In: *Proceedings of the Fifth Neural Computation and Psychology Workshop*, 52-63. London: Springer-Verlag, 1999.

Connectionist Dissociations, Confounding Factors and Modularity

John A. Bullinaria

Department of Psychology, The University of Reading
Reading, RG6 6AL

Abstract

Although much has been written on this subject, there still seems to be considerable confusion in the literature concerning dissociations, double dissociations and what they really mean, especially when connectionist or neural network models are involved. In this paper I attempt to clarify matters by looking at the subject from the point of view of patterns of learning rates in neural network models.

1 Introduction

Neuropsychological data places strong constraints on mental structure (e.g. [1]). In particular, Double Dissociation (DD) is traditionally taken to imply modularity within language processing systems (e.g. [9, 10]) and in many other areas such as visual processing and memory (e.g. [13]). Bullinaria & Chater [3] have investigated whether strong (cross-over) DDs are also possible after damaging fully distributed connectionist systems and concluded that without modularity only single dissociations are possible (assuming one successfully avoids small scale artefacts). Moreover, these single dissociations were seen to be natural regularity effects with the regular mappings more robust than the irregular. These general arguments have 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 [1, 2]. Marchman [8], however, has looked at related models of past tense production and seemingly found dissociations with the irregular items more robust, and Plaut [11] claims to have found a connectionist DD without modularity. Naturally, these apparent contradictions have caused a certain amount of confusion, particularly amongst researchers unfamiliar with the detailed workings of connectionist models. Here I shall extend the work of Bullinaria & Chater [3] with view to minimising future confusion in this area.

The models we shall look at will all have the same simplified structure of a feed-forward network with one hidden layer trained by some form of gradient descent 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 hence they get learnt more quickly and accurately and require 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

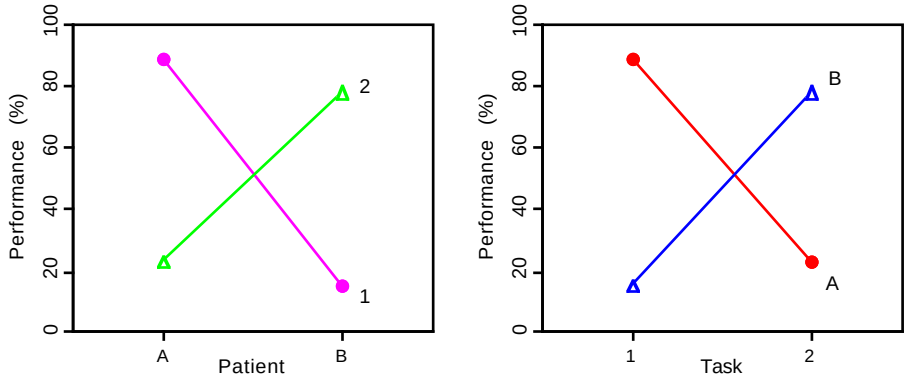

Figure 1: A strong cross-over double dissociation for Tasks 1 & 2, Patients A & B.

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.) I shall present explicit neural network simulation results and argue that once one controls for such confounding effects, the basic conclusion of Bullinaria & Chater [3] holds and the opposite 'regularity' effect found by Marchman [8] is more correctly called a 'frequency' effect. We shall also see how Plaut's DD is consistent with our findings. In passing, I will show how it is possible to obtain a valid strong DD 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.

I will begin by reviewing a few things that *should* be well known: the traditional inference from double dissociation to modularity, the problem of *resource artefacts*, the types of system that may exhibit DDs, and some basic properties of connectionist models. Then I will present some explicit neural network lesion simulations. First, the general effects of network damage as discussed by Bullinaria & Chater [3], then the apparently contradictory results of Plaut [11] and Marchman [8], and finally 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 general discussion and some conclusions.

2 Double Dissociation ⇒ Modularity ?

The various types of dissociation have been discussed in detail by Shallice [13] and Dunn & Kirsner [6]. If a patient performs very much better on Task 1 than on Task 2, then we have a strong *single dissociation*. If two patients have opposite dissociations, then we have a *double dissociation* (as in Figure 1). Such a double dissociation (DD) in performance is usually explained by the existence of separate modules. A classic example occurs in reading: the loss of exception words in surface dyslexics and the loss of non-words in phonological dyslexics constitutes a

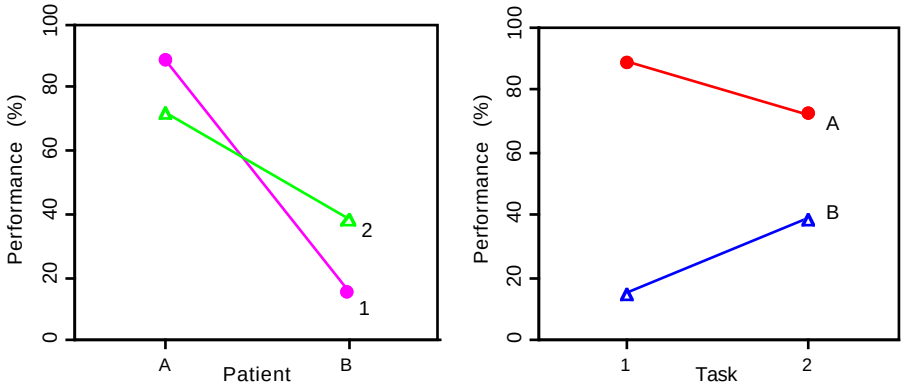

Figure 2: A weak double dissociation for Tasks 1 & 2, Patients A & B.

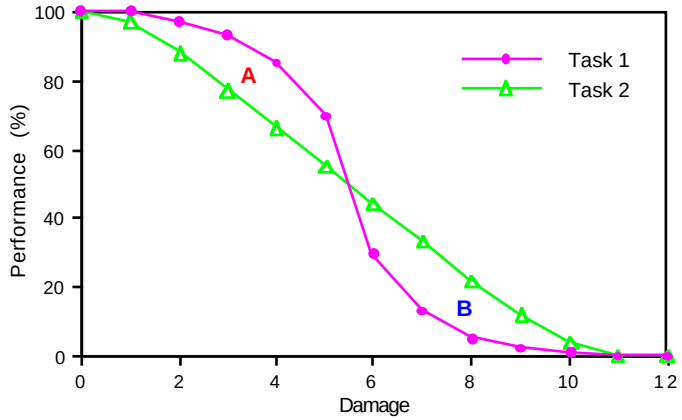

Figure 3: Tasks depending on the resources in different ways can lead to DDs.

DD which can be taken to imply separate Lexical and Rule-Based modules in a Dual Route Model of reading (e.g. by Coltheart et al. [4]). In fact, double dissociations are taken quite generally to imply modularity, even though Dunn & Kirsner [6] have shown that this cannot generally be justified. Moreover Shallice [13, p249] lists a number of non-modular systems that can produce dissociations (even double dissociations) when damaged (e.g. topographic maps, overlapping processing regions, coupled systems). However he claims that "as yet there is no suggestion that a strong DD can take place from two lesions within a properly distributed network".

Before moving on to test this claim, we should note one final 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 and should NOT be taken to imply modularity [13, p 234]. As we shall see later, such DDs are easily obtainable in connectionist models. Devlin et al. [5] present an interesting example involving a connectionist account of category specific semantic deficits.

3 Neural Network Models

In this paper I will be primarily concerned with simple feed-forward networks that map between our chosen input and output representations via a single hidden layer. Extensions to more complicated systems will be readily apparent. The important feature is that the network *learns* to perform its given task by iteratively adjusting the connection weights (e.g. by gradient descent) to minimise the output errors for appropriate sets of input-output pairs. We can then compare the development of the networks performance and its final performance (e.g. its output errors, generalization ability, RTs, 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 we are particularly interested in simulating neuropsychological effects by lesioning our trained networks.

The output of each network unit *i* for each training pattern *P* will be the sigmoid of the sum of the weighted activations flowing into it from the previous layer:

$$Out_i(P) = sigmoid(Sum_i(P)) \quad , \quad Sum_i(P) = \sum_j w_{ij} Prev_j(P) \ .$$

To train the network we define a typical sum-squared output error measure:

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

and iteratively update the weights by gradient descent to reduce this error:

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

It then follows straightforwardly from adding up the weight change contributions due to individual training patterns that:

1. Regular items will get learnt more quickly than irregular items.
2. High frequency items get learnt more quickly than low frequency items.
3. Ceiling effects will arise as the sigmoids saturate and the Δw's tend to zero.

It is feasible to explore this explicitly in small networks trained simultaneously to perform sets of regular and irregular mappings of varying frequency.

4 Learning and Damage Curves

Consider first a simple feedforward net with 10 inputs, 100 hidden units and 10 outputs trained by back-propagation with binary targets on two sets of 100 regular items (permuted identity mappings) and two sets of 10 irregular items (random mappings). For each regularity we have one set appearing during training with a frequency of 20 times the other. These frequency differences can be implemented over many epochs by manipulating the probability that a given pattern is used for training in a given epoch, or within a single epoch by scaling the weight change contributions. As long as the weight changes per epoch are kept small it seems to

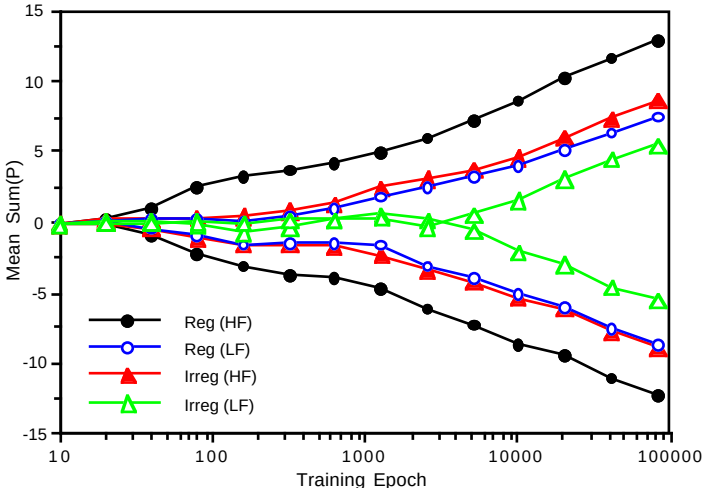

Figure 4: Learning curves for a simple network trained on quasi-regular mappings.

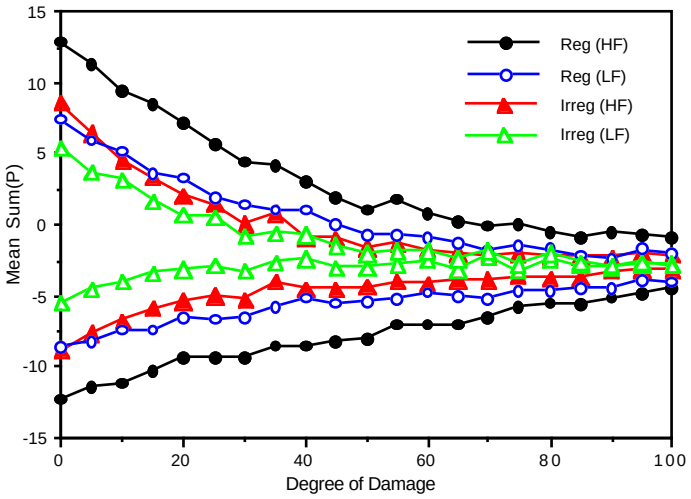

Figure 5: Damage curves corresponding to Figure 4 due to removal of connections.

make little difference. The predicted regularity and frequency effects are seen clearly in Figure 4 which shows how the mean output $Sum_i(P)$'s develop during training for each of the four item types and two target activations.

Bullinaria & Chater [3] found that damaging trained networks by removing random hidden units, removing random connections, adding random noise to the weights, or globally scaling 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 5 shows the effect of removing increasingly large numbers of connections from our network – we see the reverse of the pattern of learning in Figure 4. If we set a particular correct response threshold for the $Sum_i(P)$'s, e.g. ±2.2

corresponding to output activations 0.1 and 0.9, we see that the first items to be learnt during training tend to be the last to be lost during damage, and hence we get clear dissociations with the regulars more robust than the irregulars. These basic effects extend easily to more realistic models, for example, surface dyslexia in the reading model of Bullinaria [1, 2]. They also allow us to understand various developmental and reaction time effects.

The general point to be made is that some items are naturally learnt more quickly and more accurately than others and the effects of subsequent network damage follow automatically from the patterns of learning. There are actually many factors which can cause the differing learning and damage rates and we can explore them all in a similar manner, for example: Regularity and Frequency (as discussed above), Neighbourhoods and Consistency, (which we noted above are related to regularity), Pattern Strength or Sparseness (e.g. as used by Plaut & Shallice [12] and Plaut [11] to distinguish concrete and abstract semantics), Correlation, Redundancy and Dimensionality (e.g. as used by Devlin et al. [5] to model the semantics of natural things versus artefacts), and so on. At some level of description, they can all be regarded as forms of regularity, but clearly, if one wants to make claims about one effect, one needs to control for the others.

5 Disappearing Double Dissociations

I have glossed over many issues discussed in more detail by Bullinaria & Chater [3], but one complication is of such importance that it is worth repeating here. The damage curves of Figure 5 are relatively smooth because we have averaged over many output units and many training items, and because our network has many more 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. However, as we scale up to larger networks with sufficient numbers of hidden units and connections that individual contributions each have a small effect on the outputs, the 'regulars lost' dissociations disappear. We always find the apparent double dissociations dissolve into single dissociations as we make the network more distributed. We are then left with a simple 'regularity effect' as discussed above.

So how many hidden units do we need for reliable results? In effect, we need to make sure that individual processing units are not acting as 'modules' in their own right. We can check that a network is fully distributed by seeing that the individual contributions $w_{ij}Prev_j(P)$ to an output unit i are small compared to the total $Sum_i(P)$, i.e. that the ratios of the contributions to the corresponding totals are less than one. Figure 6 shows the distribution of 10000 typical individual contribution ratios 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 larger than one, but with only 30 hidden units, many contributions are much greater than their corresponding total and their removal will

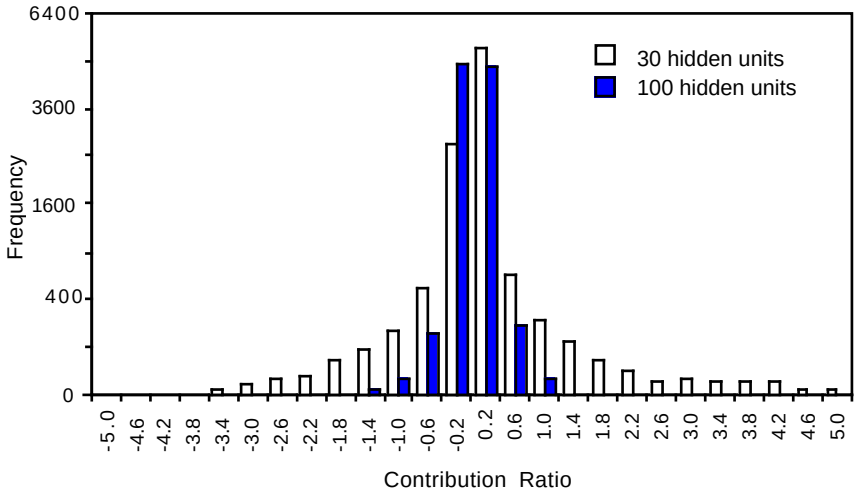

Figure 6: The distribution of output contribution ratios for two typical networks.

result in wild fluctuations in the outputs. Unfortunately, in general, it seems that we really do need lots of hidden units – many times the minimal number required to learn the given task. I'm often asked what can be done if limited computational resources make this impossible. Well, consider the effect of network damage on the histograms in Figure 6. Naturally, 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 $\alpha\, Sum_i(P) = \alpha \sum w_{ij} Prev_j(P)$. Note that we achieve exactly the same result by just 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 corresponds to a number of real neurons, then the weight scaling can be regarded as removing a fraction α of these corresponding neurons.

6 But what about Plaut, 1995 ?

Plaut's well known paper entitled "Double dissociation without modularity: Evidence from connectionist neuropsychology" [11] is often taken as evidence that there must be something wrong with the above discussion. Lesions in two different locations in an attractor network were found to produce a DD between concrete and abstract word reading if the concreteness is 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 reduced performance on the concrete

words. Although the two damage locations do not constitute modules in the conventional sense, we can easily see how they contribute different degrees to the processing of the two word types and will give opposite dissociations when damaged [11, 12]. The performance of each location is fully consistent with the above discussion, and the only disagreement is over the appropriateness of the use of the word 'module' to describe the two locations.

7 But what about Marchman, 1993 ?

Marchman [8] presented a past tense model that is in more direct conflict with the above. She used back-propagation to train a feedforward network with 45 inputs (for the stem phonology), 45 hidden units and 60 output units (for the corresponding past tense phonology). In contradiction to the above, 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?

The crucial feature of her simulation was that irregular items were presented to the network with frequencies up to 15 times that of the regular items. This might seem reasonable, given that irregular items *are* more frequent than regular items in English. However, it presents us with two 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.

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

8 Regularity and Frequency Performance Effects

We have already noted that high frequency and high regularity both increase the rate of network learning and the robustness to damage. We can also see in Figures 4 and 5 that, in terms of the $Sum_i(P)$'s, it is possible to compensate for low regularity by high frequency. By setting appropriate correct response thresholds on the output activations it is straightforward to translate these results into correct performance curves. Figure 7 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 what Marchman found [8] – though she repeatedly refers 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, Figure 8 shows the patterns of damage following from the pattern of learning as discussed above. Again Marchman [8] gets her "frequency effect" but generates unnecessary confusion by calling it a "regularity effect". Interestingly, by carefully matching the frequency ratio to the degree of regularity,

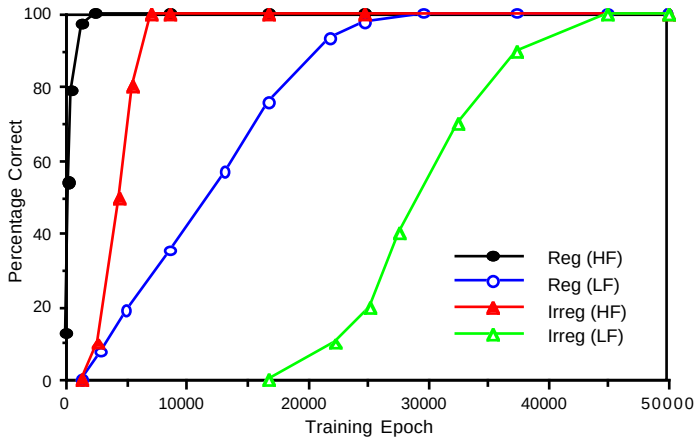

Figure 7: Performance improvements during the course of learning.

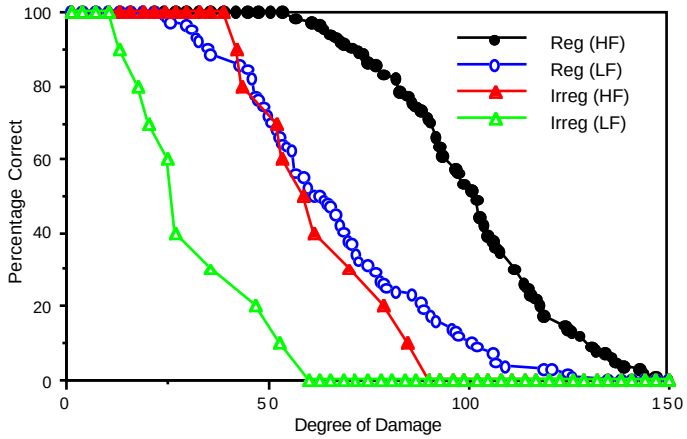

Figure 8: Performance loss due to increasing degrees of network damage.

we get a crossover of the frequency and regularity effects. We see that there is potential for a weak double dissociation caused by the frequency and regularity effects coming into play at different rates (remember that resource artefact graph of Figure 3). But can we get stronger double dissociations in a similar way?

9 Dissociations as a Function of Frequency Ratio

It is not difficult to explore the effect of the frequency ratio. Again we take our simple feedforward network with 10 inputs, 100 hidden units and 10 outputs, but now we train it by back-propagation 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

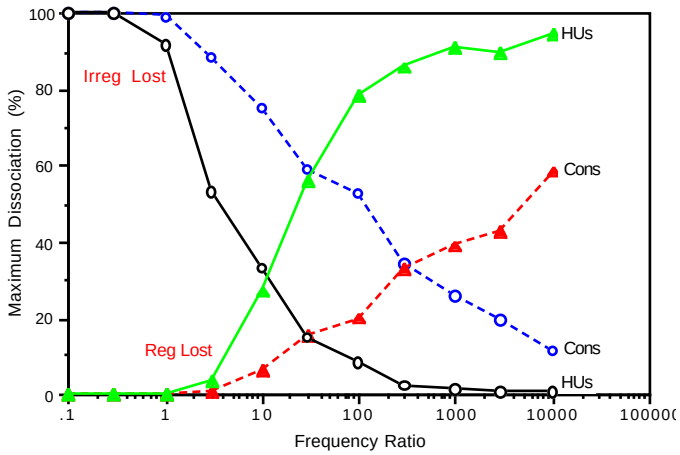

Figure 9: The dissociation depends on the frequency ratio and the damage type.

familiar form of Figure 7 and lesioning the network continues to produce damage curves like Figure 8. The only unexpected result from this more systematic study is that the relative rates of fall off in performance turn out to depend on the type of damage inflicted.

For a given trained network we can define the maximum dissociation of each type as the maximum 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 9 shows the maximum dissociations obtained for hidden unit removal versus connection removal as the frequency ratio varies over five orders of magnitude.

We see that it is possible to get any dissociation we want by picking an appropriate frequency ratio. The precise ratio at the cross-over point will, of course, depend on the details of the regularity, and in more realistic training sets there will be a whole distribution of different regularities and frequencies to complicate matters. The cross-over ratio will also be affected by allowing retraining after damage (as Marchman did [8]) – especially if the networks are near minimal (as Marchman used [8]).

10 A Strong Connectionist Double Dissociation ?

Made to measure single dissociations are one thing, but can we get double dissociations in this manner? Consider the cross-over point in Figure 9 where we have strong dissociations in both directions (i.e. around a frequency ratio of 30). The actual performances here are plotted in Figure 10. For both damage types, the pattern of dissociation reverses as we increase the number of removals. We begin with a regulars lost dissociation but, after many removals, end with an irregulars lost dissociation. We see that, for particular degrees of damage, it is possible to obtain a strong cross-over double dissociation between high frequency irregulars and low frequency regulars. However, to get it we need an interaction between two carefully

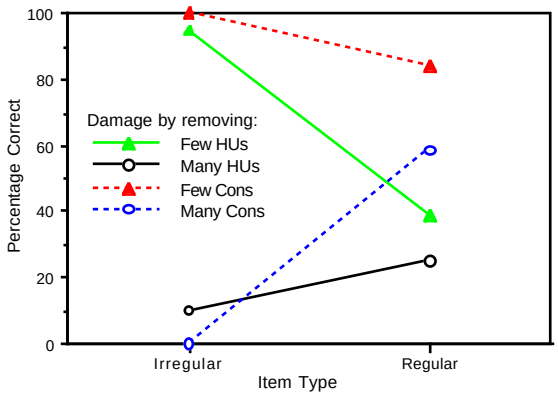

Figure 10: Carefully chosen parameters can result in a connectionist cross-over DD.

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. connection and hidden unit removal) that "act" in the same way but at different rates.

11 Conclusions

This work grew out of repeated questions and confusion over the consistency of the conclusions of Bullinaria & Chater [3] with those of Marchman [8] and Plaut [11]. I have hopefully convinced the reader that the network simulation results are in agreement – and that the apparent inconsistency is purely in the terminology.

Regularity and frequency and various related factors (such as consistency, strength and correlation) are all seen to result in increased rates and accuracy of learning which in turn results in an increased resilience to network damage. The problem is that the causes of these differential effects are easily confused. Clearly, if one wants to make reliable claims about regularity, one has to be very careful about controlling for frequency and the other factors. Marchman [8], for example, didn't control for frequency and ended up with a "regularity effect" that is really a "frequency effect".

We have also seen that by carefully balancing factors like regularity and frequency, one can get strong double dissociation without modularity. Given the discussion of Dunn & Kirsner [6], this should not be too much of a shock, but it does complicate our modelling endeavors. We are left with the question: Can this happen in real life, and if so, what does it mean? In reality, word frequencies are not just random. Hare & Elman [7] have shown how language evolution naturally results in a correlation between irregularity and frequency and so a balancing of the effects of frequency and regularity does really happen. It seems that real language does have an inbuilt confound. If we are not careful, we will be equally justified in claiming separate modules for high and low frequencies as we are for regulars and irregulars.

Finally, it is probably worth commenting here that the problematic confounds we have been discussing will automatically follow through to secondary measures such as reaction times and priming. Moreover, this is not just a problem for connectionist

modellers, it is at least equally as problematic for experimenters on human subjects. As noted by Shallice [13, p239], even basic questions, such as what frequency distributions did a given subjects learn from, are generally unanswerable. And even if we did know, it would inevitably be such that it would be virtually impossible to control for all the potential confounds.

Acknowledgements

This paper presents extensions of work originally carried out in collaboration with Nick Chater while we were both at Edinburgh University and funded by the MRC. The work is currently funded by the EPSRC.

References

1. 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.
2. Bullinaria, J.A. (1997). Modelling Reading, Spelling and Past Tense Learning with Artificial Neural Networks. *Brain and Language*, **59**, 236-266.
3. Bullinaria, J.A. & Chater, N. (1995). Connectionist Modelling: Implications for Cognitive Neuropsychology. *Language and Cognitive Processes*, **10**, 227-264.
4. 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.
5. 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.
6. Dunn, J.C. & Kirsner, K. (1988). Discovering Functionally Independent Mental Processes: The Principle of Reversed Association. *Psychological Review*, **95**, 91-101.
7. Hare, M. & Elman, J.L. (1995). Learning and Morphological Change. *Cognition*, **56**, 61-98.
8. Marchman, V.A. (1993). Constraints on Plasticity in a Connectionist Model of the English Past Tense. *Journal of Cognitive Neuroscience*, **5**, 215-234.
9. Pinker, S. (1991). Rules of Language. *Science*, **253**, 530-535.
10. Pinker, S. (1997). Words and Rules in the Human Brain. *Nature*, **387**, 547-548.
11. Plaut, D.C. (1995). Double Dissociation Without Modularity: Evidence from Connectionist Neuropsychology. *Journal of Clinical and Experimental Neuropsychology*, **17**, 291-321.
12. Plaut, D.C. & Shallice, T. (1993). Deep Dyslexia: A Case Study of Connectionist Neuropsychology. *Cognitive Neuropsychology*, **10,** 377-500.
13. Shallice, T. (1988). *From Neuropsychology to Mental Structure*. Cambridge: Cambridge University Press.