Why Double Dissociations Don't Mean Much

Patrick Juola (PATRICK.JUOLA@PSY.OX.AC.UK)[1]
Kim Plunkett (KIM.PLUNKETT@PSY.OX.AC.UK)
Department of Experimental Psychology
University of Oxford
Oxford, OX1 3UD UNITED KINGDOM

Abstract

The conventional interpretation of double dissociations is that they are almost irrefutable evidence of distinctions in both function and type of mental processes, or of separation of cognition into modules. We present a connectionist model that demonstrates apparent double dissociations within a single-route, single-mechanism network and argue that these apparent dissociations are simply the expected tails of a standard bell curve describing network performance. We conclude that within a connectionist model, the appearance of double dissociations may not be evidence for functional or mechanistic separation, and that similar caveats apply to the interpretation of double dissociations in human cognitive behaviour.

Introduction

Dissociations, and specifically double dissociations, are widely considered to be one of the more powerful tools in a cognitive neuropsychologist's arsenal. This phenomenon occurs when "[cognitive psychologists] can find one patient who can perform task A but not task B and a second patient who can perform task B but not task A." (McLeod, Plunkett, & Rolls, 1998, p.254) The philosophical position is simple, but far-ranging. By examining behaviour, and specifically how behaviour breaks down, the goal is to fractionate the components of cognition into their logical and behavioural constituents. If two styles of processing are logically and behaviourally separable (for example, the recognition of people by their faces and the person-independent recognition of facial expressions), one is tempted to conclude that the two processes are independent and don't rely on one another. Furthermore, if the two processes appear to do radically different things, or to do things in radically different ways, one can argue for a complete processing difference. This can further be argued to be not just a difference between separate systems doing the same sort of processing, but between separate systems doing completely different kinds of processing, as the needs of the separate constituents demand — for example, the basic features relevant to facial recognition are different from the basic features for expression recognition, and so there's reason to think that the two processes operate in completely different ways.

Inflectional morphology in English has been a fruitful area for this sort of study and analysis. It is relatively easy to fractionate, for example, the process of inflection into two separate processes of regular and irregular inflection. Even within a single syntactic category, the inflection of regular forms (e.g. *house* → *houses*) seems intuitively different from the inflection of irregular forms (*mouse* → *mice*). Regular inflections appear to generalize easily to novel or nonword forms, to unusual productions, and even to word forms that are usually irregular but are used in unusual or atypical ways. In contrast, irregulars seem to generalize significantly less when they generalize at all (few people are tempted to produce **hice* as the plural of *house*).

Daugherty and Seidenberg (1992) have demonstrated that irregular forms are much more sensitive to frequency effects in reading aloud than are regular forms. Prasada and Pinker (1993) have shown differences in similarity effects between regular and irregular forms in terms of generalization performance levels. The crowning piece of evidence for some sort of separation, or against any theory of a unified mechanism, should be the production of case studies showing a double dissociation — for example, studies of Williams syndrome patients (Bellugi, Hoeck, Lillo-Martin, & Sabo, 1988) seem to show that regular forms are relatively preserved (in comparison with performance on irregular forms), while studies of SLI patients (Gopnik & Crago, 1994) show that *irregular* forms are relatively preserved. Finally, Marslen-Wilson and Tyler (1997) present a case of two aphasics with different lexical decision performances on regular and irregular words, and claim, specifically, that "this is evidence for functional and neurological distinctions in the types of mental computation that support these different aspects of linguistic and cognitive performance." This is the classic form and conclusion of a double dissociation.

This argument, we claim, is incorrect. Plaut and Shallice (1994) have produced a connectionist network that does show functional separation between various sets of units, and thus damage to one particular set of units (or interconnecting weights) produces predictably different error patterns than damage to other sets. Similar effects could be expected to be found in, for instance, the (Mareschal, Plunkett, & Harris, 1995) model of object permanence or the (Miikkulainen, 1993) model of story understanding. In general, any of the generalized

[1]Current address: Department of Mathematics and Computer Science, Duquesne University, Pittsburgh, PA 15282 (JUOLA@MATHCS.DUQ.EDU)

pipe-fitting complex connectionist models show as much functional separation between their units and connections as the box-and-arrow diagrams that underly them and are often used to explain their functioning. Despite this degree of functional separation, however, this sort of structure does not provide evidence for "distinctions in the *types* of ... computation," as the type of processing is the same for all units within this type of network.

We present here a model and associated connectionist simulation that may explain some forms of double-dissociation as simple variance from a stochastic norm within a single-system, single-mechanism, associator. In this model, the effects of damage are unpredictable, and further, these effects may differentially affect different words or word categories. Our experiments show, for example, that a single network can be damaged randomly in such a way as to have very good performance on a particular class of words, or very bad performance, despite the level of damage being the same in either case. We argue that the mere observance of a double dissociation, particularly as a rare or pathological case, is not sufficient evidence to conclude a separation of processing, especially in cases where the damage itself can only be observed crudely and the function lost is highly complex.

Simulation

Network definition

The network that we chose to damage is a standard connectionist simulation, constructed as a multi-layer perceptron network using backpropagation of error (Rumelhart, Hinton, & Williams, 1986). The simulation was built using the PlaNet simulator (Miyata, 1991) using 130 units for the input layer, 160 units for the output layer, and 200 units as the hidden layer.

Five random sets of starting weights were trained over a gradually expanding corpus of training data eventually encompassing 2280 noun types and 946 verb types of varying frequencies representing their token frequencies as found in the Brown corpus (Kučera & Francis, 1967). The training data for the simulations were taken from the CELEX corpus (Baayan, Piepenbrock, & Rijn, 1993); we extracted from this database all words which were monosyllabic, which contained no "foreign" sounds in their pronunciation (according to the Moby Pronunciator database (Ward, 1997)), and for which we had evidence that they could be used as nouns or verbs. This yielded a total corpus of 2626 stems, which encompassed 3226 total inflected types (2280 nouns and 946 verbs). Of these types, 26 were irregular nouns and 122 were irregular verbs. For these words, we took the corresponding token frequencies (of the stems) from the Brown corpus (Kučera & Francis, 1967) as a rough measure of token frequencies in running speech. The token frequencies of words were individually tabulated as nouns and verbs, then the function $\log_2(freq^2 + 1)$ applied to these frequencies to flatten them into something more presentable to the network. The final variance was between 1 and 21 tokens/inflected type, meaning that the most frequent words appeared just over twenty times as often as the least. These token frequencies were also heavily dominated by nouns. Of the 17129 tokens in the training set 13045 were noun tokens (204 of them irregular) and 4084 were verb tokens (997 of them irregular).

The training corpus was prepared by converting the Moby symbolic pronunciation (Ward, 1997) into a large binary vector using a modification of the PGPfone alphabet representation (Juola & Zimmermann, 1996). Each phoneme was represented as a cluster of 16 binary phonetic features including aspects such as place, manner, and height of articulation. Each word was divided into onset–nucleus–coda constituents and right-justified within a CCCVVCCC template (e.g. the word "cat" (/kAt/) would be represented by the training pattern ##k#A##t, where '#' represents an absent sound). To this 128-bit pattern, two additional bits were appended representing the syntactic form to be inflected into, either the past tense (of a verb) or the plural (of a noun). The desired outputs were a similar encoding of the phonology of the inflected form, including an optional epenthetic vowel and final consonant. An incremental training regime was applied, where training started out with a small number (20 types) of high frequency words. The training set was then gradually expanded (5% type expansion per epoch) to include words of decreasing frequency until the entire corpus is absorbed. This training schedule is intended to capture the distinction between input to and uptake by the child (Plunkett & Marchman, 1993).

Each of the five starting points yielded a unique weight configuration after 115 training increments and was used as the basis for the lesioning experiments. Increment 115 is the earliest point at which the network had been exposed to the entire training corpus and, as might be expected, is the point with the worst overall performance on the training corpus. Because of the high error rate under "normal" circumstances, it is reasonable to assume that it would be the most sensitive to damage and therefore an appropriate time to lesion in search of interesting error patterns. The exact details of the acquisition and loss profiles are reported elsewhere in this volume, but can be briefly summarized by the results that nouns are, in general, superior to verbs and regulars superior to irregulars, in keeping with their relative frequencies within the corpus. We focus here not on the average level of loss, but instead on the variation in loss.

Analysis

Each (lesioned) subject was presented with the training corpus and the outputs interpreted by taking the closest phoneme string to the output units' activation pattern. This pattern was simply evaluated as "correct" or "incorrect," and the number of correct types of each category (e.g. regular nouns, irregular nouns, regular verbs, and irregular verbs) was taken as a measure of network performance.

We first note that the level of performance is "random," in at least the limited sense of not entirely predictable. Because of this unpredictability, we therefore can "expect" unexpected behaviour, both unexpectedly good and unexpectedly bad.

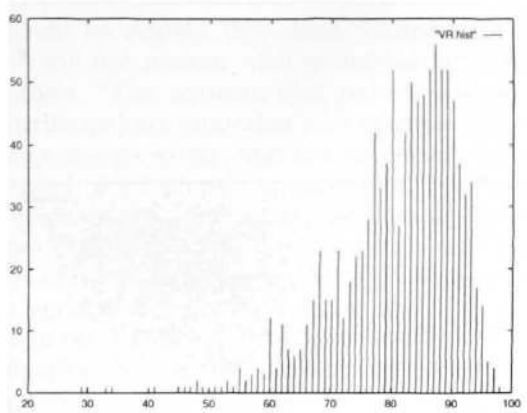

Figure 1: Histogram of regular verb performance (percentage of "normal")

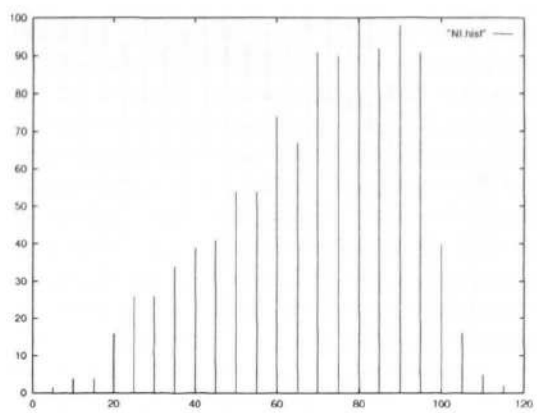

Figure 2: Histogram of irregular noun performance (percentage of "normal")

To confirm this, we performed 1065 separate lesions, all at the 97% level (in other words, leaving a random 97% of the connections intact) of one of the networks developed in the prior experiment. In a medical context, this might represent an exploration of the range of damage that could be expected after a particular patient received an injury of some particular severity.

Some degree of normalization is necessary, as the training set itself only included 26 irregular noun types but 2254 regular noun types — and the undamaged networks correctly inflected a higher percentage of regular noun tokens than irregular ones. (The network, for example, achieved 99% on regular nouns, 97% on regular verbs, but only 90% on irregular verbs and only 77% [20 out of 26] on irregular nouns. This is typical for all networks we studied.) We therefore normalized performance by calculating it as a percentage of "baseline" performance of the undamaged network used as a base for each subject. Because in many cases, especially for irregular nouns and verbs, the baseline performance included errors, it occasionally happened that the performance of an individual subject on an individual category would exceed the baseline, resulting in apparently paradoxical performance levels that exceed 100%. In other words, under certain circumstances, damage not only does not degrade performance, but will actually increase it.

Results

Figure 1 presents a histogram of the performance level on the inflection of regular verb types. Even after variation in patient and physical severity of damage have been controlled for, the outlines of a skewed bell curve can be seen. In other words, depending upon the exact nature of the lesion, an individual network/patient may display "little" impairment or "severe" impairment, relative to the expected performance level, although most networks will display "moderate" or "average" impairments. A similar curve, although with a different mean and median level of impairment, can be seen in figure 2, showing the same lesions' performance on irregular noun types.

Furthermore, these performance levels, although correlated, are at least partially independent, as can be seen

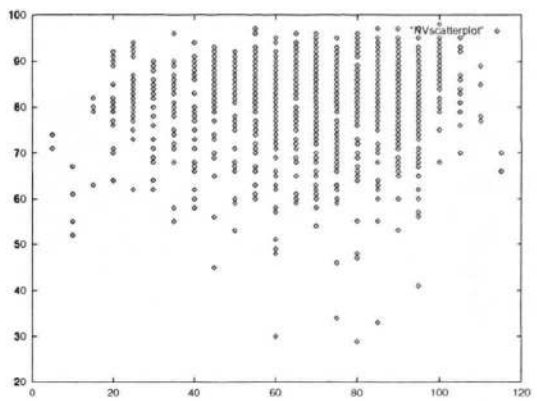

Figure 3: Scatter plot of irregular noun vs. regular verb performances (%ages).

on the scatter-plot (Figure 3). Obviously, the lesions at the upper left corner of the graph will display an apparent dissociation of performance on irregular nouns relative to regular verbs, and the converse holds for the lesions in the lower right. However, *these represent identically-severe lesions for the same network!* These lesions demonstrate a "double dissociation" within the same, unified-route, unified-mechanism processor.

These dissociations can be seen more clearly in the following set of figures (figures 4–8). In all these figures, points at the upper left of the figure are doubly dissociated with points at the lower right. Such pairs are easy to find in all figures except for figure 7, which appears to be relatively empty at the extreme lower right corner. This apparent emptiness, however, is only an emptiness at the extreme; points still exist where the performance on irregular verbs is "significantly" better than on regular verbs.

Discussion

This argument, then, may explain both the occurrence as well as the rarity of some neurological impairments. If one considers the case of a neurologist sitting in an emergency ward and examining patients as they come

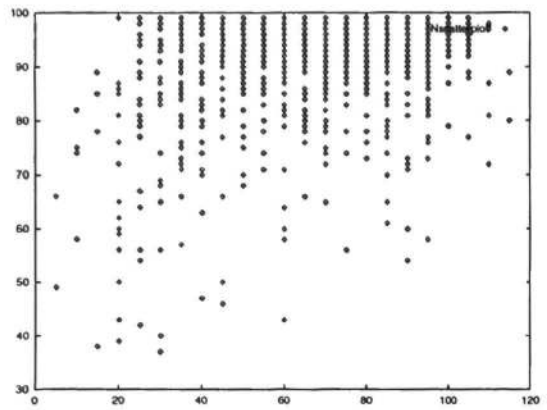

Figure 4: Scatter plot of irregular vs. regular noun performance

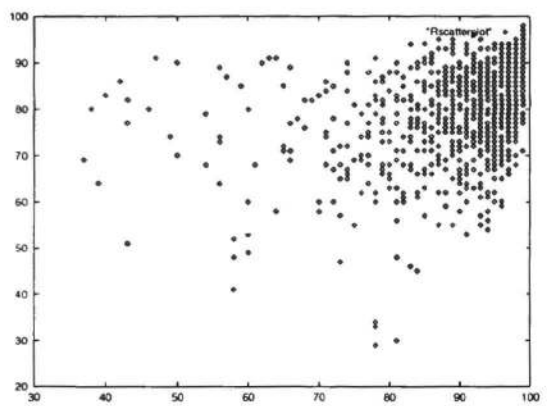

Figure 5: Scatter plot of regular noun vs. regular verb performance

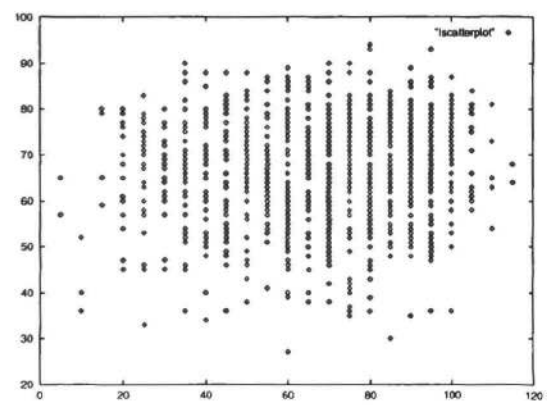

Figure 6: Scatter plot of irregular noun vs. irregular verb performance

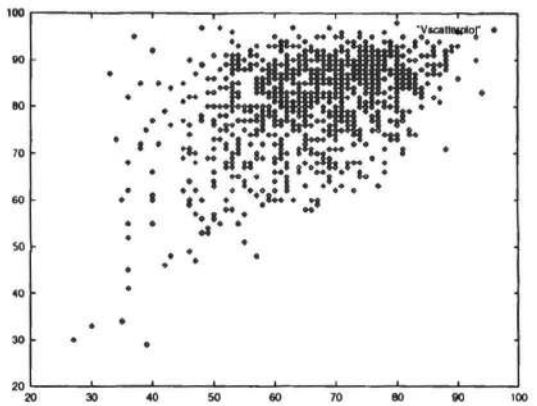

Figure 7: Scatter plot of irregular vs. regular verb performance

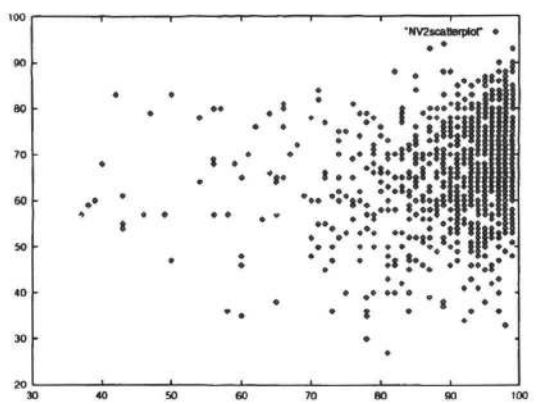

Figure 8: Scatter plot of regular noun vs. irregular verb performance

in, it should be evident that most "closed head injury patients" will not present with an interesting collection of symptoms. The patients that receive interest (and journal writeups) are somewhat rare examples that show how things *can* go wrong, and not necessarily how they are expected to go wrong. In particular, the change in output behaviour of a sufficiently complex system might not appropriately be described as just a performance "loss". A better description would be a change or alteration in performance; as we describe above, sometimes the change can improve (some aspects of) performance by eliminating factors that have contributed to errors. There appears to be no hard and fast boundary between routes, or modules, such that the module fails to function — instead, the outputs are subtly altered (by the insertion of unpredictable "noise") and the noisy outputs themselves may be subject to further noisy processing.

In particular, this model demonstrates that double dissociations may be possible, even in the absence of specific modular differences in type and function of processing. These double dissociations are instead the result of stochastic processes. The appearance of such a dissociation may not be sufficient evidence to conclude that such a separation exists. By extension, this sort of evidence *in humans* may not be sufficient evidence to conclude that an equivalent functional or neurological separation exists.

Conclusions

We have presented a model and explanation for some kinds of double dissociation that does not require a different method of processing or even a functional separation between modules in the underlying processing. We argue instead that, because the effects of damage to as complex a system as inflectional morphology are somewhat unpredictable, in some cases "random" damage will result in surprisingly good performance on some aspects of a task and surprisingly bad performance on other aspects, merely as a result of the task complexity exceeding our understanding of the system underlying it. Specifically, because we are unable to understand the exact differences between the representation of one type and another in connectionist networks, the differences in representation may occasionally conspire (under damage) to produce variance among some representational groups, whether these groups are "irregular nouns" or "words with even parity." These apparent conspiracies will (stochastically) produce double dissociations at the extreme tails of the probability distribution, irrespective of the functional and computational makeup of the system.

This argument might be extended as a suggestion against the extensive use of individual case studies in the general psychological literature; if connectionist networks are regarded as complex, how much more so is the human brain? It would be unusual indeed if we could predict the exact behavioural result from a particular injury or genetic makeup, meaning that in some cases, the results will be surprisingly devastating, while in other cases (or other tasks) the results will be surprisingly preserved. By sorting through enough patients or networks, one can "expect" that one's expectations will sometimes be woefully misleading. So the appearance of individual cases, absent some analysis of how characteristic or uncharacteristic they are, may not be significant in the fractionating of cognition. Paradoxically enough, then, the double dissociations produced by neural networks *may* actually be better evidence for how mental processes might break down, because the network developers can lesion the same network over and over until the results can be described, not in terms of idiosyncratic cases, but in terms of means and expectations.

Acknowledgements

This work was supported by a research project grant from the ESRC to Kim Plunkett.

References

Baayan, H., Piepenbrock, R., & Rijn, H. van. (1993). *The CELEX lexical database (CD-ROM).* University of Pennsylvania, Philadelphia: Linguistic Data Consortium.

Bellugi, U., Hoeck, K. van, Lillo-Martin, D., & Sabo, H. (1988). Dissociation between language and cognitive function in Williams syndrome. In D. Bishop & K. Mogford (Eds.), *Language develoment in exceptional circumstances.* Edinburgh: Churchil Livingstone.

Daugherty, K., & Seidenberg, M. S. (1992). Rules or connections? the past tense revisited. In *Proceedings of the fourteenth annual conference of the Cognitive Science Society.* Hillsdale, NJ: Erlbaum.

Gopnik, M., & Crago, M. (1994). Familial aggregation of a developmental language disorder. *Cognition, 39,* 1-50.

Juola, P., & Zimmermann, P. (1996). Whole-word phonetic distances and the PGPfone alphabet. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP-96).* Philadelphia, PA.

Kučera, H., & Francis, W. N. (1967). *Computational analysis of present-day American English.* Providence, RI: Brown University Press.

Mareschal, D., Plunkett, K., & Harris, P. (1995). Developing object permanence : A connectionist model. In J. D. Moore & J. E. Lehmann (Eds.), *Proceedings of the seventeenth annual conference of the Cognitive Science Society* (pp. 170–5). Hillsdale, NJ: Erlbaum.

Marslen-Wilson, W. D., & Tyler, L. K. (1997). Dissociating types of mental computation. *Nature, 387,* 582–4.

McLeod, P., Plunkett, K., & Rolls, E. T. (1998). *Introduction to connectionist modelling of cognitive processes.* Oxford, UK: Oxford University Press.

Miikkulainen, R. (1993). *Subsymbolic natural language processing: An integrated model of scripts, lexicon, and memory.* Cambridge, MA: MIT Press.

Miyata, Y. (1991). *A user's guide to PlaNet version 5.6 : A tool for constructing, running, and looking into a PDP network.*

Plaut, D., & Shallice, T. (1994). *Connectionist modelling in cognitive neuropsychology : A case study.* Hove, UK and Hillsdale, US: Lawrence Erlbaum.

Plunkett, K., & Marchman, V. (1993). From rote learning to system building: Acquiring verb morphology in children and connectionist nets. *Cognition, 48,* 21–69.

Prasada, S., & Pinker, S. (1993). Generalizations of regular and irregular morphology. *Language and Cognitive Processes*(19), 207–96.

Rumelhart, D., Hinton, G., & Williams, R. (1986). Parallel distributed processing: Explorations in the microstructure of cognition. In (Vol. 2: Psychological and Biological Models, pp. 318–362). The MIT Press.

Ward, G. (1997). *Moby pronunciator.* 3449 Martha Ct., Arcata, CA, USA. (Also available at *http://www.dcs.shef.ac.uk/research/ilash/Moby/index.html*)