DISSOCIATION IN CONNECTIONIST SYSTEMS

John A. Bullinaria

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

Connectionist techniques are increasingly being used to model cognitive function with a view to providing extensions, elaborations or replacements of earlier "box and arrow" models. Networks of simplified processing units loosely based on real neurons are set up with architectures based on known physiology, trained to perform appropriately simplified versions of real tasks, and iteratively refined by checking their performance against humans. Such systems can still be linked together as in the old box and arrow models, with all the old explanations of patient data carrying through, but now we can examine the details of the degradation of the various components, and removing neurons or connections constitute natural analogues of real brain damage. We can also question the validity of the old assumptions of neuropsychological inference, and explore the possibility that processing is actually more distributed and interactive than the older models implied (Bullinaria, 2002). Here I shall outline what I consider to be the essential ideas.

Connectionist models *learn* to perform by iteratively updating their weights (e.g. by gradient descent) to minimise the output errors for appropriate training sets of input-output pairs. Adding up the network weight change contributions due to individual training patterns explains why:

1. High frequency items are learned more quickly, because the appropriate weight changes get applied more often.
2. Regular items are learned more quickly, because consistent weight changes combine while inconsistent weight changes cancel.
3. Ceiling effects arise when items are mastered.

These effects are easily demonstrated in a standard feed-forward network with 10 inputs, 100 hidden units and 10 outputs. Training on two sets of 100 regular items and two sets of 10 irregular items, with one regular set and one irregular set presented 20 times more frequently than the other, results in the learning curves of Figure 1 (Bullinaria, 1999).

Simulating brain lesions in connectionist systems was discussed by Small (1991). Bullinaria and Chater (1995) found that very similar patterns of deficits arose by randomly removing hidden units, randomly removing connections, globally scaling the weights, or adding random noise to the weights. If small scale artefacts were avoided, and all other factors controlled for, only single dissociations were found. Each damage type results in the activation feeding into

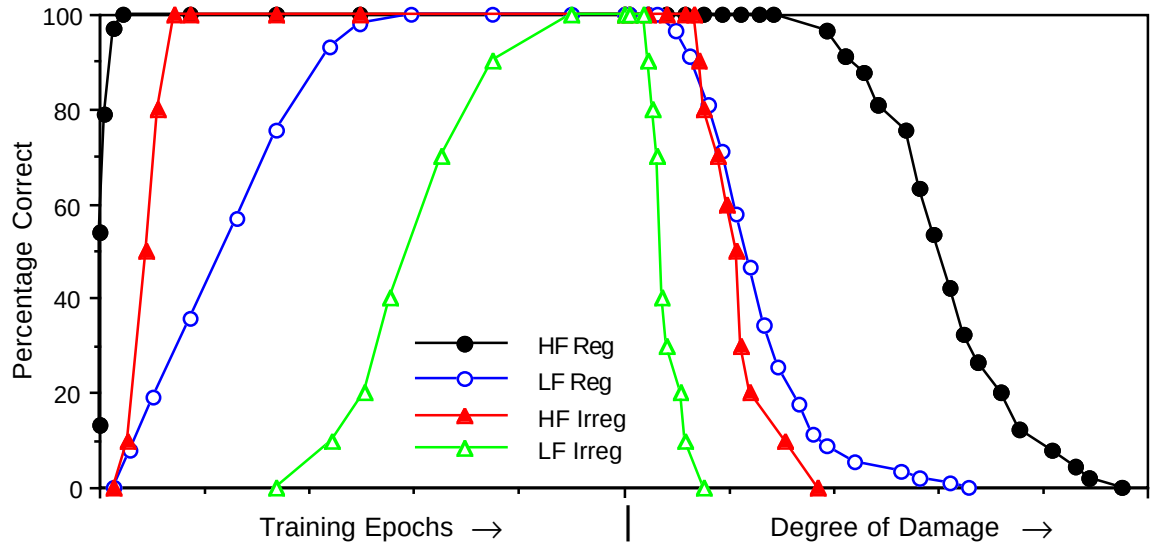

Figure 1: Frequency and regularity effects during training and lesioning of a simple neural network model (HF: High frequency; LF: Low frequency; Reg: Regular; Irreg: Irregular).

each output unit either drifting at random or falling to zero. Items that are learned first during training tend to end up furthest past the correct response thresholds when the training is stopped. Consequently they tend to be the last to cross over again and result in output errors during increasing degrees of damage. We see this in Figure 1: clear dissociations with the regulars more robust than frequency matched irregulars, and high frequency items more robust than regularity matched low frequency items. These basic effects extend easily to more realistic cases, such as surface dyslexia in the reading model of Bullinaria (1997) where not only are the relative error proportions for the various word categories simulated, but also the types of error produced.

There are clearly many factors, in addition to regularity and frequency, that can cause differing learning rates and corresponding deficits on damage. Consistency and Neighbourhood Density are commonly found in models of language tasks such as reading and spelling (e.g. Bullinaria, 1997). Representation Sparseness or Pattern Strength are often used to distinguish between concrete and abstract semantics, as in models of deep dyslexia (e.g. Plaut and Shallice, 1993). Correlation, Redundancy and Dimensionality have been used to distinguish the semantics of natural things versus artefacts, as in models of category specific semantic deficits (e.g. Devlin et al., 1998). These factors act in a similar manner to frequency and regularity, and their effects can easily be confounded. To make claims about neuro-psychological deficits involving one of them, we must be careful to control for the others.

We can see in Figure 1 the performance on high frequency irregulars crossing the low frequency regulars. With increasing damage there is first dissociation with better performance on the irregulars, and later the reversed dissociation. Devlin et al. (1998) find a similar pair of dissociations in their connectionist account of category specific semantic deficits. Such "double dissociations" are *resource artefacts* well known not to imply

underlying modularity (Shallice, 1988, p234), so there is no conflict with conventional neuropsychological inference. However, the finer grain of detail connectionist modelling affords here allows accounts of human deficits difficult to accommodate in older "box and arrow" models.

Modelling massively parallel brain processes by simulating neural networks on serial computers is only rendered feasible by abstracting the essential details and scaling down the size of the networks. The damage curves of Figure 1 are relatively smooth because the network has many more hidden units and connections than required to perform the task, and individual connections or hidden units make only small contributions to the network's outputs. For smaller networks, individual damage contributions can be large enough produce wildly fluctuating performance on individual items, and this can result in dissociations in arbitrary directions. Such small scale artefacts are often sufficient to produce convincing looking double dissociations (Shallice, 1988, p254). Bullinaria and Chater (1995) showed that as we scale up to more realistically sized networks, the processing becomes more distributed and these apparent double dissociations dissolve into single dissociations.

Many successful models of human performance and their associated neuropsychological deficits have been based on attractor networks rather than simple feed-forward networks, but still, the resilience to damage follows the story outlined above. Probably the most successful models of this type are the Plaut and Shallice (1993) models of *deep dyslexia*. Lesions at two different locations in their trained networks produce double dissociation between concrete and abstract word reading (Plaut, 1995). Although the two damage locations do not constitute modules in the conventional sense, it is not difficult to understand how they contribute to the processing of the two word types to different degrees, and give opposite dissociations when damaged. The robustness of each location in the network is fully consistent with the general discussion above.

REFERENCES

BULLINARIA JA. Modelling reading, spelling and past tense learning with artificial neural networks. *Brain and Language, 59*: 236-266, 1997.

BULLINARIA JA. Connectionist dissociations, confounding factors and modularity. In D Heinke, GW Humphreys and A Olsen (Eds), *Connectionist Models in Cognitive Neuroscience*. London: Springer, pp. 52-63, 1999.

BULLINARIA JA. Lesioned networks as models of neuropsychological deficits. In MA Arbib (Ed.) *The Handbook of Brain Theory and Neural Networks, Second edition*. Cambridge, MA: MIT Press, 2002.

BULLINARIA JA and CHATER N. Connectionist modelling: Implications for cognitive neuropsychology. *Language and Cognitive Processes, 10*: 227-264, 1995.

DEVLIN JT, GONNERMAN LM, ANDERSEN ES and SEIDENBERG MS. Category-specific semantic deficits in focal and widespread brain damage: A computational account. *Journal of Cognitive Neuroscience*, *10*: 77-94, 1998.

PLAUT DC. Double dissociation without modularity: Evidence from connectionist neuropsychology. *Journal of Clinical and Experimental Neuropsychology*, *17*: 291-321, 1995.

PLAUT DC and SHALLICE T. Deep dyslexia: A case study of connectionist neuropsychology. *Cognitive Neuropsychology*. *10*: 377-500, 1993.

SHALLICE T. *From Neuropsychology to Mental Structure*. Cambridge, UK: Cambridge University Press, 1988.

SMALL SL. Focal and diffuse lesions in cognitive models. In *Proceedings of the Thirteenth Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Erlbaum, pp. 85-90, 1991.