Submitted to *Cortex*, Forum on the Logic of Double Dissociation
August 2002

Interpreting Double Dissociations in Connectionist Networks

David C. Plaut
Departments of Psychology and Computer Science
Center for the Neural Basis of Cognition
Carnegie Mellon University

A common motivation for studying the cognitive impairments of brain-damaged patients is to determine the "functional architecture" of the cognitive system. But what constitutes a functional architecture? This question used to have a straightforward answer: a set of discrete components with communication pathways among them, with each component assigned a specific function or type of representation. With the additional assumption that brain damage can (and occasionally does) impact individual components or pathways while leaving the rest of the system intact, it becomes possible to use patterns of dissociations in the performance of behavioral tasks by brain-damaged patients to determine the identity and organization of the functional components of the cognitive system (Shallice, 1988).

But what if the cognitive system is not composed of discrete components? Do neuropsychological dissociations still inform cognitive theories, and if so, how? One important class of system to consider in this context is distributed connectionist networks. In such systems, different types of information are represented by distributed patterns of activity over different groups of neuron-like processing units. Mappings from one type of information to another (e.g., mapping a written word to its meaning and/or pronunciation) are accomplished by interactions across weighted connections, either directly or via additional groups of intermediate or "hidden" units that learn representations gradually in response to task demands. In general, these hidden representations come to reflect a blend of the similarities among the "visible" representations they mediate (Plaut & Gonnerman, 2000). Consequently, functions typically ascribed to individual components in modular theories are distributed across multiple groups of units—potentially the entire network—on a connectionist approach.

Even though the entire network may participate in processing each stimulus, different parts of the system typically make unique contributions or are differentially important for particular aspects of task performance. A variety of factors can contribute to this learned "functional specialization," including architectural biases on the sizes and patterns of connectivity within and among groups of units, as well as the statistical structure within the task information to be learned. As a result, damage to different parts of the system can result in relatively selective deficits in task performance, including double dissociations between two tasks or between two classes of stimuli within a single task.

As a case in point, Plaut and Shallice (1993; Plaut, 1995) demonstrated a double dissociation in reading aloud concrete versus abstract words (Warrington, 1981) following two types of damage to a network trained to pronounce written words via their meanings. In the simulation, concrete words were assigned far more semantic features than abstract words under the assumption that isolated concrete words evoke richer, more consistent semantic representations than abstract words (Jones, 1985; Schwanenflugel, 1991). This difference in statistical structure led the network to learn much stronger supportive interactions between semantic units and associated "clean-up" units for concrete words than for abstract words, so the latter had to rely largely on the direct pathway from orthography to semantics (see Figure 1a). As a result, random damage to subsets of the connections within this direct pathway produced a concrete word advantage (i.e., positive values for the measure plotted in Figure 1b). By contrast, because concrete words had learned to rely on the interactive support from the clean-up units, severe damage to connections in the clean-up circuit produced the opposite effect on average, with better performance on abstract than on concrete words. Thus, although the entire network participates in processing all types of item, learned functional specialization in subregions of the network led to relatively selective impairments following damage to those subregions.

The data plotted in Figure 1b were generated by testing the network after 1000 instances of each lesion type in which, for each instance, a specified percent of a particular set of connections were selected at random and removed. As would be expected given the random sampling involved, the resulting performance measures show a certain degree of variability across lesion instances. Indeed, the variability is sufficiently large to give rise to statistically reliable double dissociations *within* each condition—that is, between pairs of individual lesions at the same location and severity (Plaut, 1995). The occurrence of double dissociations among quantitatively equivalent lesions is potentially problematic for their use in informing cognitive theories insofar as, under such conditions, the dissociations are essentially chance occurrences that provide no information about the functional organization of the system. Indeed, based on analogous results in the domain of English inflectional morphology, Juola and Plunkett (2000) have argued (in a paper entitled *Why Double Dissoci-*

Financial support was provided by NIH Grant MH55628. Correspondence be sent to Dr. David Plaut, Department of Psychology, Carnegie Mellon University, Pittsburgh PA 15213–3890; email: plaut@cmu.edu.

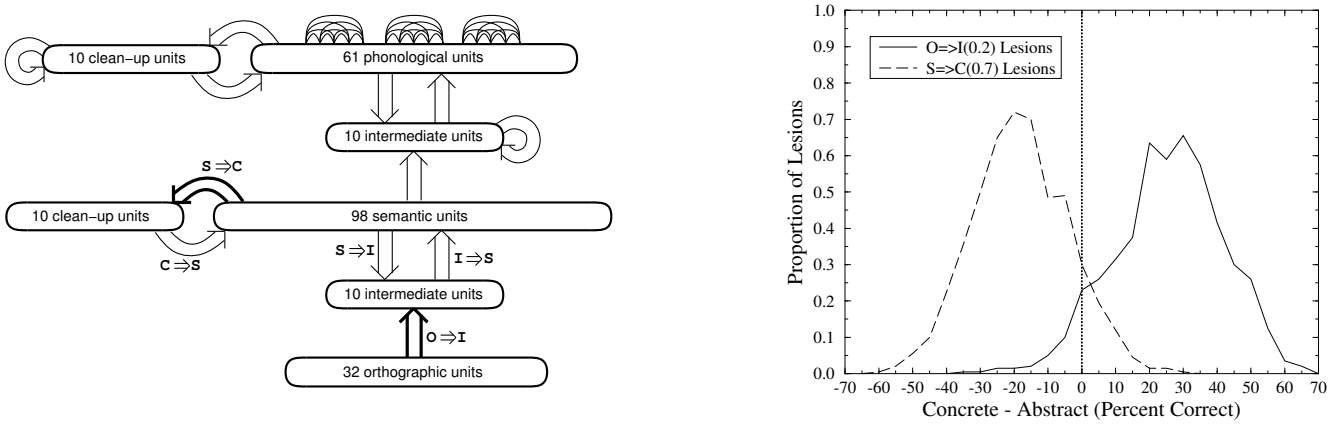

Figure 1. (a) The architecture of the Plaut (1995) network that mapped orthographic input to phonological output via semantics; and (b) distributions of differences in percent correct on concrete versus abstract words after lesions of 20% of orthographic-to-intermediate (O⇒I) connections and lesions of 70% of semantic-to-cleanup (S⇒C) connections—these sets of connections are shown in bold in (a). Adapted from Plaut (1995) with permission.

ations Don't Mean Much) that this same concern applies to the interpretation of single-case studies in human neuropsychology.

However, caution is warranted in drawing conclusions based on the *variance* of effects caused by random lesions in connectionist networks. The reason is that even the largest connectionist simulation is vastly smaller in scale then the brain systems it approximates. Each lesion provides a noisy estimate of the mean effect of quantitatively equivalent lesions. According to the Central Limit Theorem, the variance among these estimates decreases as a function of the number of samples entering into each estimate. In the case of lesions to a network, this sampling corresponds to the set of probabilistic choices of whether or not to remove individual units or connections. If it is assumed that actual brain damage is random at the scale of individual neurons, then a given brain lesion involves "sampling" over orders of magnitude more variables and, hence, the expected variance among the effects produced by quantitatively equivalent lesions is highly reduced. Put simply, sampling over hundreds or thousands of things (e.g., units/connections) is far more likely to yield idiosyncratic effects than sampling over hundreds of millions or billions of things (e.g., neurons). Thus, without some evidence that the granularity of sampling applied in lesions to connectionist networks is matched to the granularity of sampling inherent in brain damage, interpreting idiosyncratic effects of individual lesions to networks is likely to be misleading. On the other hand, effects of damage in connectionist networks that are based on the mean rather than the variance resulting from multiple lesions (as illustrated in Figure 1b) can be informative for identifying the nature of functional specialization in the system, even when this specialization does not correspond to the structure of the system in a "transparent" way (Caramazza, 1986), as modular theories typically assume (see also Van Orden, Jansen op de Haar, & Bosman, 1997).

References

Caramazza, A. (1986). On drawing inferences about the structure of normal cognitive systems from the analysis of patterns of impaired performance: The case for single-patient studies. *Brain and Cognition*, *5*, 41-66.

Jones, G. V. (1985). Deep dyslexia, imageability, and ease of predication. *Brain and Language*, *24*, 1-19.

Juola, P., & Plunkett, K. (2000). Why double dissociations don't mean much. In G. Cohen, R. A. Johnston, & K. Plunkett (Eds.), *Exploring cognition: Damaged brains and neural networks: Readings in cognitive neuropsychology and connectionist modelling* (p. 319-327). Hove, UK: Psychology Press.

Plaut, D. C. (1995). Double dissociation without modularity: Evidence from connectionist neuropsychology. *Journal of Clinical and Experimental Neuropsychology*, *17*(2), 291-321.

Plaut, D. C., & Gonnerman, L. M. (2000). Are non-semantic morphological effects incompatible with a distributed connectionist approach to lexical processing? *Language and Cognitive Processes*, *15*(4/5), 445-485.

Plaut, D. C., & Shallice, T. (1993). Deep dyslexia: A case study of connectionist neuropsychology. *Cognitive Neuropsychology*, *10*(5), 377-500.

Schwanenflugel, P. J. (1991). Why are abstract concepts hard to understand? In P. J. Schwanenflugel (Ed.), *The psychology of word meanings.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Shallice, T. (1988). *From neuropsychology to mental structure.* Cambridge: Cambridge University Press.

Van Orden, G. C., Jansen op de Haar, M. A., & Bosman, A. M. T. (1997). Complex dynamic systems also predict dissociations, but they do not reduce to autonomous components. *Cognitive Neuropsychology*, *14*, 131-165.

Warrington, E. K. (1981). Concrete word dyslexia. *British Journal of Psychology*, *72*, 175-196.