If this approach is pursued, we need to be careful not to lose sight of what type of brain damage we are modelling, and what the weight scaling really represents.

6 Plaut's Double Dissociation Without Modularity

Given that we have just concluded that valid DD does not arise in fully distributed connectionist systems, it is not surprising that Plaut's well known paper entitled "Double dissociation without modularity: Evidence from connectionist neuropsychology" (Plaut, 1995) is often taken as evidence that there must be something wrong with the above discussion. His work was based on the models of deep dyslexia of Plaut & Shallice (1993),

which in turn were extensions of the earlier models of Hinton & Shallice (1991). Deep dyslexia is a well known acquired reading disorder characterized by semantic errors such as reading "forest" as "tree". Of particular relevance to us are two patients who provide a DD between abstract and concrete word reading. Patient CAV was able to read correctly 55% of abstract words but only 36% of concrete words (Warrington, 1981), whereas patient PW could read 67% of concrete words but only 13% of abstract words (Patterson & Marcel, 1977).

The Plaut & Shallice (1993) models consist of attractor networks that map from orthography to semantics via a layer of hidden units, and then from semantics to phonology via another set of hidden units, with additional layers of "clean-up" units at the semantics and phonology levels. The particular model used to investigate concreteness had 32 orthography units corresponding to letters at particular word positions, 61 phonology units corresponding to phonemes in a similar manner, and 98 semantic units corresponding to a hand-crafted set of semantic micro-features. Each hidden layer and clean-up layer contained 10 units. The network was trained on 40 words, using back-propagation through time, until it settled into the correct semantics and phonology when presented with each orthography.

Lesions at two different locations in the trained network were then found to produce a DD between concrete and abstract word reading if the concreteness was coded as the proportion of activated semantic micro-features. Specifically, removal of orthographic to hidden layer connections resulted in preferential loss of abstract word reading, whereas removal of connections to the semantic clean-up units primarily impaired performance on the concrete words. Although the two damage locations do not constitute modules in the conventional sense, it is not difficult to understand how they contribute to different degrees to the processing of the two word types and will give opposite dissociations when damaged. It is simply a consequence of the sparser representations of the abstract words making less use of the semantic clean-up mechanism, and depending more on the direct connections, than the richer representations of the concrete words (Plaut & Shallice, 1993). The performance of each location is fully consistent with the general discussion above, and the only disagreement concerns the appropriateness of using the word "module" to describe the two damage locations.

As Plaut (1995) himself points out, one of the problems when discussing "modularity" is that different authors use different definitions of the term. A Fodor (1983) module, for example, is hard-wired, innate and informationally encapsulated, whereas a Coltheart (1985) module is defined to have none of those properties. Moreover, the definitions provided are often imprecise, and sometimes they are even left totally implicit. A cynic, such as myself, might therefore suggest that the situation would be less confusing if we all confined ourselves to describing our models and their ability to account for the neuropsychological data, and made a conscious effort to avoid using words like "module" altogether.

7 Regularity and Frequency Confounds

Another connectionist model, that appears to be in even more direct conflict with the above discussion, is the past tense model of Marchman (1993). She used back-propagation to train a feedforward network to map from 45 input units representing the stem phonology, to 60 output units representing the corresponding past tense phonology, via 45 hidden units. In contradiction to all our previous arguments, she concluded that "the acquisition of regular verbs became increasingly susceptible to injury, while the irregulars were learned quickly and were relatively impervious to damage". So what is at the root of this opposite conclusion?

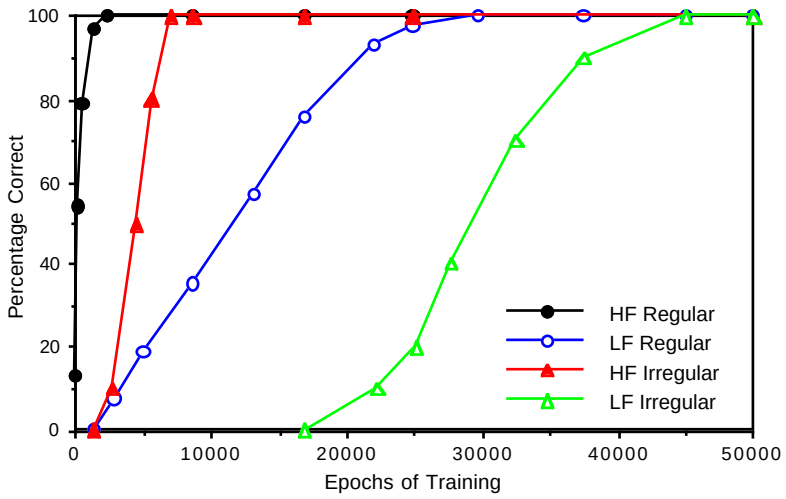

Figure 9: Performance improvements during the course of learning.

The crucial feature of her simulations was that the irregular items were presented to the network with frequencies of up to 15 times those of the regular items. On the face of it, this might seem eminently reasonable, given that the irregular items *are* more frequent than the regular items in English. However, it presents us with two fundamental problems:

1. It is far from obvious how the real word frequencies should map to training pattern frequencies in our over-simplified network models.

2. It is clearly going to confound the regularity and frequency effects that are observed in the models.

Fortunately, it is not difficult to investigate the regularity-frequency confound in our model, and hence understand what is happening in her model.

We have already noted that high frequency and high regularity both increase the rate of network learning and the subsequent robustness to damage. We can also see in Figures 5 and 6 that, in terms of the $Sum_i(P)$'s, it is possible to compensate for low regularity by higher frequency. By setting appropriate correct response thresholds on the output activations, it is straightforward to translate those $Sum_i(P)$ results into correct performance curves. Figure 9 shows how the performance of our simple model varies for the four item types during the course of learning. We see that our frequency ratio of 20 is sufficient for the frequency effect to swamp the regularity effect and allow the high frequency irregulars to be learnt more quickly than the low frequency regulars. This reversal of the natural regularity effect is exactly what Marchman found – though she confuses the issue by repeatedly referring to it as a "regularity effect" rather than a "frequency effect".

Taking global weight scaling as a smooth approximation to the removal of random network connections results in the pattern of performance loss shown in Figure 10. We see the patterns of damage follow from the patterns of learning as discussed above, with the low frequency regulars more susceptible than the high frequency irregulars. Again we have replicated Marchman's result – a "frequency effect" that is often inappropriately called a "regularity effect". Interestingly, by our careful matching of the frequency ratio to the degree of regularity, we have generated a crossover of the frequency and regularity effects. We see that there is potential for a weak double dissociation here, caused by the frequency and regularity effects coming into play at different rates (remember the resource

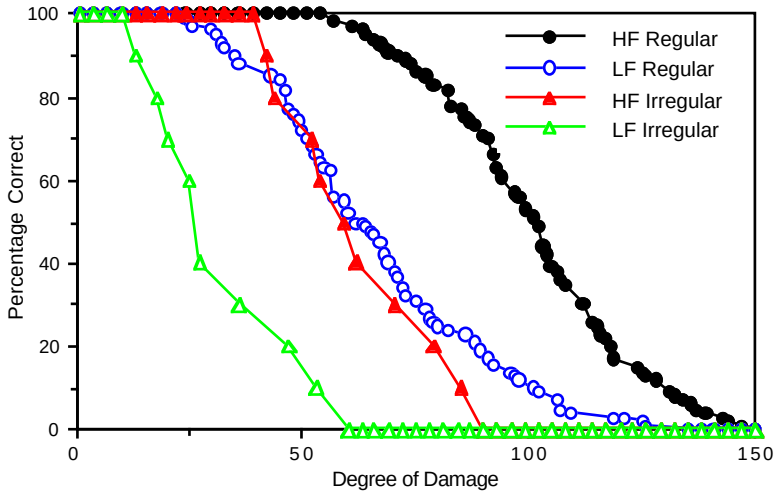

Figure 10: Performance loss due to increasing degrees of network damage.

artefact graph of Figure 3). But it remains to be seen if we can get stronger double dissociations in a similar manner.

It is actually quite straightforward to explore further the effect of different frequency ratios by explicit simulation. Again we take our simple feedforward network with 10 input units, 100 hidden units and 10 output units, but now we train it by gradient descent on just one set of 200 regular items and one set of 20 irregular items with a variable (Irregular/ Regular) frequency ratio. (This guarantees that we avoid any potential confounds caused by having two sets of each regularity type.) For each frequency ratio, we find that the learning curves take the familiar form of Figure 9, and lesioning the network continues to produce damage curves like Figure 10. The only unexpected result from this more systematic study is that the relative rates of fall off in performance turn out to be rather dependent on the type of damage inflicted.

If each trained network corresponds to a typical normal subject, then the network after different degrees of damage can be regarded as corresponding to a typical series of patients with different degrees of brain damage. Naturally, it is the data from the patients with the clearest dissociations that are the most well known, as they place the clearest and strongest constraints on our models, and these cases will inevitably correspond to the largest dissociations. It therefore makes sense for us to look for the largest dissociations in damage curves such as those of Figure 10. For a given trained network we can define the maximum dissociation in each direction as the maximum absolute percentage difference in performance between the two item types as the performance on them is reduced by damage from 100% to 0%. We can then determine how this varies with the frequency ratio and the type of damage. Figure 11 shows the maximum dissociations obtained for hidden unit removal versus connection removal as the frequency ratio varies over five orders of magnitude in our model.

We see that, by picking an appropriate frequency ratio, it is possible to get any dissociation we want. It is important not to take the frequency scale too literally though. First, the precise frequency ratio at the cross-over point will, of course, depend on the details of the regularity, which will rarely be as regular as the identity map that has been used in the simulations. In practice, in more realistic training sets, there will be a whole distribution of different regularities and frequencies to complicate matters. Secondly, matching real frequencies to appropriate training data distributions for networks that

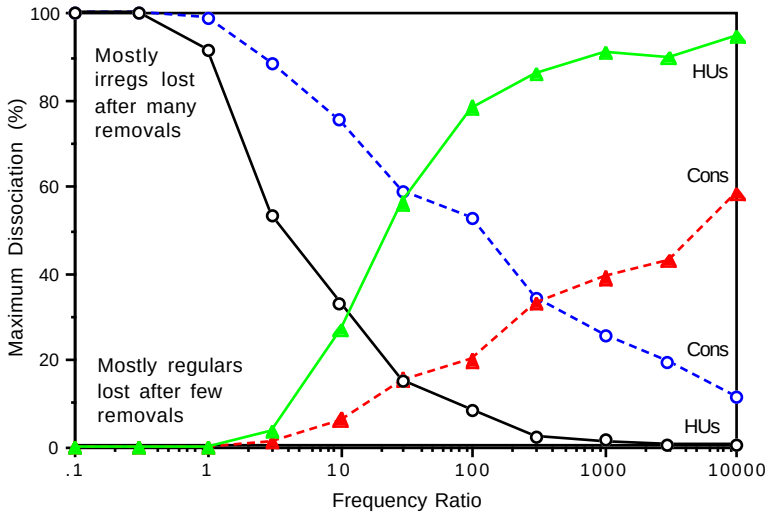

Figure 11: The dissociations depend on the frequency ratio and the damage type.

employ representations and learning algorithms of dubious biological plausibility is notoriously difficult. Seidenberg & McClelland (1989), for example, argued that a logarithmic compression of the real word frequencies was appropriate for use in the training data of their reading model, and this did produce good results. Later, more successful, reading models (e.g. Plaut et al., 1996; Bullinaria, 1997a) provide a better account for the empirical data if actual word frequencies are used. As Plaut et al. (1996) discuss in some detail, obtaining realistic interactions between frequency and regularity in a given model relies on coordinating the choice of input and output representations appropriate for the regularity, with the choice of training data frequencies. Getting this wrong can easily reverse the dissociation found in a given model. Conversely, a model that only gets the right dissociation by invoking doubtful choices of representations and frequencies should be viewed with some suspicion. Fortunately, there is plenty of non-neuropsychological data, such as reaction times and ages of acquisition, that will assist in constraining our models in this respect.

There are two further factors which will also affect the cross-over frequency ratios in Figure 11, and whether we get the resource artefact style DD seen in Figure 10. First, as noted above, relearning after damage will tend to enhance a dissociation between types or frequencies. In cases involving opposing frequency and regularity effects, the consequence of relearning will depend on the details. For example, in the case of Figures 9 and 10, the high frequency irregulars are learnt more quickly than the low frequency regulars, so the regulars lost dissociation will be enhanced and the irregulars lost dissociation reduced by relearning. In extreme cases, then, relearning may convert a double dissociation into a single dissociation or vice-versa, and in a single patient, a dissociation could be reversed. The second factor is the number of hidden units employed in the model, especially if it is near the minimal number required to perform the task in hand. It is well known from the field of modelling developmental disorders that dissociations can occur with poor performance on the last learned items if resources are limited, such as by restricting the number of hidden units (e.g. Plaut et al., 1996; Bullinaria, 1997a). This will again be model dependent, but is likely to have differential effects on frequency and regularity. It is perhaps worth noting that both these two factors apply to the past tense model of Marchman (1983) discussed above.

8 Connectionist Double Dissociation

Given that "box and arrow" models have provided good accounts of all manner of DD, and since one can always implement modular "box and arrow" models in terms of neural networks, it is clearly possible to obtain DD as a result of damage to connectionist systems. Exactly how the modules emerge in biological neural networks is still a matter of some debate, but this is another area where connectionist modelling may be of assistance (e.g. Jacobs, 1999; Bullinaria, 2001). However, all this still leaves the question of whether connectionist models can allow DD without modularity. In fact, since we know that there exist non-connectionist systems that can exhibit DD without modularity (e.g. Shallice, 1988; Dunn & Kirsner, 1988), and that these non-modular systems can too be implemented in terms of neural networks, it is clearly also possible to obtain connectionist DD without modularity (for appropriate definitions of the word "modularity").

One kind of non-modular system that sits particularly naturally within the framework of connectionist modelling, and yet can result in double dissociation when damaged, involves a continuum of processing space or topographic maps (Shallice, 1988, p249). These are not fully distributed systems in the sense used above, and so are perfectly consistent with our preceding conclusions. A particularly transparent example described by Shallice is that of the visual cortex. Damage resulting in deficits in different parts of the visual field can constitute a DD, yet there is no natural separation into modules. Such DD without modularity may also result from any other representations in the brain that take on a similar topographic form, and it is not difficult to see how these representations may arise naturally from restrictions on the neural connectivity with other sub-systems. For example, if semantic representations are of this form, then it is easy to see how localized damage could result in all manner of category specific and concrete-abstract deficits (Warrington & Shallice, 1984). An interesting connectionist model of optic aphasia involving topographic biases within semantics has recently been presented by Plaut (2002). More specific details of this type of system are highly problem dependent, and would take us too far from our discussion of general principles, so I will not present any explicit models here. The challenge is not just to get the models to produce dissociations, as we have seen that this is fairly straightforward, but to justify the chosen representations and relative degrees of connectivity necessary to give dissociations that match the patients. This is another area where explicit connectionist models can take us to a new level of refinement beyond the old "box and arrow" models.

We are finally left with the question of whether we can get DD in fully distributed models that have no non-connectionist analogue. We have seen above how it is possible to generate made to measure single dissociations in fully distributed networks, but it is not clear whether it is also possible to get double dissociations in this manner. Bullinaria & Chater (1995) suggest not, but they did not allow regularity-frequency confounds of the type discussed above. Consider the cross-over point in Figure 11 where there are strong dissociations in both directions (i.e. around a frequency ratio of 30). The actual network performance levels here are plotted in Figure 12. For both lesion types, the pattern of dissociation reverses as we increase the amount of damage. We begin with a mostly regulars lost dissociation but, after many removals, end with a mostly irregulars lost dissociation. We see that, for particular degrees of damage, it is possible to obtain a cross-over double dissociation between high frequency irregulars and low frequency regulars. However, to get it we need an interaction between two carefully balanced factors (e.g. regularity and frequency) that "act" in the same way but at different rates, and two different types of damage (e.g. hidden unit and connection removal) that "act" in the same way but at different rates.

So, by carefully balancing factors like regularity and frequency, one *can* get cross-over

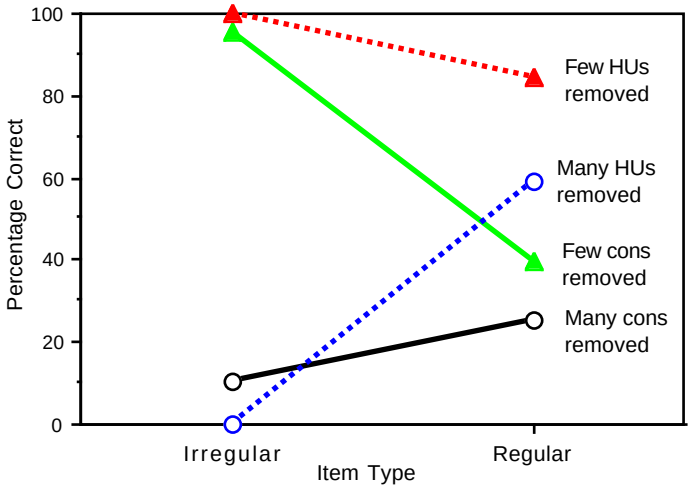

Figure 12: Carefully chosen parameters can result in a connectionist cross-over DD.

DD in a fully distributed system, and it is inevitable that other factors known to give dissociation (such as representation sparseness) will also be able to result in DD in a similar manner. Given the discussion of Dunn & Kirsner (1988), this should not be too much of a shock, but it does complicate our modelling endeavors. We are left with a number of questions: Is this not rather unnatural? Can it really happen like this in real life? And if so, what does it mean? In language, for example, the relation between word regularities and word frequencies is not just random. Hare & Elman (1995) have shown how language evolution naturally results in a correlation between irregularity and frequency because irregular words tend to get lost from the language or regularized unless they are high frequency. In this way, a balancing of the effects of frequency and regularity can really happen, and thus it seems that real language does have a built in confound. Similar natural confounds are likely to arise as a result of evolution in other areas as well. Whether this is the right way to account for particular real DDs is something that will need to be investigated on a case by case basis. As always, the modeller simply needs to take each set of empirical data at face value and examine how it might be modelled, irrespective of any confounds the experimenter has failed to remove.

9 Conclusions

This work grew out of repeated questions and confusion concerning the consistency of the conclusions of Bullinaria & Chater (1995) with the properties of explicit network simulations by other researchers that apparently gave conflicting results. The work of Marchman (1993) and Plaut (1995) seemed to provide particularly strong counter-examples. Hopefully the above discussion has convinced the reader that all the network simulation results are actually in agreement – and that the apparent inconsistencies are merely in the terminology.

We have seen that a general feature of neural network models is that regularity and frequency and various related factors (such as representation consistency, strength and correlation) all result in increased rates and accuracy of learning, and these in turn result in increased resilience to network damage. This simple fact is at the root of most of the results that have come out of connectionist lesion studies. A major problem in comparing our connectionist models with empirical patient data is that the causes of these differential

effects are easily confused. Clearly, if one wants to make reliable claims about one factor, such as regularity, one has to be very careful about controlling for frequency and the other factors. The model of Marchman (1993), for example, has a regularity effect that has been reversed by a larger frequency effect. Moreover, it is also probably worth noting here that the problematic confounds we have been discussing will automatically follow through to secondary measures such as reaction times and priming. Unfortunately, this is not just a problem for connectionist modellers, it is at least equally problematic for experimenters on human subjects. As noted by Shallice (1988, p239), even basic questions, such as what frequency distributions did a given subject learn from, are generally unanswerable. And even if we did know, the nature of many natural tasks, like language, is such that it would be virtually impossible to control for all the potential confounds anyway.

In conclusion, it seems clear that connectionism has much to offer in the fleshing out of the details of earlier "box and arrow" models, or even in replacing them completely, to provide more complete accounts of cognitive processing. The resulting enhanced models and the new field of connectionist neuropsychology are not only producing good accounts of existing empirical data, but are also beginning to suggest more appropriate experimental investigations for further fine tuning of these models, and an ethical approach for exploring potential remedial actions for neuropsychological patients.

Acknowledgements

This chapter presents extensions of work originally carried out in collaboration with Nick Chater while we were both at the University of Edinburgh and supported by the MRC. I continued the work while at Birkbeck College London supported by the ESRC, and at the University of Reading supported by the EPSRC.

References

Bishop, C.M. (1995). *Neural Networks for Pattern Recognition*. Oxford, UK: Oxford University Press.

Bullinaria, J.A. (1994). Internal Representations of a Connectionist Model of Reading Aloud. *Proceedings of the Sixteenth Annual Conference of the Cognitive Science Society*, 84-89. Hillsdale, NJ: Erlbaum.

Bullinaria, J.A. (1997a). Modelling Reading, Spelling and Past Tense Learning with Artificial Neural Networks. *Brain and Language*, **59**, 236-266.

Bullinaria, J.A. (1997b). Modelling the Acquisition of Reading Skills. In A. Sorace, C. Heycock & R. Shillcock (Eds), *Proceedings of the GALA '97 Conference on Language Acquisition*, 316-321. Edinburgh: HCRC.

Bullinaria, J.A. (1999). Free Gifts from Connectionist Modelling. In R. Baddeley, P. Hancock & P. Földiák (Eds), *Information Theory and the Brain*, 221-240. Cambridge: Cambridge University Press.

Bullinaria, J.A. (2001). Simulating the Evolution of Modular Neural Systems. In *Proceedings of the Twenty-third Annual Conference of the Cognitive Science Society.* 146-151. Mahwah, NJ: Lawrence Erlbaum Associates.

Bullinaria, J.A. & Chater, N. (1995). Connectionist Modelling: Implications for Cognitive Neuropsychology. *Language and Cognitive Processes*, **10**, 227-264.

Bullinaria, J.A., Riddell, P.M. & Rushton, S.K. (1999). Regularization in Oculomotor Adaptation. In *Proceedings of the European Symposium on Artificial Neural Networks*, 159-164. Brussels: D-Facto.

Caramazza, A. (1986). On Drawing Inferences About The Structure Of Normal Cognitive

Systems From Analysis Of Patterns Of Impaired Performance: The Case Of Single-Patient Studies. *Brain and Cognition*, **5**, 41-66.

Caramazza & McCloskey, (1989). Number System Processing: Evidence from Dyscalculia. In N. Cohen, M. Schwartz & M. Moscovitch (Eds), *Advances in Cognitive Neuropsychology*. New York, NY: Guildford Press.

Coltheart, M. (1985). Cognitive Neuropsychology and the Study of Reading. In M.I. Posner & O.S.M. Marin (Eds), *Attention and Performance XI*. Hillsdale, NJ: Erlbaum.

Coltheart, M., Curtis, B., Atkins, P. & Haller, M. (1993). Models of Reading Aloud: Dual-Route and Parallel-Distributed-Processing Approaches. *Psychological Review*, **100**, 589-608.

De Renzi, E. (1986). Current Issues in Prosopagnosia. In H.D. Ellis, M.A. Jeeves, F. Newcome & A. Young (Eds), *Aspects of Face Processing*. Dordretcht: Martinus Nijhoff.

Devlin, J.T., Gonnerman, L.M., Andersen, E.S. & Seidenberg, M.S. (1998). Category-Specific Semantic Deficits in Focal and Widespread Brain Damage: A Computational Account. *Journal of Cognitive Neuroscience*, **10**, 77-94.

Dunn, J.C. & Kirsner, K. (1988). Discovering Functionally Independent Mental Processes: The Principle of Reversed Association. *Psychological Review*, **95**, 91-101.

Farah, M.J. (1990). *Visual Agnosia: Disorders of Object Recognition and What They Tell Us About Normal Vision*. Cambridge, MA: MIT Press.

Farah, M.J. (1994). Neuropsychological Inference with an Interactive Brain: A Critique of the Locality Assumption. *Behavioral and Brain Sciences*, **17**, 43-104.

Fodor, J.A. (1983). *The Modularity of the Mind*. Cambridge, MA: MIT Press.

Funnell, E. (1983). Phonological Processing in Reading: New Evidence from Acquired Dyslexia. *British Journal of Psychology*, **74**, 159-180.

Geshwind, N. (1985). Mechanisms of Change After Brain Lesions. *Annals of the New York Academy of Sciences*, **457**, 1-11.

Hare, M. & Elman, J.L. (1995). Learning and Morphological Change. *Cognition*, **56**, 61-98.

Harley, T.A. (1996). Connectionist Modeling of the Recovery of Language Functions Following Brain Damage. *Brain and Language*, **52**, 7-24.

Hinton, G.E. (1989). Connectionist Learning Procedures. *Artificial Intelligence*, **40**, 185-234.

Hinton, G.E. & Sejnowski, T.J. (1986). Learning and Relearning in Boltzmann Machines. In D.E. Rumelhart & J.L. McClelland (Eds) *Parallel Distributed Processing: Explorations in the Microstructures of Cognition*, Vol. 1, 282-317. Cambridge, MA: MIT Press.

Hinton, G.E. & Shallice, T. (1991). Lesioning an Attractor Network: Investigations of Acquired Dyslexia. *Psychological Review*, **98**, 74-95.

Jacobs, R.A. (1999). Computational Studies of the Development of Functionally Specialized Neural Modules. *Trends in Cognitive Science*, **3**, 31-38.

Lavric, A., Pizzagalli, D., Forstmeir, S. & Rippon, G. (2001). Mapping Dissociations in Verb Morphology. *Trends in Cognitive Science*, **5**, 301-308.

Marchman, V.A. (1993). Constraints on Plasticity in a Connectionist Model of the English Past Tense. *Journal of Cognitive Neuroscience*, **5**, 215-234.

McCarthy, R. A. & Warrington, E. K. (1986). Phonological Reading: Phenomena and Paradoxes. *Cortex*, **22**, 359-380.

Patterson, K., & Marcel, A. (1977). Aphasia, Dyslexia and the Phonological Coding of Written Words. *Quarterly Journal of Experimental Psychology*, **29**, 307-318.

Patterson, K., & Marcel, A. (1992). Phonological ALEXIA or PHONOLOGICAL Alexia?

In J. Alegria, D. Holender, J. Junça de Morais & M. Radeau (Eds), *Analytic Approaches to Human Cognition*, 259-274. Amsterdam: Elsevier.
Pinker, S. (1991). Rules of Language. *Science*, **253**, 530-535.
Pinker, S. (1997). Words and Rules in the Human Brain. *Nature*, **387**, 547-548.
Plaut, D.C. (1995). Double Dissociation Without Modularity: Evidence from Connectionist Neuropsychology. *Journal of Clinical and Experimental Neuropsychology*, **17**, 291-321.
Plaut, D.C. (1996). Relearning After Damage in Connectionist Networks: Towards a Theory of Rehabilitation. *Brain and Language*, **52**, 25-82.
Plaut, D.C. (2002). Graded Modality-Specific Specialisation in Semantics: A Computational Account of Optic Aphasia. *Cognitive Neuropsychology*, **19**, 603-639.
Plaut, D.C., McClelland, J.L., Seidenberg, M.S. & Patterson, K.E. (1996). Under-standing Normal and Impaired Word Reading: Computational Principles in Quasi-Regular Domains. *Psychological Review*, **103,** 56-115.
Plaut, D.C. & Shallice, T. (1993). Deep Dyslexia: A Case Study of Connectionist Neuropsychology. *Cognitive Neuropsychology*, **10,** 377-500.
Sartori, G. (1988). From Neuropsychological Data to Theory and Vice Versa. In G. Denes, P. Bisiacchi, C. Semenza, E. Andrewsky (Eds), *Perspectives in cognitive neuropsychology*. London: Erlbaum.
Seidenberg, M.S. & McClelland, J.L. (1989). A Distributed, Developmental Model of Word Recognition and Naming. *Psychological Review*, **96**, 523-568.
Sejnowski, T.J. & Rosenberg, C.R. (1987). Parallel Networks that Learn to Pronounce English Text. *Complex Systems*, **1**, 145-168.
Shallice, T. (1979). Neuropsychological Research and the Fractionation of Memory Systems. In L.G. Nilsson (Ed.), *Perspectives on Memory Research*. Hillsdale, NJ: Erlbaum.
Shallice, T. (1988). *From Neuropsychology to Mental Structure*. Cambridge: Cambridge University Press.
Shoben, E.J., Wecourt, K.T. & Smith, E.E. (1978). Sentence Verification, Sentence Recognition, and the Semantic-Episodic Distinction. *Journal of Experimental Psychology: Human Learning and Cognition*, **4**, 304-317.
Small, S.L. (1991). Focal and Diffuse Lesions in Cognitive Models. *Proceedings of the Thirteenth Annual Conference of the Cognitive Science Society*, 85-90. Hillsdale, NJ: Erlbaum.
Teuber, H.L. (1955). Physiological Psychology. *Annual Review of Psychology*, **9**, 267-296.
Warrington, E.K. (1981). Concrete Word Dyslexia. *British Journal of Psychology*, **72**, 175-196.
Warrington, E.K. (1985). Agnosia: The Impairment of Object Recognition. In P.J. Vinken, G.W. Bruyn & H.L. Klawans (Eds), *Handbook of Clinical Neurology.* Amsterdam: Elsevier.
Warrington, E.K. & Shallice, T. (1984). Category Specific Semantic Impairments. *Brain*, **107**, 829-853.
Wilson, B. & Patterson, K.E. (1990). Rehabilitation for Cognitive Impairment: Does Cognitive Psychology Apply? *Applied Cognitive Psychology*, **4**, 247-260.
Wood, C.C. (1978). Variations on a Theme of Lashley: Lesion Experiments on the Neural Model of Anderson, Silverstein, Ritz & Jones. *Psychological Review*, **85,** 582-591.