THE INFLUENCE OF IMPLEMENTATION ON "HUB" MODELS OF SEMANTIC COGNITION

Olivia Guest, Richard P. Cooper, and Eddy J. Davelaar

Department of Psychological Sciences, Birkbeck, University of London
Malet Street, London, WC1E 7HX, United Kingdom

Rogers et al. (2004) present a model of semantic cognition – the "hub" model – that reproduces the behaviour of neurologically healthy and neurologically impaired individuals on a range of tests of semantic knowledge. The model and associated theory provide a comprehensive explanation for deficits, such as semantic dementia, by appealing to the breakdown of attractors within a recurrently connected system following damage. We report findings from an attempted replication of the Rogers et al. model. While normal behaviour was reproduced, lesioning the reimplementation did not fully replicate the behaviour of the original model, meaning that the reimplementations contain healthy semantic representations which are in line with the hub theory, but the effects of damage on the structure of the semantic representations are not theoretically accounted for. The hub theory predicts that after damage semantic representations must decay in certain ways in order to give rise to patient behaviour. Our results show that the reimplementations do not fully exhibit these symptoms. This suggests that while semantic impairments reminiscent of patients may arise following lesioning of the hub model, such patterns are not a necessary consequence of the model as initially described. We discuss the implication of this apparently negative result for the hub theory of semantic cognition, focusing on differences between our reimplementation and the implementation of Rogers et al., and on the theory-model relationship more generally.

1. Introduction

Semantic cognition comprises a set of cognitive processes that give percepts meaning, allowing for the formation of relations over both concepts and percepts. Various semantic tasks have been developed that test a subjects' ability to access and use semantic concepts, given percepts. These tasks are designed to be administered to both neurologically healthy individuals and patients with semantic deficits (e.g., semantic dementia, herpes simplex virus encephalitis, semantic aphasia, etc.) to determine the exact nature of patient deficits. Patient and healthy participant data may further be

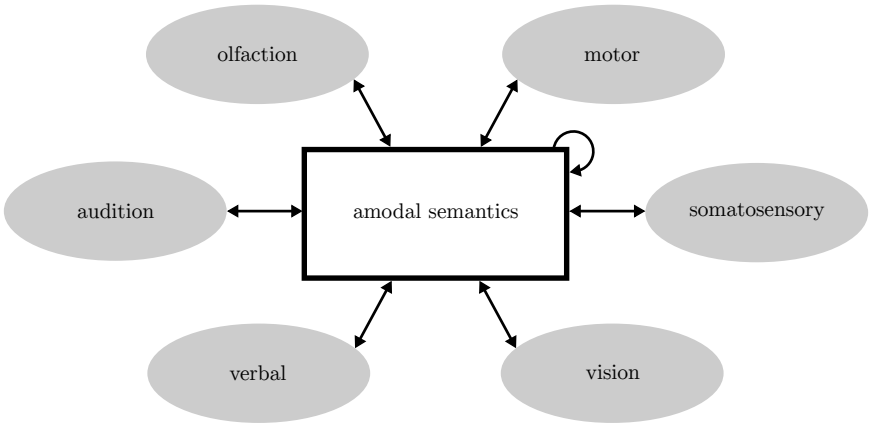

Figure 1: An overview of the semantic hub and its modal spokes, based on figure 2 in Lambon Ralph et al. (2007).

used to inform and benchmark cognitive modelling efforts, which aim to account for the ability to perform semantic cognition tasks and predict the breakdown of semantic memory.

One influential account of semantic cognition is the "hub" theory of Rogers et al. (2004). According to this theory, modality-specific perceptual inputs (e.g., visual, aural, motor, somatosensory, etc.) are reciprocally connected to a central amodal hub, as shown in figure 1. The information passed between the hub and its spokes allows for retrieval of semantic associations (e.g., visualising a dog based on hearing a bark), identification (e.g., calling a picture of a dog "dog"), categorisation (e.g., classifying a poodle as a "dog", "mammal" and "animal"), and generation (e.g., describing, drawing, or imitating a "dog"). Damage to the connectivity within the amodal semantic hub, and between the hub and the modal spokes, is proposed to give rise to the deficits seen in patients. The model thus aims to provide an explanation for both normal and impaired semantic cognition.

Rogers et al. (2004) present a connectionist implementation of the hub theory. Their model, shown in figure 2, consists of a subset of the possible modalities, allowing for visual, verbal and name input/output, and is specifically designed to account for the effects of neurodegeneration as seen in semantic dementia (SD) patients. When undamaged, the Rogers et al. model performs at ceiling on four tests of semantic cognition, as do neurologically healthy participants. However, after removing a random subset of all connections (by setting the corresponding weights to zero), the model shows the same qualitative patterns as SD patients. Thus, the model is

held to capture the complexity of semantic cognition required to explain both normals and SD patients.

The mechanism that underpins concepts – both in the hub theory and in the connectionist implementation – is the emergence of attractor states. Such states arise in dynamical systems that have recurrently connected components. Given partial input the system state gravitates towards the centre of a basin of attraction, thus recreating the full multi-modal experience of the concept. The hub theory proposes that, as a result of lesioning connections, neighbouring attractor basins coalesce, creating larger more generalised concepts. Attractors that are proximal in semantic space merge to represent a more general concept, as reflected in SD patients' responses on semantic tasks.

Based on three reimplementations, we explore the claim that the hub theory as described by Rogers et al. (2004) is sufficient to yield models with the required attractor dynamics. In the following sections, the effect of implementation differences on model performance on semantic tasks is discussed with the aim of illuminating the relationship between the hub theory and its implementation. We conclude by considering the general relationship between models and theories.

2. Three implementations of the hub theory

2.1. General architecture

The hub model, as presented in Rogers et al. (2004), is a real-valued recurrent neural network consisting of three pools of input/output units: 40 name units, 64 visual units, and 111 verbal units (further subdivided into 61 perceptual, 32 functional, and 18 encyclopaedic units). Name units represent natural language labels (e.g., "car"), visual units code for visual perceptual features (e.g., "is blue"), and verbal units assume the role of general verbal properties that are perceptual (e.g., "makes noise"), functional (e.g., "can cut"), and encyclopaedic (e.g., "is living"). These units are bidirectionally connected to 64 fully recurrent hidden units, as shown in figure 2. The input/output pools represent the sensory spokes, and the hidden units represent the amodal semantic hub. Activation spreads from one or more spokes to the hub and from the hub back to every input/output pool, thus functioning as a pattern-completing auto-associator.

2.2. Pattern set

Rogers et al. (2004) provide a probabilistic template for generating appropriate training sets. We use this template (cf. fig. 3, Rogers et al.) to create a pattern set, equivalent in structure to the original Rogers et

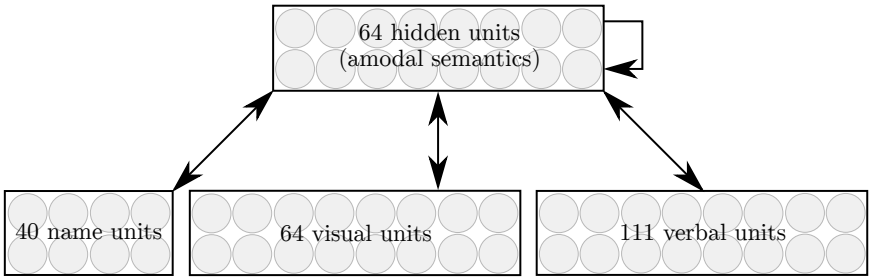

Figure 2: The hub model's neural network topology, based on figures 1 and 3 in Rogers et al. (2004).

al. (as evidenced by hierarchical cluster dendrograms), for training and testing our reimplementations. According to the template, mutually exclusive subsets of visual and verbal features underpin the main distinction between man-made and inanimate objects, as shown in figure 3. Other structural properties are: that the two domains are subdivided into 6 categories (mammals, birds, fruit, vehicles, household objects and tools); that verbal sub-patterns include a single feature present to denote category and domain membership; and that names consist of a single uniquely activated unit, thus creating 40 orthogonal name bit vectors. Some names are shared between certain visual/verbal sub-patterns in order to create category-level names, thus giving rise to archetypal patterns (e.g., labelling an animal "dog" as opposed its breed name).

The elements of the training set are binary vectors each with 215 bits. Each vector has the following bits set: *a*) the individual visual or verbal features it possesses (e.g., "is red", "has legs"); *b*) the localist orthogonal bit vector that constitutes the name sub-pattern (e.g., "robin"); and *c*) the localist category and domain membership units within the verbal sub-pattern (e.g., "is mammal", "is tool"; in figure 3 these are represented by the last 7 units). Based on this structure, we created a pattern set consisting of 48 items that abides by the above constraints. In other words, each pattern consists of a name, which contains no intrinsic information, followed by a set of visual and verbal properties, which contain shared and distinctive features that enable the network models to infer a similarity structure.

2.3. Training algorithms

We report three implementations of the hub theory using the architecture and pattern set described above.

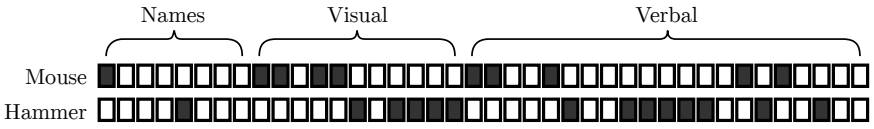

Figure 3: Two examples of simplified patterns. Solid rectangles represent activated features in the visual and verbal sub-patterns (e.g., "has fur"), while empty ones represent features that are not present.

2.3.1. $\mathbf{BPTT}_1$

The first network was trained using epochwise back propagation through time (BPTT: Williams & Zipser, 1989, 1995), following the procedure of Rogers et al. (2004) where specified. BPTT is a variant of back propagation that involves "unrolling" a multi-layered feedforward version of the recurrent network and training the weights using standard back propagation within this new unrolled network. When the learning phase is completed the network is reverted back to its normal recurrent state. Following Rogers et al., the network was settled for 28 steps during training. As in Rogers et al., the input units were clamped (i.e., forced to take on their target values) for twelve of these steps. We refer to this method of training as BPTT_1.

2.3.2. $\mathbf{BPTT}_2$

An alternative method, which we refer to as BPTT_2, is to clamp the targets to the outputs for the full 28 settling steps – every other aspect of this training procedure is identical to BPTT_1. This reduces the noise in the error signal during training resulting in an order of magnitude fewer epochs to learn the training set.

2.3.3. BM

A radically different way of implementing the model is to use a Boltzmann machine (BM). BMs are a type of binary-valued recurrent stochastic neural network. This kind of network is able to conform to the topology required by the hub theory and permits the emergence of attractors (Hinton & Sejnowski, 1986). Training involves minimising the difference in unit activations between the network settled with all inputs clamped, known as the plus state, and the network settled on each sub-pattern (e.g., just the verbal features clamped), called the minus states (Ackley, Hinton, & Sejnowski, 1985).

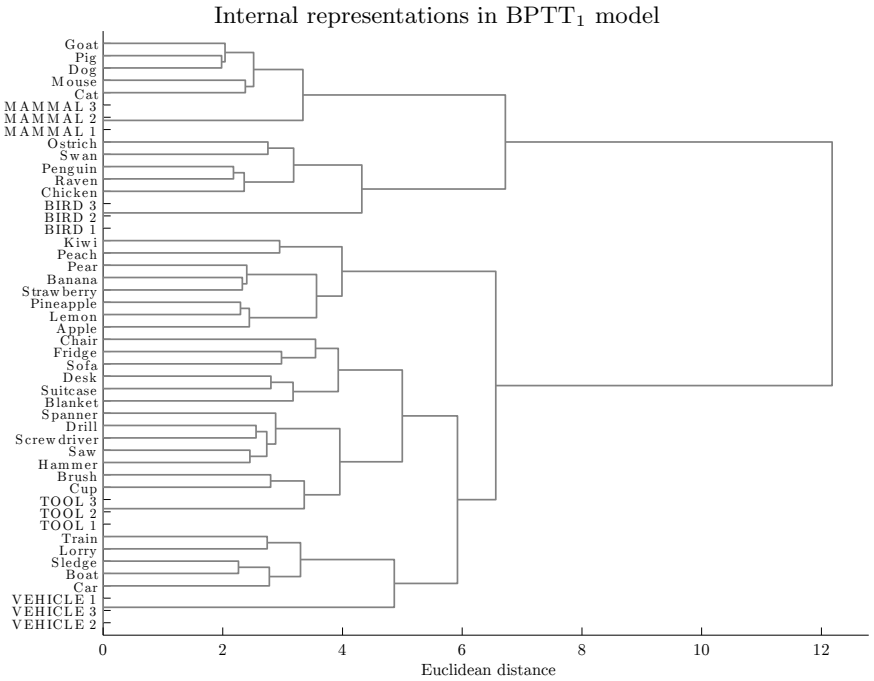

Figure 4: Dendrogram for BPTT$_1$ internal semantic states sampled 100 times per name sub-pattern input. Category names, shared over three patterns, are capitalised.

3. Simulation results

3.1. Normal behaviour

Semantic cognition can be assessed using: *a*) confrontation naming, which involves giving an appropriate linguistic label to a line-drawing; *b*) word-to-picture matching, pairing a card with a word on it to one with a drawing of the same animal or artefact; *c*) sorting words and pictures, categorising the aforementioned word and picture cards into two piles for each domain (animate/inanimate objects) and into six categories (mammals, birds, fruit, vehicles, household objects, and tools); and *d*) drawing, copying, and delayed copying, which requires sketching from memory and copying line-drawings either directly or after a time delay.

Once trained, healthy naming and sorting are possible in all models except the BM. This is due to the inherent stochastic nature of BMs, the need for extra training to better learn the mapping between visual input

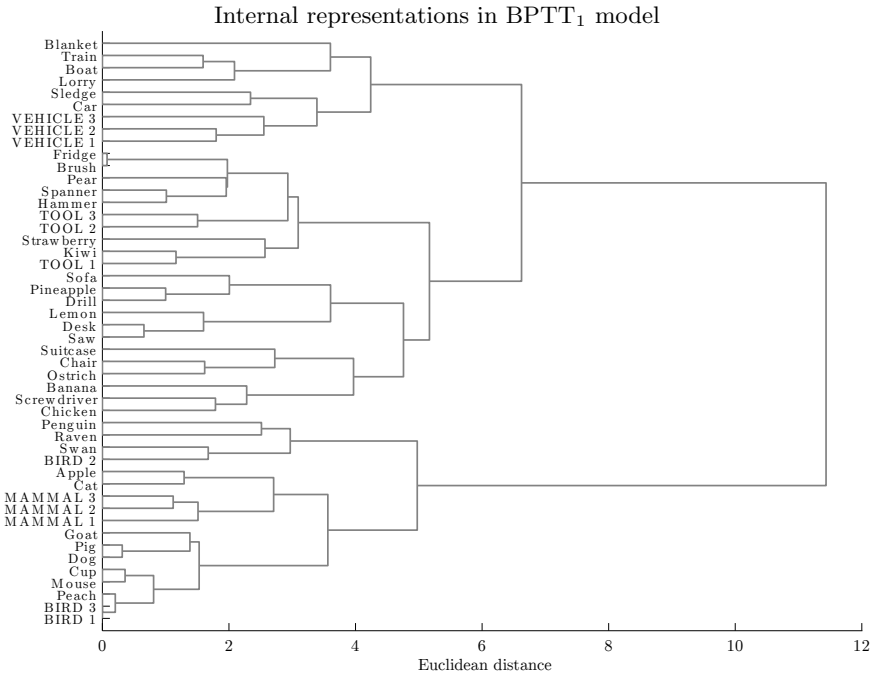

Figure 5: Dendrogram for BPTT$_1$ internal semantic states after 30% connection weights lesioned sampled 100 times per name sub-pattern input.

and name output, and the design of the two tasks. The naming and sorting tasks make use of localist interpretations for names and category membership, which are not part of the BM's learning strategy. This issue can be addressed given greater training, although with the training given in this study the BM does score at ceiling on the word-to-picture matching and the drawing, copying, and delayed copying tasks, which use distributed representations.

As required by the hub theory, all the networks have internal states that allow for the mapping of the perceptual inputs to the output modalities, thus completing each of the four semantic tasks. Fundamentally, the internal semantic space must mirror the categorical and domain structure of the training set. This attractor-space can be represented using a dendrogram as in figure 4, which shows the Euclidean distance between both individual concepts and between categories and domains. This allows a comparison between the intended categories and those that arise from the structure of the learned attractor states (cf. figure 5, Rogers et al., 2004).

Rogers et al. (2004) provide a list of qualitative properties that their model's internal representations possess. As shown in figure 4, our versions also conform to this list. Firstly, the two domains, animals and artefacts, are clearly separated from each other, as are to a lesser extent the six categories. Secondly, the model's representation of category-level names (e.g., "BIRD 1") are classed within their category cluster. And finally, fruit are classed under the domain of inanimate objects, but are in a distinct cluster to the the rest of the artefacts.

3.2. Damaged behaviour

Since our reimplementations have healthy internal representations as found in the original hub model, damage can be applied to cause disruptions to the attractor basins. SD-like damage is modelled by setting increasing proportions of all connection weights to zero. This causes the network to be less adept at completing semantic tasks, as propagation of activations both within the hub and between it and its spokes is impaired. Disconnection has a pronounced effect on the semantic attractor landscape; the network can now only manage to represent a subset of the previous 48 concepts, which can be seen in figure 5. The clusters corresponding to concepts, categories, and domains are now deformed, e.g., the attractors for "cup" and "mouse", from opposing domains, are now in the same semantic cluster. This merging of conceptual representations from different domains, as opposed to categories within the same domain of knowledge, appears to signal a deviation from the hub theory's requirements.

In the next few sections, we focus on behaviour in the naming and sorting task, as these two tasks are the most problematic in terms of accounting for the patient data in the reimplementations[1]. These two tasks results depend heavily on the network settling to the correct internal state after lesioning damage.

3.2.1. Confrontation naming

The confrontation naming test involves producing an appropriate word when given a visual depiction of an object. For the group of patients this involves a line-drawing and an experimenter to record their response. For the models, following the original task design, naming involves clamping the visual units (representing the input to semantics when looking at a picture) and then allowing the network to cycle for twelve settling steps (cf. Rogers et al., 2004, p. 217). After that the visual units are unclamped and the

[1]Rogers et al. (2004) also consider behaviour of the hub model on word-to-picture matching and in the drawing, copying and, delayed copying task. Our reimplementations replicate these results.

network is allowed to settle until equilibrium. When the network reaches a stable end-state the name output is inspected and whichever name unit is found to be most active, above a threshold of 0.5, is taken to be the model's response. If no unit has an activation above the threshold, the response is classified as an omission. This method is not applicable to the binary unit states of the BM. Hence, for the BM, the response is derived by finding the

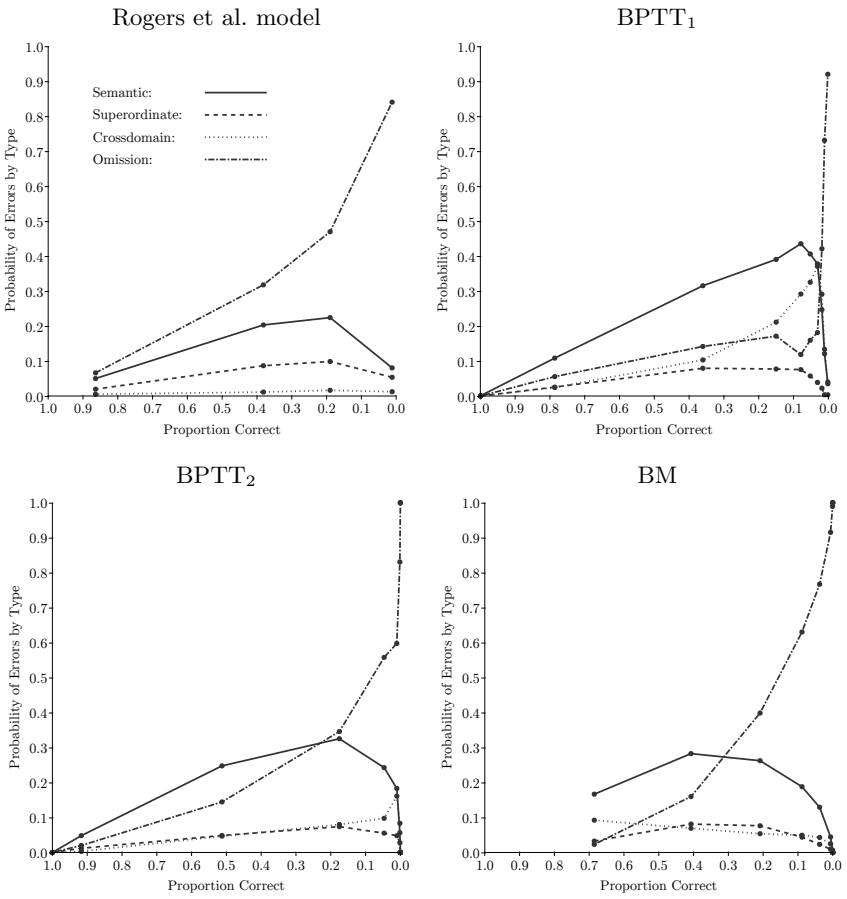

Figure 6: Results of the naming task for each model. Each data point represents the proportion of error for each type of error at a percentage of connections lesioned: from 0% to 90% in increments of 10% sampled 500 times; in the case of the Rogers et al. results are at 10%, 20%, 25%, and 35% connections lesioned taken from fig. 6, Rogers et al. (2004).

pattern closest in Euclidean space to the name output. If more than one name unit is activated the confidence of the BM is taken to be the reciprocal of the number of active name units; any reciprocal under 0.5 is an omission.

There are four important qualitative features of the naming scores found in both the longitudinally tested SD patients (cf. figure 6, Rogers et al., 2004) and in the original model (reproduced in the top left sub-graph of figure 6). Firstly, the overwhelming proportion of errors consists of *omissions*, meaning the patient is anomic or the model is unable to activate any name units above the threshold. Omission errors are seen to increase with the progress of neurodegeneration. Secondly, *semantic errors*, which involve confusing an item with another from the same semantic category (e.g., calling a mouse "dog"), initially start off low, then grow to about a quarter of responses, and finally return to a lower proportion. Thirdly, *superordinate errors*, giving a category name to an item that would not be labelled as such by a healthy participant (e.g., calling a dog "mammal"), show the same pattern as semantic errors. Although at all levels of lesioning superordinate errors are lower than semantic errors, reaching only about a tenth of all responses at their highest proportion. Fourthly, *crossdomain errors*, giving an item a name from the opposing semantic domain (e.g., calling a hammer "dog") are extremely rare in both the fifteen Rogers et al. patients and in the original model.

The $BPTT_1$, $BPTT_2$, and BM naming graphs in figure 6 show only a partial replication of the naming task scores as discussed above. Firstly, omissions are lower than semantic errors, but in fact they should be consistently above all other errors. Secondly, semantic errors are proportionally the highest error type. Thirdly, superordinate errors are qualitatively a good fit. Fourthly, crossdomain errors occur, when instead they should be at floor levels. This pattern of responses persists even if the value of the threshold, which determines the proportion of responses that are classified as omissions, is varied.

The results of the confrontation naming task run on the three different implementations show that internal representations do not decay in a way that replicates the patients' scores. So while healthy naming is possible, at least within the BPTT reimplementations, the predictions made by Rogers et al. (2004) are not met. Specifically, they claim that "[w]ith increasing damage, the model becomes unable to generate any information that individuates items from the same broad domain, and representations within a given domain collapse into a single general attractor from which the model produces only those properties common to the majority of items in the domain. [That is to say, t]he model never names an object with a completely unrelated label, because such names apply only to objects

with very distal internal representations" (Rogers et al., p. 218). However, we can see from both the damaged semantic representations $BPTT_1$ has, shown in figure 5, and from the naming scores in figure 6, that concepts from opposing domains can become much closer to each other than (what should be) neighbouring concepts. This is why a larger proportion of cross-domain errors are produced: attractor dynamics do not necessarily follow the predictions set out by the hub theory.

3.2.2. Sorting words and pictures

The sorting task is used to determine the preservation of hierarchical conceptual knowledge in patients. It is carried out by classifying words and pictures into the five categories (Rogers et al., 2004, exclude fruit during testing) and into the two domains, respectively named *specific sorting* and *general sorting*. This semantic task is modelled in the same way as the naming task with regards to settling, by clamping the target for twelve settling steps, then removing the target and allowing the model to reach equilibrium (Rogers et al. use this method for all the tasks). Once the network is in a stable state, the verbal units which represent category or domain membership are examined (cf. Rogers et al., p. 220). For general sorting, the domain unit for animals or for artefacts is used to determine the response of the network, and for specific-level sorting category units are inspected. Whichever unit is most active is taken to be the model's response.

Figure 7 shows the the original model behaviour in the top left corner. This reflects the pattern of the twelve patients tested by Rogers et al. (2004), in particular: *a*) the sorting of pictures is more preserved than that of words; *b*) sorting at a general level is retained more so than specific sorting; and *c*) the ability to classify pictures into their respective domains is largely unaffected by lesioning.

The $BPTT_1$ results, shown in the top left sub-graph of 7, indicate that the last of these properties is absent; scores in general picture sorting should be near or at ceiling even after substantial (40%) lesioning. The graphs for the $BPTT_2$ and BM models do not display this property either, but nor do they consistently show the other two qualitative effects. In contrast to the original model, the scores of the three reimplementations for all types and levels of sorting tend towards baseline values (chance for category-level sorting is 0.2 and for domain-level is 0.5 – any slight deviation from these is due to the values of the bias units). Rogers et al. (2004) propose that their model of the sorting task is able to follow the patients' scores because "the effect of damage must be quite severe before the system begins to generate incorrect verbal information about such properties" (Rogers et al., p. 220).

The $BPTT_1$ reimplementation manages to show a partial replication, however the $BPTT_2$ and BM do not reflect any aspect of the patient scores consistently. So while in the original model the "difference in the nature of the mapping between surface form and conceptual representations [...] underpins the difference in performance for word and picture sorting" (Rogers et al., 2004, p. 221), this does not hold as strongly for the $BPTT_2$ and BM.

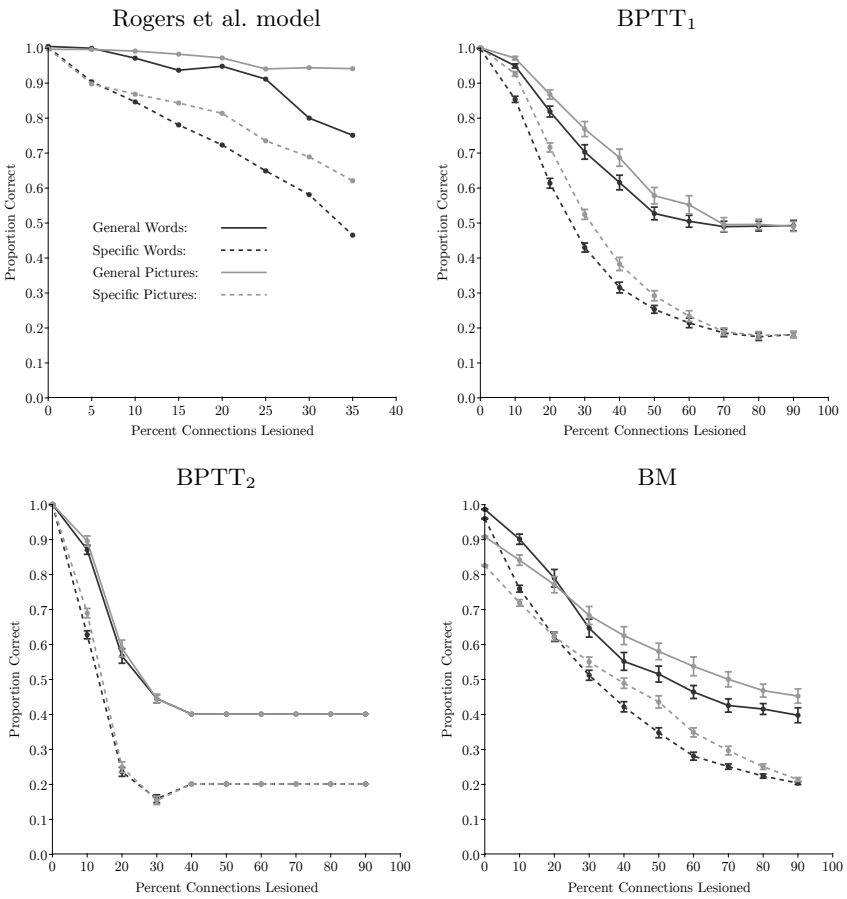

Figure 7: Results of the sorting task on words and on pictures at a general and specific level of categorisation. Each data point represents the score at the level of lesioning sampled 500 times. Error bars – where present – indicate one standard error about the mean. NB: difference in x-axis scale between the Rogers et al. (2004) model and the three reimplementations.

In addition, in the original hub model "[a]rbitrary mappings are more vulnerable to damage than are systematic mappings" (ibid, p. 221), meaning that word sorting is more fragile than picture sorting; however, this also does not generalise to all our reimplementations.

It should be noted that qualitatively equivalent naming and sorting scores as those seen in the reimplementations in figures 6 and 7 are found over many instances of $BPTT_1$, $BPTT_2$, and BM networks (i.e., the results are not an artifact of one set of trained weights). In addition, the training algorithm of the BPTT networks has been varied between epoch-wise, pattern-wise and sub-pattern-wise (weights updated after each name, verbal, visual sub-pattern) and it has been found to also produce qualitatively equivalent naming and sorting graphs.

4. Discussion

McClelland argues that:

> When a model fails to capture some aspect of human performance, it represents both a challenge and an opportunity. The challenge is to determine just what aspect or aspects of the model are to blame for the failure. Because the model is an exploration of a set of ideas, it is not clear which members of the set are at fault for the models shortcomings. Model failures also present an opportunity: When a model fails it allows us to focus attention on where we might have been wrong, allowing real progress to arise from further investigations.
>
> McClelland (2009, p. 21)

The work presented here demonstrates that the implications of the ideas embodied in the hub theory do not necessarily capture some aspect of human performance. In other words, because reimplementations do not show the same pattern of errors, as a consequence of not showing an equivalent decay of attractor basins, our results represent both a challenge and an opportunity for further research. The original hub model and the reimplementations we present here constitute an exploration of a set of ideas, some of these ideas might lead to conclusions or give rise to phenomena that might not have been uncovered a priori. By running different models based on a theory, as McClelland claims, the repercussions of the ideas explicitly and implicitly contained in the theory can be illuminated and explored.

Within the hub theory, attractors break down in ways that are predictable and this breakdown is the phenomenon proposed to account for the semantic impairments documented in patients. However this does not

appear to hold for all implementations of the hub theory. Specifically, while our three reimplementations are adept at patient modelling on the word-to-picture and drawing and delayed copying tasks, they do not fare well when reproducing patient scores in the naming and sorting tasks; nor do our models exhibit the required pattern of breakdown in their internal representations. This means that the ideas encapsulated within the hub theory can lead to models that are not fully in line with the higher level aims of the theory, i.e., to explain the effects of the neurodegeneration caused by semantic dementia on the semantic cognitive system.

In the original hub model, Rogers et al. (2004) describe the breakdown in performance of the hub model following damage as arising because "small amounts of drift may lead the network into an inappropriate proximal attractor, [thus making the model] produce incorrect responses appropriate to a semantically related object[, meaning that the attractor space is] robust even to relatively large amounts of damage, because the system's internal representations must be severely distorted before they drift out of the region to which such properties apply" (Rogers et al., p. 229). This has been shown not to hold for our reimplementations, as errors have been documented that are not semantic relations of the target response, instead they are from the opposing domain of knowledge. This is not documented in the original model, or the patients. In the reimplementations presented here it occurs even given relatively small amounts of lesioning damage.

Why might our reimplementations, when damaged, fail to reproduce the behaviour reported by Rogers et al. (2004)? One possibility is that the pattern of breakdown of attractors as required by the hub theory is not a necessary consequence of a recurrent neural network trained with the structure of the training set. The hub theory assumes that attractors drift apart and merge in certain ways, as a consequence of the underlying recurrent neural network substrate, without *requiring* this at a theoretical level. But this assumption does not always hold. Based on this disparity between models and theory, it appears that the hub theory is underspecified as different implementations behave differently. Therefore, some additional theoretical constraint is required if models that implement the hub theory are to be consistent with the patients' behaviour. In our view this constraint should concern the behaviour of attractors following lesioning.

Acknowledgements

We are grateful to Jay McClelland, Tim Rogers, Matt Lambon Ralph, and Anna Schapiro for clarifying aspects of the original hub model.

References

Ackley, D. H., Hinton, G. E., & Sejnowski, T. J. (1985). A Learning Algorithm for Boltzmann Machines. *Cognitive Science*, *9*, 147–169.

Hinton, G., & Sejnowski, T. (1986). Learning and relearning in boltzmann machines. *MIT Press, Cambridge, Mass.*, *1*, 282–317.

Lambon Ralph, M. A., Lowe, C., & Rogers, T. T. (2007). Neural basis of category-specific semantic deficits for living things: evidence from semantic dementia, HSVE and a neural network model. *Brain*, *130*, 1127–1137.

McClelland, J. (2009). The place of modeling in cognitive science. *Topics in Cognitive Science*, *1*(1), 11–38.

Rogers, T., Garrard, P., McClelland, J., M.A. Lambon Ralph, Bozeat, S., Hodges, J., et al. (2004). Structure and deterioration of semantic memory: A neuropsychological and computational investigation. *Psychological Review*, *111*(1), 205–235.

Williams, R., & Zipser, D. (1989). A learning algorithm for continually running fully recurrent neural networks. *Neural Computation*, *1*(2), 270–280.

Williams, R., & Zipser, D. (1995). Gradient-based learning algorithms for recurrent networks and their computational complexity. *Backpropagation: Theory, architectures, and applications*, 433–486.