Semantic Cognition: A Re-examination of the Recurrent Network "Hub" Model

Olivia Guest (O.Guest@bbk.ac.uk) and **Richard P. Cooper (R.Cooper@bbk.ac.uk)**
Department of Psychological Sciences, Birkbeck, University of London
Malet Street, London WC1E 7HX, UK

Abstract

This paper explores a model of "semantic cognition" first described in Rogers et al. (2004). This model was shown to reproduce the behaviour of neurological patients who perform poorly on a variety of tests of semantic knowledge; thus purporting to provide a comprehensive explanation for semantic deficits as found in patients with semantic dementia and, as extended in Lambon Ralph, Lowe, and Rogers (2007), individuals with herpes simplex virus encephalitis. Therefore, not only does the model emulate these semantic impairments, it also underpins a theoretical account of such memory disturbances. We report preliminary results arising from an attempted reimplementation of the Rogers et al. model. Specifically, while we were able to successfully reimplement the fully-functioning model and recreate "normal" behaviour, our attempts to replicate the behaviour of semantically impaired patients by lesioning the model were mixed. Our results suggest that while semantic impairments reminiscent of patients may arise when the Rogers et al. model is lesioned, such impairments are not a necessary consequence of the model. We discuss the implications of these apparently negative results for the Rogers et al. account of semantic cognition.

Keywords: semantic memory model; semantic dementia; backpropagation through time.

Introduction

Several connectionist models of "semantic cognition" have been developed. The goal of such models is to reproduce results obtained from testing both healthy and semantically compromised individuals on tests held to tap semantic knowledge. This is accomplished by first teaching the model to function like a healthy semantic system and then by "damaging" the model in a way that parallels the lesions seen in patients. Typically, models implement a theoretical framework that aims to explain the semantic system in both a normal and degenerate state, and evidence in support of the framework is adduced by appealing to the behaviour of the fully-functioning and lesioned model.

A complete understanding of semantic deficits, and of semantic memory in general, has not yet been reached. However sophisticated attempts to account for the deficits of some of the patient populations have been made, such as the Rogers et al. (2004) model. This model proposes that a central amodal semantic "hub" is reciprocally linked to modality-specific "spokes", which themselves extend into so-called modal pathways. This connectivity allows fully grounded perceptual input to give rise to amodal abstract concepts. In other words, the hub itself creates semantic representations via its recurrent connections to sensory regions of the brain.

The hub model developed by Rogers et al. (2004), and then extended in Lambon Ralph et al. (2007), is one of the most complete models of human semantic deficits, boasting both an account of semantic dementia, a global semantic memory disorder, and herpes simplex virus encephalitis, a cause of intra-semantic deficits. This paper reports results derived from an attempted reimplementation of the hub model, performed initially as a step towards extending the model and underlying theoretical framework to provide an account for additional semantic disorders. During the process of exploring this model, it became apparent that some of the results reported in Rogers et al., obtained from modelling clinical tests of semantic knowledge, were not robust. That is, the Rogers et al. results are not a necessary consequence of the model as described. This suggests that the computational-level description of the human semantic system offered by Rogers et al. is under-specified.

Semantic Cognition

The *semantic memory system* refers to a part of human long term memory consisting of a collection of abstract facts about the world. Semantic knowledge underpins linguistic meaning, providing a substrate for reasoning and inference, for categorisation, and for the creation of prototypes or exemplars. It intuitively appears that semantic memory is an abstraction or generalisation over a set of experiences collected gradually over time and organised hierarchically, as first proposed by Collins and Quillian (1969).

Semantic cognition pertains to the process by which a non- or pre-semantic percept (e.g., a drawing of a dog, or the word "dog") gives rise to a collection of related semantic memories (e.g., dogs are fury, and have four legs) that endow the percept with meaning. The reflex-like recollection of this knowledge produces a response related to the specific concept (e.g., identifying a line-drawing by saying: "dog"). Such a reaction is only possible if the relationship between the purely perceptual stimulus and its meaning has already been instantiated in the mind (e.g., an image of a dog is linked to the phonetics of the word "dog"). This definition implies semantic cognition can be explored using tasks that require a correct interpretation, and thus response, when probed with an appropriate stimulus. Four such tasks are used by Rogers et al. (2004) to assess both their participants' and their model's aptitude; these are: *confrontation naming*, where an appropriate verbal name must be provided for a picture; *word-to-picture matching*, where a linguistic label must be paired with its corresponding picture from a selection that includes distractors; *sorting*, where a selection of words or pictures must be classified under hierarchical categories; and *drawing, copying and delayed copying*, where three sketches must be created, the first recreated purely from memory in response to a word, the second by direct copy from a line-drawing, and the third from memory a short time after the direct copying subtask.

Patients with dramatically low scores on such tests were

first described by Warrington (1975). Her patients, who were in their early sixties, were tested on many aspects of their cognitive functioning in order to isolate their deficit as one of pure semantics and not one of an intellectual, perceptual, or linguistic nature. The set of behavioural symptoms found, coupled with progressive bilateral neurodegeneration of the anterior temporal lobes, are characteristic of a disorder that has come to be known as *semantic dementia* (SD), a variant of frontotemporal dementia. As seen in Warrington's study and in the many more that followed, SD causes a severe impairment of semantic knowledge, with patients performing better when tested on familiar or typical items as opposed to novel or exceptional ones.

The degenerative nature of SD appears to cause patients' semantic skills to disappear in a process of akin to the reverse of learning. This, along with the characteristics of other semantic disorders, hints at some form of functionally distinct hierarchical system in which structural damage is intrinsically linked with, and gives rise to, functional deficiencies.

The Hub Model

Overview

A central claim of the hub model of Rogers et al. (2004) is that the interactions of *attractors*, which develop through learning to represent amodal concepts within semantic space, can account for both healthy and deficient semantic cognition. Attractors are stable network states that emerge following training if recurrent connectivity exists within a connectionist network. When activation is allowed to propagate throughout the trained network in a cyclic fashion, the network's state (as represented by the set of hidden and visible unit states) will converge to one such stable configuration. These stable network states exercise attractive power over a set of neighbouring network states, collectively known as their basin of attractor, such that if the network is in any of these "nearby" states it will ultimately settle to the attractor itself. These properties, according to Rogers et al., are also found in semantic memory.

To evaluate their framework Rogers et al. (2004) develop a recurrent connectionist network model. A set of stimuli is created based on statistically analysed features of common percepts (McRae & Cree, 2002), which the model is taught to auto-associate. Post-training, the model scores on tests of semantic cognition in accordance with healthy participants. After lesioning, the impaired model exhibits deficits comparable to those of SD and HSVE patients. Thus, Rogers et al. conclude that their architecture captures some level of the internal mechanisms and sophistication of the human semantic system.

Structure and Processing in the Hub Model

The recurrent connectionist network of Rogers et al. (2004) consists of one layer of 215 visible units and one layer of 64 hidden units. The latter are fully connected both to themselves and to the visible units, which are divided into three in/output pools each consisting of: 40 name units, 64 visual feature units, and 111 verbal (61 perceptual, 32 functional, and 18 encyclopaedic) descriptor units. All units have real-valued time-varying activations with a range of $[0, 1]$ and a bias set to -2. The hidden units, through learning, come to represent a kind of amodal semantics associated with feature patterns represented at the visible units.

As discussed above, the stimuli on which the network is trained and tested are binary patterns with co-variance that reflects statistical properties of real-world concepts. These are directly applied to the name, verbal, and visual units. Name sub-patterns are a set of binary digits, of which only one unit may be active per pattern, i.e., they are defined orthogonally. Rogers et al. (2004) argue that this labelling strategy parallels natural language in as much as, for example, the word "robin" does not in itself carry any information about the bird to which it refers. In contrast, the visual and verbal sub-patterns represent perceptual and linguistic information, and therefore must conform to predefined prototypes. Visual properties and verbal descriptors represent statements like "has a red breast", "can fly", and facts such as "is a bird" and "is living".

To produce a response given a sub-pattern the network effectively performs pattern completion. It propagates activations until it reaches a stable state in which hidden unit states do not change on successive cycles. Once the trained network has settled, its semantic state conforms to the real-valued pattern of an implicitly learned attractor, an internal configuration that is reachable due to the recurrent connectivity of the hidden units. This in turn activates the output units, thus completing the input pattern.

Training Strategy

Pattern Set The set of patterns used by Rogers et al. (2004) to train the hub model has some very particular properties. Specifically, it contains some patterns in which visual and verbal sub-patterns are mapped onto the same name. The sharing of name sub-patterns is held to be analogous to the way a chicken, a robin, and a sparrow can all be called birds, both individually and collectively. What this amounts to here is, for example, 3 nondescript birds sharing the superordinate level name "BIRD"; forming a unidirectional 3-to-1 mapping from the three pairs of visual and verbal sub-patterns to a single name label. Conversely, if given "BIRD" their network "learned to generate visual and verbal properties common to most [birds]" (Rogers et al., 2004, p. 214). Based on the statistical properties of visual and verbal co-occurrences within various categories reported by McRae and Cree (2002), Rogers et al. constructed a set of 48 patterns, with 8 patterns for each of 6 categories (mammals, birds, tools, vehicles, household objects and fruits, although only the first four have associated category-level exemplars), and 40 unique names.

In order to replicate the hub model, we constructed a statistically equivalent set of patterns, based on the probabilistic prototype for pattern creation given in fig. 3 of Rogers et al.

(2004). A comparison of the resulting dendrogram showing pattern similarity with that of fig. 2 of Rogers et al. confirms the two pattern sets are equivalent in structure.

Learning Algorithm The learning algorithm used by the original network is described only as "a variant of the back-propagation learning algorithm suited to learning in a recurrent network" (Rogers et al., 2004, p. 208). J. L. McClelland (personal communication, 2011) confirmed that this was a variant of backpropagation through time (BPTT), with the network "unrolled" (allowed to run) for 28 time-steps (Rogers et al., 2004, p. 215).

In the work reported here we adopt classic epochwise BPTT (Williams & Zipser, 1995, p. 447, eq. 18-19), with a learning rate of 0.001 and with time-averaging applied to post-synaptic unit states (McClelland, 2011). Time-averaging is a statistical method of noise reduction that may be applied over any time-varying property of a dynamic system. It has the ability to increase the signal-to-noise ratio and, in this case, results in a decrease in training epochs and for more complex mappings to be internalised, given the training details in Rogers et al. (2004)

Healthy Behaviour of Hub Model

After training for 15,000 cycles, our replication of the Rogers et al. (2004) network robustly maps names to visual and verbal sub-patterns. Thus, given a name such as "chicken", the visual and verbal units of the network take on patterns (once the network has settled) that correspond to the visual and verbal features associated with "chicken". Similarly, when given the visual features of that pattern, the other visible units take on values associated with the name and verbal features of the pattern. More critically, when given a superordinate name the sets of units corresponding to visual and verbal sub-patterns take on states that amount to the weighted average of the three nondescript patterns that share that same name. Conversely, when provided with the visual or verbal descriptors the network activates the general-level name. This demonstrates that the network has created stereotypes or archetypes for each category.

Semantic Tasks

Overview

We tested our network on each of the four tasks described in the introduction. In each case the method used to probe the network consists of: keeping the relevant input constant while running the network for 12 time-steps; then allowing the network to settle without any externally applied input until equilibrium is reached; and finally comparing the states of the units in the pool currently of interest to those in the relevant pattern. This is as described in Rogers et al. (2004). Following training, our network functions in its healthy state at the same general levels as Rogers et al., both in terms of training error and on all four tasks. It is therefore appropriate to consider the network's behaviour following lesioning.

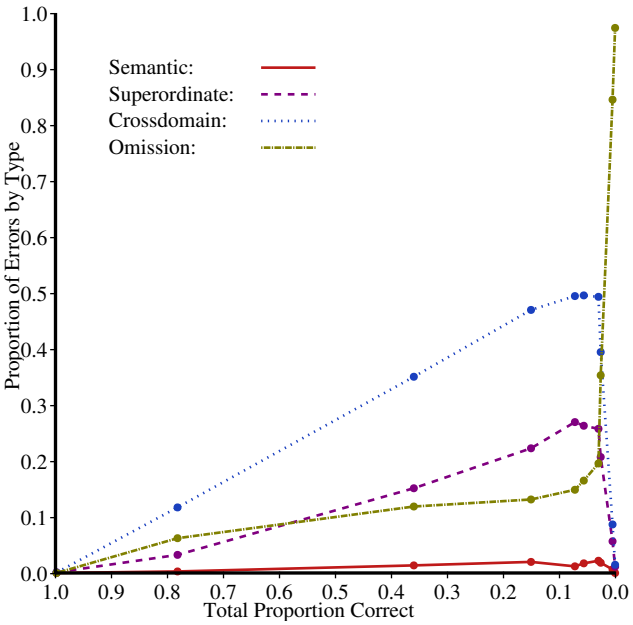

Figure 1: Results of the confrontation naming task: each data point represents the proportion of error for each type at a percentage of connections lesioned: from 0% to 90% in increments of 10%. (Compare with Rogers et al., 2004, fig. 6.)

Lesioning the Model

Rogers et al. (2004) lesioned the original hub model by indiscriminately globally severing connections between units. This zeroing of the weights is claimed to be a sufficient analogue to the damage seen in the temporal lobes of SD patients. By removing randomly selected connections in increasing percentages Rogers et al. maintain that the network displays neurodegeneration-like behaviour reflecting the decay in SD. This approach is mirrored in our replication, with semantic testing performed on three of the four standard tests described previously. Each lesion begins from a random selection of weights and is then increased, as would occur over time to an SD sufferer. For each task this is done 50 times at 10 levels of damage (i.e., from 0% to 90% of connections removed); paralleling 50 SD patients tested at 10 stages of progressive degeneration. Once the network is lesioned, settling becomes increasingly difficult and may result in dramatically different responses given the same input; thus the all results are based on sampling the current implementation 10 times for each of the sub-patterns it is tested on.

Confrontation Naming

Recall that confrontation naming requires subjects to generate verbal labels (names) from visual input (pictures). Rogers et al. (2004) report data on this task from 15 SD patients. At the earlier stages of degeneration, *omissions*, when the participant gives no answer, are relatively few but they increase dramatically as time goes by, until the only errors are omissions, i.e., the individual is completely anomic. *Superordinate er-*

rors, so called because the response is not the expected name (e.g., "owl"), but something more general (e.g., "bird"), seem to follow a similar trend to omission errors. However, at the most severe stages of the disease, superordinate errors drop off due to anomia. *Semantic errors* occur when the response is from the same category as the line-drawing presented (e.g., "dog", when the correct answer is "horse"); these errors are low initially, then rise, and finally return to a low level (again due to anomia). *Cross-domain errors*, where a response is given from the opposing domain to that which the stimulus belongs to (e.g., calling a "horse" a "car"), are almost never documented in the SD sample.

The results of our replication of the confrontation naming task are shown in fig. 1; however, the trends shown in the behaviour of the SD patients described above and the modelling results of Rogers et al. (2004) are not shown here. In regards to our model, the largest proportion of errors from 10% to 70% of weights lesioned are cross-domain errors. This means that name units corresponding, for example, to artifacts are activated when an animal is visually presented to the network and vice versa. Omission errors are defined by Rogers et al. to occur when the network fails to activate any name unit beyond a threshold of 0.5. Changing this threshold affects the relations between the error types, but does not result in a better fit to patient data. The greater the threshold the more errors are classified as omissions, and thus the remaining three kinds of naming error (semantic, cross-domain, and superordinate) are fewer; the inverse also holds. In conclusion, the reimplementation of the hub model on the naming task does not recreate the error pattern seen in the patients.

Sorting Words and Pictures

This task requires the network to classify name and visual sub-patterns into their respective categories and domains. In fig. 2, a graph of the network's performance at sorting at increasing levels of lesioning is shown. The scores for the two general levels of sorting (represented as solid lines), for words and for pictures, follow a descent from correct to chance levels. This is expected due to the architecture of the patterns: there are two encyclopaedic units that represent the mutually exclusive facts "is an animal" and "is an artifact". In much the same way, the network's scores on the two specific sorting tasks also appear to deteriorate to chance level, this time as there are 5 categories to choose from chance is at 0.2 (as in Rogers et al., 2004, fruit is excluded in the testing phase).

These results are relatively similar to those produced by the 12 patients tested by Rogers et al. (2004), however, there appears to be an important difference: the SD patients retain the ability to classify pictures into their respective domains well into their illness. Thus, while sorting into lower level categories is a skill that is largely lost, the two main semantic domains remain intact in SD; this also can be seen in fig. 8 of Rogers et al. While the original hub model appears to capture this dissociation, the current implementation does not. Arguably, the sorting of pictures is slightly more preserved than that of words, in fig. 2, but the SD patients are all at ceiling.

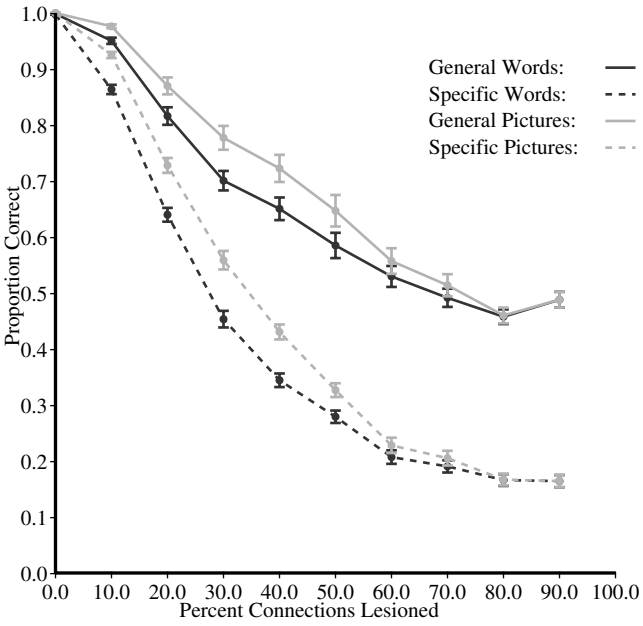

Figure 2: Results of the sorting task on words and on pictures. Error bars indicate one standard error (SE) about the mean. (Compare with Rogers et al., 2004, fig. 8.)

Again, the model is unable to fully capture this pattern of SD patient performance.

Drawing and Delayed Copying

This semantic test involves creating drawings given a name and copies based on visual sub-patterns. The results obtained from running the drawing and delayed copying semantic test on the reimplementation (see fig. 3) appear to qualitatively match those in fig. 11 of Rogers et al. (2004). Both SD patients and the model show an increase in the errors they make when drawing and copying. Also the difference between drawing and delayed copying, that the former is more difficult than the latter per patient, is reflected in both the original model and our reimplementation.

However, when the results are further analysed, as in figs. 4 and 5, a different picture emerges. Rogers et al. (2004) argue that there is an underlying distinction between the scores in each domain for two kinds of error: an *omission*, a salient feature that should have been drawn but is left out by the participant (e.g., forgetting to depict a swan with wings); and an *intrusion*, a property that perhaps holds for most exemplars but is incorrectly included in the drawing (e.g., adding four legs to a swan). In the patients' drawings there are significantly more intrusions for animals than for artifacts (Rogers et al., 2004, p. 227), but no such effect for omissions. In fact, the original hub model only partially reproduces these effects, correctly showing more intrusions for animals but incorrectly showing more omissions for artifacts (see figs. 12-13 in Rogers et al., 2004). In our reimplementation of the hub model, we found that omission errors (both when copying and drawing) are higher in artifacts over animals (see

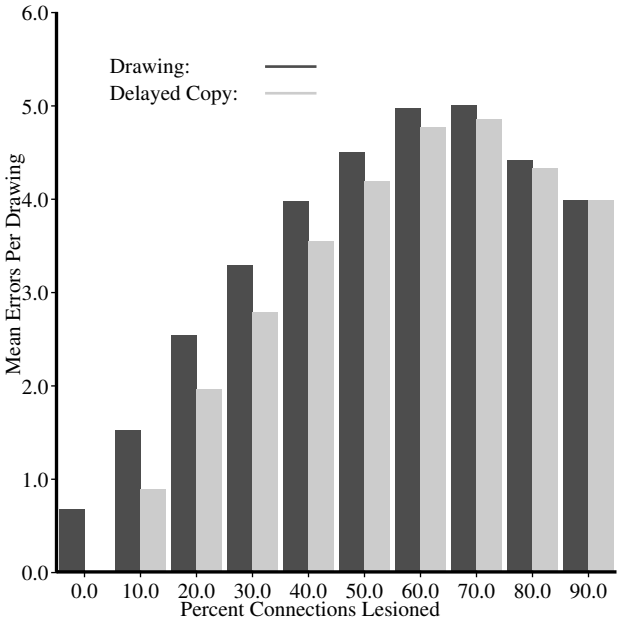

Figure 3: Mean overall feature errors per drawing for the drawing and delayed copying task for each lesioning level. Error bars not included because SE < 0.002. (Compare with Rogers et al., 2004, fig. 11.)

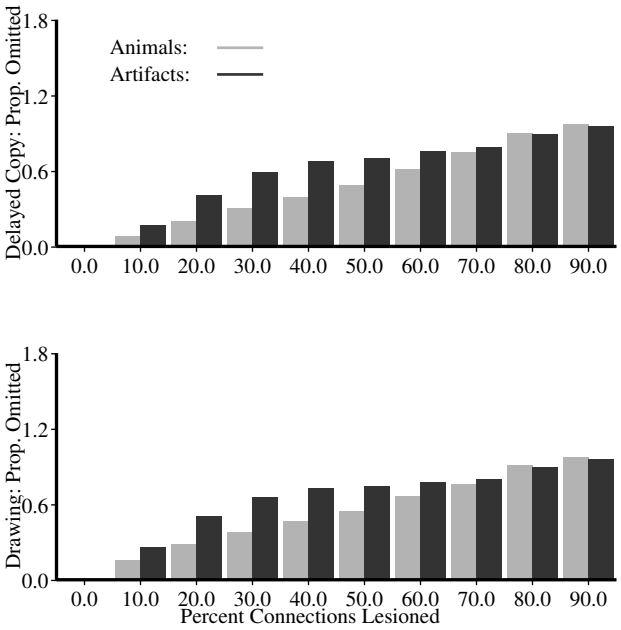

Figure 4: Proportion of errors of omission per drawing for each domain for the drawing and delayed copying task. SE $<$ 0.003. (Compare with Rogers et al., 2004, fig. 12.)

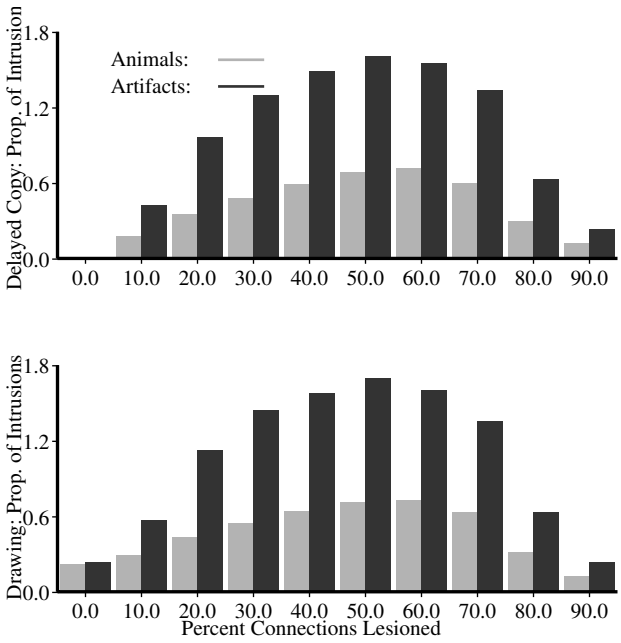

Figure 5: Proportion of intrusion errors per drawing for each domain in the drawing and delayed copying task. SE < 0.105. (Compare with Rogers et al., 2004, fig. 13.)

fig. 4), though with increased lesioning severity omission errors seem to occur equally in both domains (in contrast to patients, who show no effect; while the original hub model shows the same effect as our model). The rate of intrusion errors by domain reflects neither the patient data nor that of the original hub implementation, with more intrusion errors for artifacts than animals over most of the range of lesion severity (see fig. 5).

Discussion

Rogers et al. (2004) presented a model of the semantic system which they argued could account, when lesioned, for many of the deficits associated with semantic dementia. In support of this argument they report a number of simulations. We have attempted to replicate these simulations, but with mixed success. Thus, while we were able to recreate the basic learning performance of the model, we were unable to fully reproduce the patterns seen in the lesion studies.

Rogers et al. (2004) parallel the emergence of attractors with the learning of concepts, and propose that such knowledge is amodal: the somato-sensory input from the various modality-specific pathways is encapsulated by the hidden units, which thus form semantic representations. This basic theoretical notion is successfully captured by the hub model. For the case of the deficits seen in their SD patients, Rogers et al. appeal to the attractor basins' properties post-lesioning (zeroing of connection weights). They claim that animals are a tight cluster of similar concepts, thus consisting of many neighbouring attractors, while attractors for arti-

facts are distal (to the average central point of their domain), which means they form distinct conceptual loci in semantic space, and therefore their attractors are further apart. When connections are zeroed the attractor basins for living creatures are held to decay to form a larger super-attractor, which has a combined attractive power; meaning categorisation of input as an animal is possible, but access to individual features might be lost. Conversely, the attractor basins of non-living things do not merge; instead they maintain their individual attractors, albeit with distorted basins, allowing slightly better performance in this domain. The evidence put forward for the this phenomenon is the series of graphs generated from testing the Rogers et al. model. Yet the behaviour reported in the original hub model is not found in the network trained here. Why might this be so?

One possibility is that there is an error in our replication. We do not believe this to be the case, particularly given that we have simulated the basic learning performance of the network. A second is that the difference in results relates to some difference between, for example, the learning algorithm as implemented here and as implemented by Rogers et al. (2004). This is certainly possible, given that the algorithm is not fully described in the original publication. A third is that the attractors formed by the model are dependent upon the initial random weights of connections prior to learning or the order of exemplars in the training set. However, if either of these latter two situations is the case then it calls into question the theoretical explanation offered by Rogers et al. for their results.

An important aspect of this modelling strategy, that is related to the formation of attractors, is the claimed distribution of pre-semantic (perceptual and functional) features: animals and plants are closely perceptually related to each other (due to the fact they have evolved from a common ancestor and thus are composed of generally similar body parts); whereas tools, vehicles, and other inanimate objects are not similar to each other (as they have been created by humans to solve different problems, so by definition artifacts are distinct from both living things and from each other). Without training sets that encode patterns in this specific way, no connectionist model would be capable of producing a good fit to patient data. On this argument, the features, whose extraction from the environment itself is not modelled, play a pivotal role in giving rise to the semantic system's structure, and this is the case regardless of the network topology (be it recurrent or feedforward) or the learning algorithm. This is to say that, to a large extent, input to the semantic system should drive its organisation and dictate the way semantic knowledge will decay. Despite this fact, the patterns used here are unable to affect the internal structure of the reimplemented hub model in the way needed when the network is damaged. This means that the qualitative and consistent effects required post-lesioning are in fact *not* guaranteed merely by the structure of the training set. It appears that lesioning the recurrent network model by severing connections does not necessarily result in the kind of well-behaved breakdown and generalisation of attractors as supposed by Rogers et al.

To summarise, the differences between the models appear to be due to the results obtained in Rogers et al. (2004) depending on some unarticulated implementation detail. If this is so, then the required behaviour is not a necessary consequence of the model – the original model is underspecified (perhaps our implementation of the BPTT algorithm yields attractors with different properties to the implementation of Rogers et al.). Alternatively, it may be that the behaviour of the network when damaged depends upon, for example, some apparently irrelevant factor such as the random initialisation of the connection weights. Whatever the underlying cause of the discrepancy, further investigation is needed to discover exactly why the results obtained here differ from most of those detailed in Rogers et al. If their results are in fact reproducible, but require a very specific set-up, this suggests that the model as previously reported is insufficiently specified. Conversely, if the success of the original model is due to an artefact or randomly occurring noise then this indicates that in models of this type it is critical to present results from multiple trained models, rather than from just one, to establish whether behaviours are a necessary consequence of the model or merely one of several possible outcomes.

Acknowledgements

We are grateful to Eddy J. Davelaar for constructive input during the modelling work presented here.

References

Collins, A., & Quillian, M. (1969). Retrieval time from semantic memory. *Journal of verbal learning and verbal behavior*, *8*(2), 240–247.

Lambon Ralph, M., Lowe, C., & Rogers, 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. L. (2011). Explorations in Parallel Distributed Processing: A Handbook of Models, Programs, and Exercises [Computer software manual].

McRae, K., & Cree, G. (2002). Factors underlying category-specific semantic deficits. *Category specificity in brain and mind*, 211–249.

Rogers, T., Lambon Ralph, M., Garrard, P., Bozeat, S., McClelland, J., Hodges, J., et al. (2004). Structure and deterioration of semantic memory: a neuropsychological and computational investigation. *Psychological Review*, *111*(1), 205–234.

Warrington, E. (1975). The selective impairment of semantic memory. *Quartlerly Journal of Experimental Psychology*, *27*(4), 635–657.

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