

Introduction to special issue on computational modelling in cognitive neuropsychology

Gary S. Dell

University of Illinois, Urbana-Champaign, IL, USA

Alfonso Caramazza

Harvard University, Cambridge, MA, USA, and Center for Mind/Brain Sciences-CIMEC, University of Trento, Rovereto (TN), Italy

Recently, *Cognitive Neuropsychology* altered the statement of the journal's aims that appears in every issue, the new material expressing an interest in publishing "computational modelling research that is informed by consideration of neuropsychological phenomena". The journal's interest in computational models, however, is anything but recent. When *Cognitive Neuropsychology* was in its first decade, it devoted an entire issue to a single article reporting a model of impaired reading by Plaut and Shallice (1993). This now classic treatment of dyslexic errors from a connectionist perspective—it is currently the third most cited article in the journal—and later examples of what Coltheart (2006) calls "computational cognitive neuropsychology" have influenced the interpretation of neuropsychological data for some time now. This special issue thus comes at a time when a variety of computational models and modelling frameworks have had their say on neuropsychological data.

What is a computational model? It is a model expressed as a computational implementation, the implementation being necessary to understand the model's implications for data and for theory. What, then, is a model? The term "model" has been used in the context of accounts of impaired cognition since the inception of the field. For example, the Wernicke–Lichtheim model of aphasia (e.g., Lichtheim, 1885) was a diagram containing nodes representing mental content

(e.g., auditory lexical images) and directional arrows indicating the flow of processing among the nodes during various tasks (e.g., speech comprehension). Lesions could be associated with nodes or arrows, and, because each of these components was identified with both a cognitive function and a brain region, the model generated accounts of aphasic symptoms—how they clustered and how they were associated with brain areas. For example, the poor comprehension and repetition associated with Wernicke's aphasia was attributed to a lesion to the auditory-image node. The diagram revealed that this node is a necessary part of the path for the tasks of both comprehension and repetition. Thus, the model, in the form of Lichtheim's diagram, made concrete the links between the theory and potential patterns of data.

Modern computational models are, of course, much more than diagrams. But they have much the same function. The model's computer program specifies a cognitive architecture, including processing levels or subsystems, and a variety of operations and parameters. This program is analogous to the diagram. Running the program simulates the cognitive processes involved in the performance of some task, just as following the arrows does in a diagram. Of course, the program is necessary because the complexity of most computational models precludes working things out by hand. Various aspects of the program can be

altered, simulating lesions, and the consequences of the lesions on task performance can be determined when the altered program is run. Finally, the lesioned model's performance can then be compared to that of impaired individuals, thus providing a test of the theoretical principles behind the model. Five of the six articles in this issue test specific models of some cognitive domain by using exactly this method, and the sixth, Goldrick's analysis of how errors reflect representational structure in models (Goldrick, 2008), applies this logic at a more general level.

The first explicitly computational models of cognition treated the mind as a digital computer program (e.g., Newell & Simon, 1972). The models' processing operations were discrete and were carried out in strict serial order. There was a principled separation between representations, which were ordered strings of symbols, and the processing rules that transformed these strings into other representations. The major alternative to this digital-computer framework is the *connectionist* approach, in which processing is carried out by a network of simple processing units that operate in parallel by sending continuous quantities (activation) to one another through weighted connections. All of the models presented here are connectionist or have connectionist properties. In fact, the words "computational" and "connectionist" are often, but incorrectly, treated as synonyms because connectionism pervades modern cognitive theories to such an extent that the major models based on these theories all employ at least some connectionist features. For example, although the dual-route cascade (DRC) model of reading presented in the article by Nickels, Biedermann, Coltheart, Saunders, and Tree (2008) was derived from the classic dual-route model that has often been contrasted with connectionist reading models, it uses spreading activation through weighted connections to do its computations. Other canonical connectionist notions are fully on display in other papers: learning as adaptive weight change (Cutini, Di Ferdinando, Basso, Bisiacchi, & Zorzi, 2008; Dilkina, McClelland, & Plaut, 2008; Goldberg & Rapp, 2008), nonlinear activation functions (e.g.,

Cutini et al., 2008; Dilkina et al., 2008; Goldberg & Rapp, 2008; Nickels et al., 2008), and an interactive or recurrent flow of activation (Dilkina et al., 2008; Goldberg & Rapp, 2008; Knobel, Finkbeiner, & Caramazza, 2008). An important distinction within the connectionist framework is whether the model's representations (how patterns of activation correspond to particular cognitive elements) are *localist* (e.g., there is an abstract network unit for the lexical item CAT) or *distributed* (e.g., the lexical representation of CAT is a pattern of activation across many units). Both localist (e.g., Knobel et al., 2008; Nickels et al., 2008) and distributed (e.g., Dilkina et al., 2008; Goldberg & Rapp, 2008) models are explored here, and the function of these representational formats for cognitive neuropsychology is revealed (see, particularly, Goldrick's, 2008, paper).

The common connectionist ancestry of the papers in this issue, though, should not obscure the striking differences among them in how the models were used. In three of the papers, the model was the theoretical centrepiece, and the neuropsychological data that were examined supported, for the most part, the theories that the models implemented. In the remaining three papers, however, the models served other diverse theoretical purposes. Let us begin with the models that were supported by the data. Dilkina et al.'s (2008) analysis of lexical and semantic deficits from a single-system perspective, Cutini et al.'s (2008) study of performance on the travelling salesperson problem, and Nickels et al.'s (2008) application of the DRC model to phonological dyslexia all included successful simulations of the performance of brain-injured subjects (and also normal individuals experiencing transcranial magnetic stimulation in Cutini et al.'s study). That is, in these cases, the model worked; it fitted the data to the satisfaction of the authors.

But how can we tell whether or not a model works? There has been considerable discussion among those who apply models to patient data about how model evaluation should take place (see Rumel, Caramazza, Capasso, & Miceli, 2005; and Schwartz, Dell, Martin, Gahl, & Sobel,

2006, for recent commentary), and the first group of papers highlight a number of issues on this topic. First, should the model be matched to individual cases or to mean data from a group of cases that are assumed to be similar? When models were initially applied to impaired cognition, simulated lesions were informally compared both to data from individual cases (e.g., Patterson, Seidenberg, & McClelland, 1989) and to data from patient-group means (e.g., Haarmann & Kolk, 1991). Currently, computational cognitive neuropsychology, as illustrated by most of the papers here (e.g., Dilkina et al., 2008; Nickels et al., 2008), emphasizes matching to individual cases. There are several reasons for this choice, but a compelling one is simply that it is possible—even likely, if care is not taken in defining the group—that in a group of brain-damaged individuals mean performance may not be typical of many or even any of the individuals. Thus, one may be modelling a performance pattern that was not the product of a single brain. Second, the papers also exhibit a time-honoured feature of model-based hypothesis testing: the explicit comparison among competing models or model versions. For example, Cutini et al. (2008) implemented two different accounts of flexibility in the application of heuristics for the travelling salesperson problem, and the comparison between these and the data allowed for a valuable conclusion about the incremental nature of the planning process. Dilkina et al. and Nickels et al. explored the effect of variation in parameters and possible lesions in their models, effectively comparing different model versions as potential accounts of the data. A third issue in model evaluation concerns the role of inferential statistics when comparing models and data. Nickels et al. explicitly recommended (and did) such testing to determine whether one can reject the null hypothesis that the human data and the simulation's output are the same. Failure to reject the hypothesis of difference does indeed provide evidence of a good correspondence between the model and the data, provided that the data are extensive and precise. The alternative perspective argues that such null hypotheses are never in fact true and

that model evaluation should focus on the *degree* of fit (How full is this glass?), rather than on whether there are statistically significant deviations (Is this glass not full?). If we grant, however, that our simulations will never be perfect matches to our neuropsychological data, how are we then to decide that a simulation study has successfully supported a particular model (Ruml & Caramazza, 2000)? There is no simple answer to this question. One evaluation method that has been increasingly used in computational cognitive neuropsychology focuses on predictions—specifying or parameterizing a model based on data from some conditions or tasks and then using that model to predict performance in other conditions or tasks (e.g., Dell, Martin, & Schwartz, 2007). Dilkina et al.'s analysis of the relation between picture naming and word reading illustrates the method. First, for each patient a general extent-of-lesion parameter value was selected so that the model mimicked the patient's overall naming performance. Then, with that parameter set, the model's ability to predict the patient's reading performance for item groups varying in frequency and regularity was assessed.

The second group of papers (Goldberg & Rapp, 2008; Goldrick, 2008; Knobel et al., 2008) is a bit out of the ordinary in that the analysis offered did not explicitly support any particular model, but instead centred on a theoretical question whose answer required a computational treatment. Goldberg and Rapp started with the question of whether serially ordered behaviour—specifically in spelling—is carried out by compound chaining and answered the question in the negative. In a compound-chaining theory, the production of each item in a sequence is cued by representations of previously retrieved items. Goldberg and Rapp specifically evaluated a model of spelling that implemented compound chaining, a simple recurrent network or SRN (Elman, 1990). Not only does the SRN model clearly employ compound chaining, but SRNs are also a key component of large-scale psycholinguistic theories of production and comprehension (e.g., Chang, Dell, & Bock, 2006). Thus, their failure to account for the

spelling errors made by the two patients in the study is noteworthy.

Knobel et al. (2008) asked the question: Where is the effect of frequency in word retrieval during production? Experimental psycholinguistic studies that measure response times to retrieve picture names have led to conflicting results on the question of the locus of frequency effects, and so Knobel et al. took a different approach. They examined the production errors made by an individual with aphasia, E.C., and particularly how different *kinds* of errors were influenced by frequency. Then they used a model of word retrieval to systematically determine the implications of possible frequency loci on error patterns. For example, if frequency impacts, say, the mapping from semantic to lexical representations, what kinds of errors should and should not be affected by frequency? Knobel et al.'s comparisons between the simulated and actual influences of frequency in E.C. and other patients allowed them to conclude that frequency impacts both the interface between lexical and phonological representations and the processes that generate semantic errors. More generally, the results suggested that the effects of frequency are distributed throughout the lexical system, a conclusion that comports well with theories in which linguistic representations and the mappings among them are products of the incremental learning by which connectionist models acquire their networks (e.g., Dilkina et al.'s, 2008, analysis of how learning creates frequency and regularity effects).

Goldrick's (2008) paper stands apart from the others because no particular theoretical claim is tested. Instead, the analysis constitutes perhaps the first example of what can be called "metacomputational cognitive neuropsychology", a formal examination of the validity of the assumptions behind the use of models and other theoretical constructs in the field. The assumption in this case is that errors reflect the sharing of representational components: If cognitive item X's representation overlaps more (shares representational elements) with that of Y than it does with that of Z, then errors in which X is the target of a retrieval effort should more often

create Y than Z. So, if the representation of the phoneme /t/ overlaps with /d/ more than with /g/, then /t/s should be replaced by /d/s more than by /g/s. This assumption, which is entirely taken for granted in error-based studies of cognitive processing, was shown by Goldrick to be valid for most, but surprisingly not all, representational schemes.

Goldrick's (2008) analysis raises the thorny question of the ties between neuropsychological data, computational models, and the brain. The Wernicke-Lichtheim model was not just a model of what clusters of aphasic symptoms we should expect and what we should not expect. It also made claims about where damage in the brain should be found for each symptom cluster. These kinds of claims are, for the most part, de-emphasized in the papers in this issue. Just as "computational" does not mean "connectionist", "connectionist" does not mean "neural". The models describe cognitive rather than neural architectures, and they explain patient performance data through hypothesized lesions to the architectures. This is not to say that neural data are irrelevant to computational cognitive neuropsychology. For example, in Cutini et al.'s (2008) nested incremental modelling approach, properties of some model components were motivated by neural data (e.g., properties of simple cells in V1). Moreover, the top-down controller in their model was implicitly identified with frontal brain regions, because the lesion to this model component was matched to data obtained from individuals with frontal damage. Similarly, Dilkina et al. (2008) suggested that their patient differences may be explicable in terms of region-specific atrophy in temporal areas. Thus, their modelling study could be profitably augmented with assessments of lesion locations. In general, the theoretical claims that are advanced in the papers in this issue can be tested by a variety of data sources, behavioural or neural.

A final point: Models are simpler than the cognitive processes they represent. This truism clearly applies to the box-and-arrow models that are commonly used to illustrate cognitive architectures. But it also applies to the computational models

that are designed to open up the boxes and reveal the transformations behind the arrows. The models presented here are circumscribed in their domain (e.g., errors in spelling single words; performance on simple picture naming or reading tests), stingy in their processing assumptions (e.g., mathematically simple spreading activation rules), and limited in the number of parameters that can be varied. This simplicity is a virtue. The properties of simple models can be systematically explored, and such models can easily be compared to others that implement alternative theories. In this way, a limited, well-crafted model promotes the understanding of cognitive mechanisms and reveals the ramifications of these mechanisms for data.

REFERENCES

- Chang, F., Dell, G. S., & Bock, K. (2006). Becoming syntactic. *Psychological Review*, *113*, 234–272.
- Coltheart, M. (2006). Acquired dyslexias and the computational modeling of reading. *Cognitive Neuropsychology*, *23*, 96–109.
- Cutini, S., Di Ferdinando, A., Basso, D., Bisiacchi, S. P., & Zorzi, M. (2008). Visuospatial planning in the travelling salesperson problem: A connectionist account of normal and impaired performance. *Cognitive Neuropsychology*, *25*, 194–217.
- Dell, G. S., Martin, N., & Schwartz, M. F. (2007). A case-series test of the interactive two-step model of lexical access: Predicting word repetition from picture naming. *Journal of Memory and Language*, *56*, 490–520.
- Dilkina, K., McClelland, J. L., & Plaut, D. C. (2008). A single-system account of semantic and lexical deficits in five semantic dementia patients. *Cognitive Neuropsychology*, *25*, 136–164.
- Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, *14*, 179–211.
- Goldberg, A. M., & Rapp, B. (2008). Is compound chaining the serial order mechanism of spelling? A simple recurrent network investigation. *Cognitive Neuropsychology*, *25*, 218–255.
- Goldrick, M. (2008). Does like attract like? Exploring the relationship between errors and representational structure in connectionist networks. *Cognitive Neuropsychology*, *25*, 287–313.
- Haarmann, H. J., & Kolk, H. H. J. (1991). A computer model of the temporal course of agrammatic sentence understanding: The effects of variation in severity and sentence complexity. *Cognitive Science*, *15*, 49–87.
- Knobel, M., Finkbeiner, M., & Caramazza, A. (2008). The many places of frequency: Evidence for a novel locus of the lexical frequency effect in word production. *Cognitive Neuropsychology*, *25*, 256–286.
- Lichtheim, L. (1885). On aphasia. *Brain*, *7*, 433–484.
- Newell, A., & Simon, H. A. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice Hall.
- Nickels, L., Biedermann, B., Coltheart, M., Saunders, S., & Tree, J. J. (2008). Computational modelling of phonological dyslexia: How does the DRC model fare? *Cognitive Neuropsychology*, *25*, 165–193.
- Patterson, K., Seidenberg, M. S., & McClelland, J. L. (1989). Connections and disconnections: Acquired dyslexia in a computational model of reading. In R. G. M. Morris (Ed.), *Parallel distributed processing: Implications for psychology and neuroscience* (pp. 131–181). Oxford, UK: Oxford University Press.
- Plaut, D. C., & Shallice, T. (1993). Deep dyslexia: A case study of connectionist neuropsychology. *Cognitive Neuropsychology*, *10*, 377–500.
- Ruml, W., & Caramazza, A. (2000). An evaluation of a computational model of lexical access: Comment on Dell et al. (1997). *Psychological Review*, *107*, 609–634.
- Ruml, W., Caramazza, A., Capasso, R., & Miceli, G. (2005). Interactivity and continuity in normal and aphasic language production. *Cognitive Neuropsychology*, *22*, 131–168.
- Schwartz, M. F., Dell, G. S., Martin, N., Gahl, S., & Sobel, P. (2006). A case-series test of the interactive two-step model of lexical access: Evidence from picture naming. *Journal of Memory and Language*, *54*, 228–264.