# The Emergence of Semantic Categories from Distributed Featural Representations

**Michael J. Greer *(mgreer@csl.psychol.cam.ac.uk)**
**Maarten van Casteren ^ (maarten.van-casteren@mrc-cbu.cam.ac.uk)**
**Stuart A. McLellan *(sam26@cam.ac.uk)**
**Helen E. Moss *(hem10@cam.ac.uk)**
**Jennifer Rodd * (jrodd@csl.psychol.cam.ac.uk)**
**Timothy T. Rogers ^ (tim.rogers@mrc-cbu.cam.ac.uk)**
**Lorraine K. Tyler * (lktyler@csl.psychol.cam.ac.uk)**

*Centre for Speech and Language, Department of Experimental Psychology, University of Cambridge
^MRC Cognition and Brain Sciences Unit, Cambridge, UK

**Abstract**

This paper presents a computational model of semantic memory, trained with behaviourally inspired vectors. The results are consistent with the *conceptual structure account* (Tyler, Moss, Durrant-Peatfield & Levy, 2000), which claims that concepts can be understood, and the effects of random damage predicted, based on (i) the number of correlations its features make, and (ii) the distinctiveness of those correlated features; the former indicating category membership, the latter distinguishing between concepts. The model shows a changing direction of domain-specific deficits as damage accumulates (animals concepts lost first, then objects upon severe lesioning). Also, the pattern of error differs between domains; animals tend to be confused for other members of the same category, whilst object errors disperse more widely across categories and domain. Recent neuropsychological evidence demonstrates a similar pattern for semantically impaired patients. For both patients and the model, this can be attributed to the timing of featural loss: distinctive features are lost earlier than shared features. The model demonstrates that the relative timing of feature loss differs between domains, resulting in the emergence of domain-specific effects.

## Introduction

The neuropsychological literature on semantic memory shows patients can develop an impairment in one domain of knowledge, whilst the other is relatively spared. Most commonly, semantically impaired patients show a deficit for living things (e.g. Warrington & Shallice, 1984), with the reverse pattern being rarer (e.g. Hillis and Caramazza, 1991).

There are three main types of explanation for the double dissociation. One postulates physically separate and functionally independent stores in the brain for dissociable categories of knowledge (e.g. Goodglass, Klein, Carey and Jones, 1966; Carramazza & Shelton, 1998). Another suggestion is that concepts may vary by domain according to the type of semantic information upon which they depend, with living things depending more on sensory information and artefacts depending more on functional properties (Warrington & Shallice, 1984; Warrington & McCarthy, 1983; 1987). Selective brain damage to one type of semantic information will lead to a category-specific deficit. This account assumes neuro-anatomical specialisation for type of property rather than category per se, to permit their independent disruption by brain damage. Finally, and most recently, attempts to account for category-specific deficits suggest that they can emerge from the *internal structure of concepts* alone without any type of neural or functional specialisation. Computational models have shown that random damage to a unitary, distributed system can produce category-specific deficits (e.g. Devlin, Gonnerman, Anderson and Seidenberg, 1998; Tyler et al, 2000). These models draw on structural aspects such as property correlation and distinctiveness.

## Conceptual Structure Account

Common to all distributed accounts of semantic memory (see McRae, de Sa, Seidenberg, 1997; Devlin et al, 1998; Tyler et al, 2000) is the observation that similar concepts tend to have overlapping sets of semantic features. Properties that frequently co-occur in concepts will serve to predict each others presence, a fact that a distributed connectionist network will exploit during training, leading to mutual activation of those properties. A consequence of mutual activation is resilience to damage of those properties, and hence their continued 'availability' to a stricken network when identifying concepts. A second important factor, is the distinctiveness[1] of features (cf. Devlin et al, 1998). A feature that is present in only one concept can be used

[1] Distinctiveness is calculated as 1/number of concepts for which the property is given.

to discriminate that concept from all others. As a feature occurs in an increasing number of concepts it becomes a progressively poorer marker for each of those concepts.

The *conceptual structure account* of semantic memory (Durrant-Peatfield, Tyler, Moss, & Levy, 1997; Tyler et al 2000) recognises that these factors – correlation and distinctiveness – will interact to determine which features will survive random damage, and the usefulness of the remaining features in preventing concept loss. By their very nature correlation and distinctiveness tend to be inversely related with each other. Highly correlated properties are often present in many concepts, and hence are not very distinctive. Thus, they will be robust to damage, but their preservation will be more useful for identifying the category to which an item belongs rather than distinguishing it from other category members. However, those distinctive properties that do correlate with other properties (especially other distinctive properties) will protect the concept in which they are found. Distinctive properties that fail to make strong correlations with other properties will be very vulnerable to damage. Domain differences and dissociations arise because concepts in different domains differ in these respects. We theorize that living things concepts have many intercorrelated properties, compared to artefacts, but these tend to be less distinctive. As a consequence, artefact concepts are more robust at all but severe levels of damage when only highly correlated properties remain intact.

## Computational Model

Previous work instantiating the conceptual structure account of semantic memory (Tyler et al, 2000) used 16 vectors that incorporated the theoretical assumptions of the account. In the current model the vectors are designed to broadly reflect the observed differences between living and non-living domains, as found in a large-scale property generation study (Moss, Tyler & Devlin, In Press). The simplified vectors, homogenous within domain and of equal number between domains, ensure the model's results are readily interpretable. In addition, the model was scaled up, and trained on 96 vectors. Consequently, the training set is as sparse as the property norm data which it resembles. One might expect a distributed model to perform differently as the training set becomes more sparse, as a sensible error-reduction strategy would be to turn all units off. This model sought to confirm that a distributed model would still build internal representations reflecting correlational structure in spite of extreme sparsity.

## Property norm data

Tables 1 and 2 report the global and distributional statistics of the property norm concepts. Data is also given for the model vectors, designed to resemble the property norm concepts as far as practicable.

Table 1: Global properties of the property norm set and the model vectors

| | **Property norms** | **Model vectors** |
|---|---|---|
| Number of concepts | 93 | 96 |
| Features that are highly distinctive[2] | 78% | 78% |
| Sparsity[3] | 3.7% | 4.6% |

Table 2: Characteristics of property norm concepts across domains (figures for model vectors in brackets)

| | **Living things** | **Artefacts** |
|---|---|---|
| Mean no. properties/concept | 17.7 (20) | 11.3 (14) |
| Mean distinctiveness of properties | 0.64 (0.22) | 0.73 (0.32) |
| No. of shared properties/concept[4] | 13.7 (15) | 7.5 (6) |

Following McRae et al (1997) the Pearson product moment correlation was computed for all pairs of semantic features. For the property norms, of the 78,210 possible correlations, only 2332 scored $|r|>0.3$. Living things had more correlated property pairs (CPPs) than artefacts (4070 vs. 1612), but artefacts had proportionally more CPPs occurring between distinctive features (20.0% artefacts vs. 11.7% living things).

## Representations

The training set consisted of 96 vectors, divided into 2 domains (Animals and Objects) and 4 categories (labelled somewhat arbitrarily as Land animals, Birds, Tools and Furniture). The vectors embodied the facts outlined below:

- There were 48 Animal vectors (24 Land animals and 24 Birds).
- Each Animal vector turned on 20/368 features.
- Every Animal vector turned on 10 'Animal shared' features.

[2] The proportion of features in the set shared by just 1 or 2 concepts.

[3] Sparsity refers to the average proportion of features turned on for each vector.

[4] A shared property being defined as one held by three or more concepts, otherwise the property is distinctive.

- Land animals were distinguishable from Birds by which group of 5 shared features was turned on – 'Land shared' or 'Bird shared'.
- All animals could be distinguished from each other by which three 'Animal distinctive features' were on.
- There were 48 cross-domain features, each concept turning on two. Each concept, whether animal or object turned on a unique combination of cross-domain features (say 1 and 4; or 2 and 5 etc) such that each unit was turned on by 4 different concepts (2 animal, 2 objects). This means that having a cross-domain feature on does not predict at all which domain's concept is on, but limits the concept to one of four possibilities.
- There were 48 object concepts (24 tools and 24 furniture).
- Every Object vector turned on 14/368 features.
- Tools and Furniture were distinguishable by which group of 6 shared features they turned on.
- Object concepts were identifiable by which 2 'object distinctive triplets' were on. They are termed triplets because, within a triplet, if one feature is on then the other two must also be on (likewise when off). However, each "object distinctive triplet" is turned on both by 1 tool concept and 1 furniture concept, so having a triplet on does not perfectly predict which object is on (in contrast to "animal distinctive features").

The resultant vectors resemble the concepts analysed in the property generation study (see tables 1 and 2). The resemblance extends to the vectors' correlational structure. For the model vectors, of the 67,528 possible correlations, 1864 scored $|r|>0.3$. Animals had more correlated property pairs (CPPs) than objects (5472 vs. 2016), but objects had proportionally more CPPs occurring between distinctive features (35.7% objects vs. 2.6% animals). This reproduces the pattern of the domain effects in the property norms, but exaggerates the size of the difference.

## Architecture and training

The network consisted of three layers, a semantic input layer, a hidden layer and a semantic output layer, as shown in Figure 1. During training, with the back-propagation learning algorithm (Rumelhart, Hinton & McClelland, 1986), the network was required to reproduce the input on the output layer. 10 networks were trained with different initial random weights (±0.005), with a learning rate of 0.25 and momentum 0.5. Training was stopped when the squared error for each feature in every vector was below 0.01, occurring after a mean of 193 presentations of the complete target set.

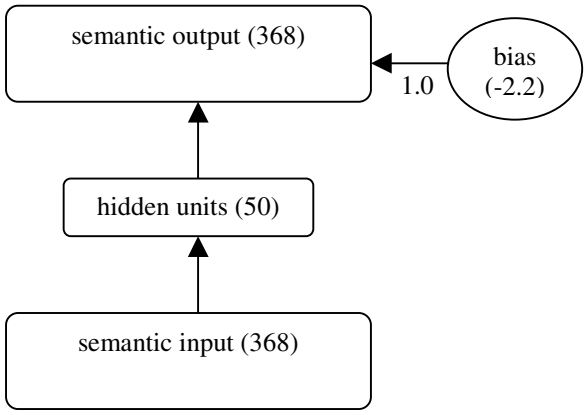


Figure 1: Model architecture: the numbers in each box indicate the number of units in that layer, while arrows indicate full connectivity between layers[5]

### Lesioning

Brain damage is simulated in this model by random deletion of semantic connections (by setting weights to 0). Initially 10% of weights were cut, then the model's performance analyzed. The proportion of damaged connections was increased by increments of 10% until all inter-layer connections were set to 0. This lesioning process was carried out 5 times on each of the 10 trained networks to produce a total of 50 networks.

## Testing

Network performance was analyzed both at the level of features and concepts. The training set was presented to the network's input layer and the pattern of activations on the output layer examined. We predict that highly shared features will be more resilient to damage than distinctive features (i.e. will still activate when they should). In the model vectors this will correspond to the greatest advantage being for 'animal shared features', then 'land', 'bird', 'tool' and 'furniture' shared features behaving similarly, with 'animal distinctive' features being least preserved. A different pattern is predicted at the conceptual level where discrimination is dependent primarily upon distinctive features. Object distinctive features cluster into 'triplets', this additional inter-correlation is predicted to enhance their robustness to damage relative to Animal distinctive features, leading to an advantage in concept identification. Only at severe levels of damage, when all distinctive features are lost to the network, will the advantage for animal shared features translate to an advantage in concept naming.

[5] Bias ensures the 100% damaged model outputs 0s (approximately); in its absence the semi-linear logistic activation function makes every semantic output unit 0.5. Bias connections were not lesioned.

## Featural analysis

For each vector, the activation of the semantic output layer was binarised – unit values <0.5 were scored 0 while values ≥0.5 were scored 1. Each unit value was compared to that of the input vector and declared correct or error. Attention focused on the subset of units that were supposed to be on for each vector, and the number of errors summed for each domain. Because each output unit represents a local feature it is possible to compare the errors across the different feature types (i.e. 'animal shared', 'object distinctive triplet' etc).

## Overall analysis

The pattern of activity over the output units was compared to all 96 vectors; both had been normalised to remove effects of concept size[6]. The closest match, by Euclidean distance, was considered the model's response. Upon network lesioning, errors occur of three types: within-category error, cross-category error, or a cross-domain error.

# Results

Figure 2 presents the results of the featural analysis averaged over 50 simulations. The general pattern is for a greater impairment of unique relative to shared features at all levels of lesioning between 20 and 80%.

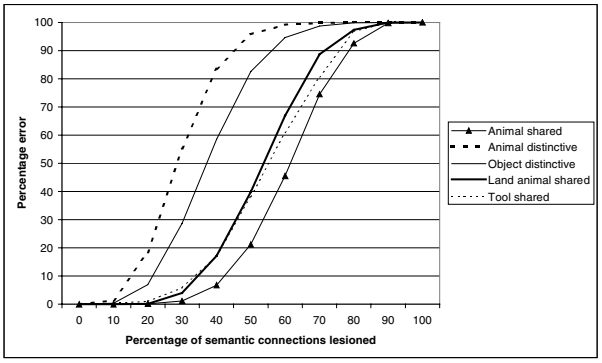


Figure 2:Featural error (failure to activate) as a function of network damage.

The effects of damage upon concept identification are shown in Figure 3, which shows an advantage for objects over animals up until 80% of connections had been lesioned. A two-way ANOVA (domain*damage) showed a main effect of domain (i.e. animals vs. objects, (F[1,539]=317, p<.0001), and a significant domain by damage interaction (F[10,539]= 5406, p<.0001).

A repeated measures t-test on the 90% damaged data-points showed that the advantage for animals, although small, was significant (t= -6.249, p<.0001).

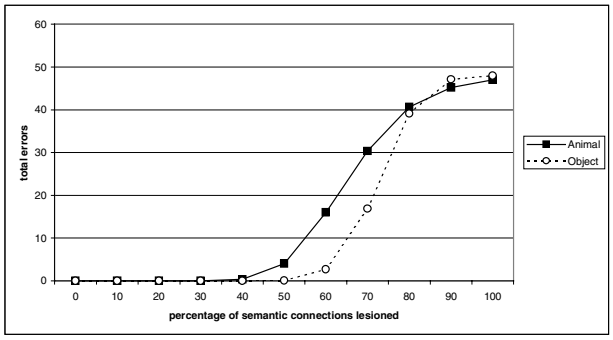


Figure 3: Identity mapping as a function of damage.

Figure 4 shows the difference in the distribution of error types when concepts were mis-identified. It attributes the early animal deficit to within-category error, with all errors involving members of the same category. Conversely, object errors are more widely dispersed between the two domains.

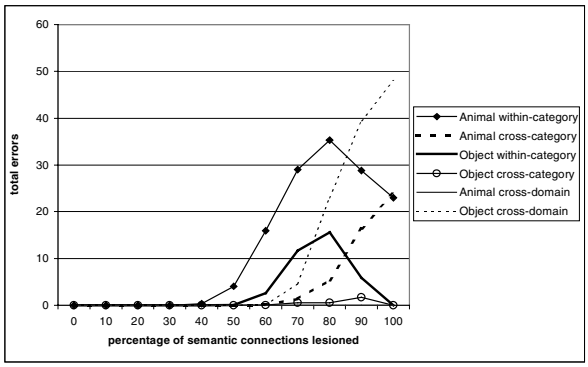


Figure 4: Error types as a function of damage.

# Discussion

This research demonstrates how category and domain-specific deficits can arise following damage to a single distributed semantic system without explicit category structure. Further, it accounts for the patterns of impairment observed in patients as resulting from a complex interaction of correlations between features and the extent to which features are shared or distinctive. In general, the behaviour of the model is very similar to the behaviour of some brain-damaged patients with category/domain specific deficits.

The preservation of individual features was dependent on the number of correlations each feature entered into, with more highly shared, correlated features being more robust. This pattern is similar to that observed in semantically-impaired patients who show better preserved knowledge of shared, category-

[6] Concept size refers to the number of features turned on.

defining information compared to distinctive, concept-identifying information (Moss, Tyler, Durrant-Peatfield, & Bunn, 1998; Moss et al, In Press).

The preservation of individual concepts showed a different pattern. Global damage, where connections between layers were randomly lesioned, produced an initial impairment for animals, followed at severe lesioning, by impairment for objects. Successfully identifying a concept relies most heavily on activating its distinctive features. With damage, object distinctive properties were more robust than animals, which can be attributed to their tendency to correlate with other distinctive properties. Crucially, this same pattern of correlations occurred in the empirically derived property norms (Moss et al, In Press). This shows that the same factor that accounts for an early animal deficit in the computational model could also account for the initial living-things deficit found in at least some semantic dementia patients (Moss & Tyler, 2000; Moss et al, In Press). Beyond 70% lesioning all distinctive features failed to activate, whether animal or object. Consequently, the model had to 'make a guess', though the odds of guessing correctly would have differed for the two domains. The animal shared features are both more numerous and more correlated, hence will remain more likely to be available to the network. Therefore the model would have been guessing from a smaller subset of possible concepts than would have been the case for objects, which could lead to a mild object deficit. This unequal distribution of shared features between domains is also characteristic of the property norm data.

The lesioning data shows that the magnitude of the early disadvantage for naming animals exceeds that of the late disadvantage for naming objects. This too seems to be reflected in semantic dementia patients where living-thing deficits tend to be both greater in size and more numerous than corresponding artefact deficits (Moss & Tyler, 2000).

The conceptual structure account, realised in the computational model, also predicts the type of error likely to be made when concepts are mis-identified. Due both to the robustness of animal shared features, and the vulnerability to damage of their distinctive features, animals will most commonly be confused with other members of the same category. Cross-domain errors should hardly ever occur. Whilst the same should be true of objects, this tendency will be less marked, and errors will be dispersed more widely between the types of error possible: within-category; cross-category; and cross-domain. There is some evidence for this pattern in longitudinal studies of picture naming. Hodges, Graham & Patterson (1995) report a semantic dementia patient, JL. For living things, he made progressively more within-category and superordinate errors, but never produced a cross-category or cross-domain error. Similarly for the progressive aphasic patient AA (Moss et al, 1999), tested on four occasions over two years, but failing to produce a living things cross-category mistake until the final testing session. For artefacts, in contrast, she occasionally made cross-category and cross-domain errors throughout the testing period.

## Epilogue: The problem of determining error

In common with other models of semantic memory, the network's output was compared to every vector in the training set, and the vector with which the normalised Euclidean distance was smallest, regarded as the network response. As a result the network was forced to make a response, irrespective of the meaningfulness or otherwise of the output. A potential limitation of this procedure is that the network could not respond "don't know", a response commonly produced by semantically impaired patients in semantic tasks. As a first step to simulating a "don't know" response a threshold for normalised Euclidean distance was introduced; if the error between the output and every vector exceeded this threshold then the output was scored incorrect. The problem then was to decide how strict the threshold ought to be. The result of some early explorations is shown in figure 5.

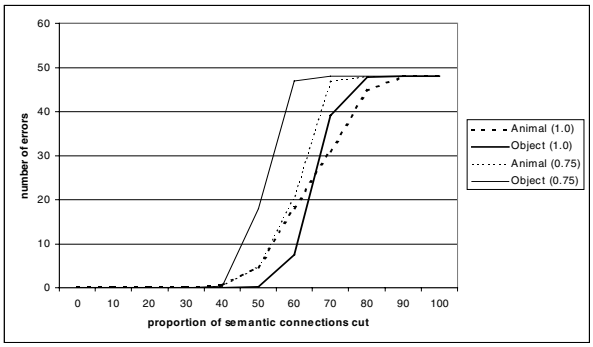


Figure 5: The variability of domain effects with different thresholds for normalised Euclidean distance.

A threshold value of 1.0 produced a strong cross-over, an early animal deficit progressing to an object deficit with damage. Reducing the threshold value to 0.75 had a catastrophic effect upon object identification, but animal identification was less impaired. As a result, the graph showed a consistent deficit in identifying objects. A threshold figure of 1.5 produced a graph identical to that when no threshold was applied (i.e. the same as figure 3). The sensitivity of object identification to varying threshold values probably reflects the smaller number of inter-correlations between object features. The representation of an object concept in semantic space will be sparser, so when the representation is damaged, and a strict threshold is applied, it will be unlikely to fall into a neighbouring concept's space.

Instead the output will be scored "don't know", an outcome which is much less likely in the denser animal semantic space.

The problem here is how we determine where the error threshold should lie? One approach might be to record the number of "don't know" responses the model makes, and relate this to patient data. This is complicated by the difficulty of relating the degree of brain damage to particular levels of network lesioning. Also, patient performance on tests of semantic knowledge varies with the demands of the task. For example, "Don't know" is a more common response for picture naming than word-picture matching.

Speculatively, the normalised Euclidean threshold could reflect a compromise position in a speed/accuracy trade off. Tasks that demand a rapid response would have a more relaxed threshold than those where time is given for a considered response.

## Conclusion

This model suggests that the data inherent in conceptual structure is sufficient to account for the domain-specific effects observed in semantically impaired patients. Categories and domains emerge when concepts are represented in a single, distributed system. Some recent neuro-imaging studies fail to show regional differences in activation for different conceptual domains (e.g. Devlin, Russell, Davis, Price, Moss, Matthews, & Tyler, 2000), consistent with the neural substrate of concepts being organised in a distributed fashion.

## Acknowledgements

Grants from the McDonnell-Pew Foundation, Medical Research Council (UK) and the Wellcome Trust supported this research.

## References

Carramazza, A.C., & Shelton, J. R. (1998) Domain-specific knowledge systems in the brain: The animate-inanimate distinction. *Journal of Cognitive Neuroscience*, 10, 1-35.

Durrant-Peatfield, M., Tyler, L.K., Moss, H. E. & Levy, J. (1997) The distinctiveness of form and function in category structure: A connectionist model. In: M.G. Shafto & P. Langley (Eds.), *Proceedings of the Nineteenth Annual Conference of the Cognitive Science Society*, Stanford University, Mahwah, NJ: Erlbaum.

Devlin, J., Gonnerman, L., Anderson, E., and Seidenberg, M. (1998). Category specific semantic deficits in focal and widespread brain damage: A computational account. *Journal of Cognitive Neuroscience*, 10, 77-94.

Devlin, J.T., Russell, R.P., Davis, M.H., Price, C.J., Moss, H.E., Matthews, P., & Tyler, L.K. (2000) Susceptibility-induced loss of signal: Comparing PET and fMRI on a semantic task. *NeuroImage*, 11, 589-600, 2000

Goodglass, H., Klein, B., Carey, P., & Jones, K. (1966). Specific semantic word categories in aphasia. *Cortex*, 2, 74-89.

Hillis, A. E., & Caramazza, A. C. (1991). Category-specific naming and comprehension impairment: A double dissociation. *Brain & Language*, 114, 2081-2094.

Hodges, J., Graham, N. & Patterson, K. (1995) Charting the progression in semantic dementia: Implications for the organisation of semantic memory. *Memory*, 3, 463-495.

McRae, K., de Sa, V., & Seidenberg, M. (1997). On the nature and scope of featural representations of word meaning. *Journal of Experimental Psychology: General*, 126, 99-130.

Moss, H. E., & Tyler, L. K. (2000) A progressive category-specific deficit for non-living things. *Neuropsycholgia, 38, 60-82.*

Moss, H. E., & Tyler, L. K. (1997) A category-specific semantic deficit for non-living things in a case of progressive aphasia. *Brain and Language*, 60, 55-58.

Moss, H. E., Tyler, L. K., & Devlin, J. (In Press) The emergence of category-specific deficits in a distributed semantic system. In E. M. E. Forde and G. W. Humphreys (Eds.) *Category-specificity in mind and brain*.

Moss, H. E., Tyler, L. K., Durrant-Peatfield, M., & Bunn, E. (1998) "Two eyes of a see-through"; Impaired and intact semantic knowledge in a case of a selective deficit for living things. Neurocase, 4, 291-310.

Rumelhart, D. E., Hinton, G. E. & McClelland, J. L. (1986) A general framework for parallel distributed processing. In D. E. Rumelhart and J. L. McClelland (Eds.) *Parallel distributed processing: Explorations in the microstructure of cognition*. Cambridge, MA: MIT Press.

Tyler, L.K., Moss, H. E., Durrant-Peatfield, M., & Levy, J. (2000). Conceptual structure and the structure of concepts: A distributed account of category-specific deficits. *Brain & Language, 75, 195-231.*

Warrington, E. K., & McCarthy, R. (1983). Category specific access dysphasia. *Brain*, 106, 859-78.

Warrington, E. K., & McCarthy, R. (1987). Categories of knowledge: further fractionations and an attempted integration. *Brain*, 110, 1273-96.

Warrington, E. K., & Shallice, T. (1984). Category specific impairments. *Brain*, 107, 829-54.