Is the Mystery of Thought Demystified by Context–Dependent Categorisation? Towards a New Relation Between Language and Thought

MICHAEL S. C. THOMAS, HARRY R. M. PURSER AND DENIS MARESCHAL

Abstract: We argue that are no such things as literal categories in human cognition. Instead, we argue that there are merely temporary coalescences of dimensions of similarity, which are brought together by context in order to create the similarity structure in mental representations appropriate for the task at hand. Fodor (2000) contends that context-sensitive cognition cannot be realised by current computational theories of mind. We address this challenge by describing a simple computational implementation that exhibits internal knowledge representations whose similarity structure alters fluidly depending on context. We explicate the processing properties that support this function and illustrate with two more complex models, one applied to the development of semantic knowledge (Rogers and McClelland, 2004), the second to the processing of simple metaphorical comparisons (Thomas and Mareschal, 2001). The models firstly demonstrate how phenomena that seem problematic for literal categorisation (such as the 'non-literal' comparisons involved in metaphor and analogy) resolve to particular cases of the contextual modulation of mental representations; and secondly prompt a new perspective on the relation between language and thought: language affords the strategic control of context on semantic knowledge, allowing information to be brought to bear in a given situation that might otherwise not be available to influence processing. This may explain one way in which human thought is creative, and distinctive from animal cognition.

1. Introduction

In this article, we pursue the thesis that there are no such things as literal categories in human cognition. Instead, we argue that there are merely temporary coalescences of dimensions of similarity, which are brought together by context in order to create the similarity structure in mental representations appropriate for the task at hand. Phenomena that seem problematic for literal categorisation (such as the 'non-literal' comparisons involved in metaphor and analogy) resolve to particular cases of the contextual modulation of mental representations. In the following, we review proposals that human cognition, and particularly categorisation, is intrinsically context sensitive. This leads us to an obstacle—the contention, advanced by Fodor (2000), that context-sensitive cognition cannot be realised by current computational

theories of mind. We address this challenge by describing a simple computational implementation that exhibits internal knowledge representations whose similarity structure alters fluidly depending on context, and explicate the processing properties that support this function. To exhibit the utility of this processing architecture, we then review two more complex examples of the same idea, one model applied to the development of semantic knowledge (Rogers and McClelland, 2004), the second applied to the processing of simple metaphorical comparisons (Thomas and Mareschal, 2001). Finally, based on the latter model, we speculate on a new way to conceive the relationship between language and thought that stems directly from the premise of context-sensitive representations and the manipulation of non-literal similarity through language.

Much of cognitive science is premised on the assumption that the discovery of literal categories (i.e. pre-existing sets of entities in the world) is crucial to human cognition (e.g. Murphy, 2003). However, there are two sorts of problems with the view that literal categories are fundamental to human cognition. First, philosophically it has proved notoriously hard to define literal categories in terms of necessary and sufficient features. This has left the concept resting on the shaky foundations of notions like 'family resemblance' or tied up in entrenched debates on whether exemplars or prototypes are more fundamental to human categories. Second, intrinsic to the idea of literal categories is *literal similarity*. But high-level human cognition is also characterised by the use of non-literal similarity, exemplified by metaphor and analogy, which rely on the non-literal similarity between categories. If there is a basic divide between literal and figurative similarity, the production and comprehension of metaphorical and analogical comparisons would seem to require additional, special cognitive mechanisms; but little psychological evidence has accrued for the existence of such mechanisms (Glucksberg, 2000).

Turning first to the problem of categorisation, it has been argued that one of the hallmarks of human categories is that they do not have to exist prior to the situation of their usage: novel categories can be created on the fly (Barsalou, 1983, 1993; Chalmers *et al.*, 1992). If novel categories can be created on the fly, perhaps all categories function this way. Other researchers have argued that human categorisation behaviour is often driven by *partial representations*, so that only some dimensions of knowledge are activated by a given situation, and different aspects of a category are activated by different situations (Mareschal *et al.*, 2007; Sirois *et al.*, 2008). Both these positions are consistent with the view that human categories are intrinsically context dependent.

Contextual effects on semantic knowledge are well established. Barsalou (1993) noted that when participants were asked to provide definitions for categories, such as *bird*, on average, more features differed across two participants' definitions than were shared by those participants, suggesting that considerable representational flexibility exists between individuals. However, there may also be marked flexibility within individual participants: if different supporting contexts are provided for the same category label, the prototypicality (or representativeness) of particular exemplars may differ wildly (e.g. Glucksberg and Estes, 2000; Murphy, 1988; Roth and

Shoben, 1983). For example, from the imagined perspective of a Chinese person, *swan* and *peacock* may be highly representative of the category of birds, whereas from the perspective of an American, *robin* and *eagle* may be prototypical. Such flexibility cannot reflect differences in underlying knowledge, because the same participants were involved in each context. Even across participants, the knowledge base may be quite uniform: Barsalou (1993) reported that when all the features produced by participants for *bird* were pooled and presented to a new group of participants, and that new group asked to judge whether each feature was potentially true of birds, the agreement across the groups was near-perfect. Thus, differences in features listed for a definition and differences in prototypicality of category exemplars do not owe primarily to differences in knowledge, but to those of context.

The idea that categories are context specific is certainly not new: William James, in *The Principles of Psychology* chapter on 'Reasoning' (James, 1890/1999), articulated the view that categories are goal-directed and context-specific: 'Now that I am writing, it is essential that I conceive my paper as a surface for inscription … But if I wished to light a fire, and no other materials were by, the essential way of conceiving the paper would be as combustible material' (p. 959). Wittgenstein, too, in his *Investigations*, states that 'how we group words into kinds depends on the aim of our classification—and on our own inclination' (Wittgenstein, 1953). Wittgenstein famously demonstrated the difficulties of defining the word 'game', not to show that to do so is impossible, but to point out that a rigid definition is not necessary for people to use the term successfully—people clearly *do* use and comprehend the word without apparent difficulty.

If categories are specific to contexts, then what is it that binds categories together? Quine (1977) made the point that simply invoking similarity as mental glue raises the very problem that it is intended to answer: things may seem similar simply *because* they belong to the same category. Murphy and Medin (1985) have criticised the prevalent focus on similarity and the associated tendency to break down concepts into constituent attributes or components, noting that such practice ignores human goals, needs and theories. An alternative account, then, is that categories with exemplars connected by structure-function relationships, or by causal schemata of some kind, will be more coherent than categories with exemplars that are not.

One might indeed have a goal-state that could connect (to some degree) objects that appear to share very few features: Barsalou (1983) investigated the properties of *ad hoc* categories, which are presumed to be formed 'on the fly' rather than retrieved from long-term memory. Two examples are 'ways to escape being killed by the Mafia' and 'vegetarian dishes to accompany *melanzane alla parmigiana*'. For *ad hoc* categories, typicality cannot be determined by similarity to a category concept, but must be driven by dimensions relevant to the goal that the category serves. Even so, *ad hoc* categories were found to show typicality gradients (i.e. some exemplars being more representative than others) as salient as those associated with 'common' categories (such as *fruit* or *birds*): *ad hoc* categories varied as much in typicality as common categories, and participants showed similar levels of agreement in typicality judgements of exemplars from each. These findings are consistent with the notion

that the same kind of mental processes underlie both *ad hoc* and common categories, with each being context-dependent and fluid.

Fodor (2000) has endorsed the view that high-level human cognition (or 'thinking') is characterised by *context sensitivity* and *globality*. Although Fodor argues that low-level sensory and motor functions are subserved by modular systems, the 'central system' is conceived as having access to the entirety of an individual's knowledge, in order that it might guide behaviour; Fodor refers to this complete access as *globality*. *Centrality* and *simplicity* are viewed as illustrative of *globality*. *Centrality* refers to the notion that an individual item of information may be central to one idea but peripheral to another (e.g. 'unmarried' for bachelors and pizza, respectively, cf. Barsalou, 1983); the *centrality* of the information is context-specific inasmuch as it is dependent on the particular idea under consideration and is therefore not intrinsic to the item of information itself. *Simplicity* refers to the idea that two different explanations drawn from the same set of representations may differ in their degree of complexity (e.g. '2+2−1=3' versus '2+3+3+1−(1+3+2)=3'); *simplicity* is a property at the level of the explanation, but not of the constituent representations. Importantly for our purposes, Fodor (2000) has expressed scepticism that current computational theories of mind are sufficient to explain the context sensitivity of human thought. His concern is that for both symbolic and connectionist approaches to cognition—two of the leading computational theories—the causal properties of reasoning systems are driven by local rather than global properties of representations. For symbolic systems, the local property is the syntactic structure of representations. For connectionist systems, the local property is the connectivity matrix of the neural network.

Fodor's comments make context-sensitivity appear a mysterious property of the mind, under current conceptions of mental processes. How could context-sensitivity operate in real representational systems with fixed causal structures? How could context sensitivity emerge within cognitive development? It is difficult to address such questions unless they can be precisely formulated. The contribution of the current article is to show how it can be done: context-sensitivity may not be particularly mysterious, but rather straightforward. This demonstration will utilise computational (and specifically connectionist) modelling: we will begin by describing a simple five-unit neural network in which the similarity structure of the internal representations is altered by context. This model will serve as a 'teaching example' to plainly demonstrate the computational mechanism we propose for context-sensitivity. We will then illustrate this mechanism in two more complex models of human cognition, first to show how semantic memory can demonstrate context dependence and then, turning to the issue of non-literal similarity, to argue that metaphor can be viewed as merely a variety of contextual modulation. Computational models have proved useful to cognitive science because they demonstrate how complex theoretical notions can work in practice. For example, Oakes, Newcombe and Plumert (2009) have argued that modelling has made a significant contribution to advancing our understanding of the concepts of interaction and emergence, even though these ideas were already present in the theories of Piaget, Gibson, and Vygotsky. In the same way, the intention here is to show that a simple computational architecture can
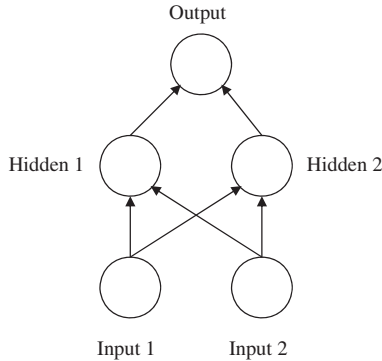
Output

Hidden 1   Hidden 2

Input 1   Input 2

Figure 1 *Exclusive-or (XOR) network.*

nevertheless show complex patterns of context-sensitive processing, and further, that this property enables similar models to account for high-level human behaviours.

2. A Mechanism for Contextual Modulation

The *exclusive-or (XOR)* logical problem was used in the early exploration of the computational properties of connectionist networks because solving it requires *internal representations* (Rumelhart, Hinton, and Williams, 1986). In neural network terms, this means that the problem cannot be solved by a network with direct connections between inputs and outputs, but requires an intermediate layer of 'hidden' processing units that develops a transitional representational state that half solves the mapping problem. The connectionist network traditionally used to learn the solution to the XOR problem has only five units: two input units, two hidden units, and one output unit (Figure 1). This network is both well known and simple, so it is ideally suited to introducing the computational principle of context-sensitivity that is the focus of this article.

The XOR problem is specified over two inputs and one output (see Table 1). Figure 2a shows the four input−output patterns comprising the problem represented in a two-dimensional 'input space', with the axes depicting, respectively, the value of input unit 1 and the value of input unit 2. The computational complexity arises because the output unit of a network can only make a single categorisation in input space, equivalent to drawing a 'decision line' through input space and responding positively to inputs falling on one side and negatively to units falling on the other. However, the two inputs that must be classified positively, namely [1,0] and [0,1], cannot be separated with a straight line from those to which it must respond negatively, [0,0] and [1,1]. Hence, the problem is termed 'linearly inseparable'. A network with a layer of hidden units can learn to re-represent the similarity structure of the problem over these hidden units, so that the problem becomes linearly separable for the output unit.

Pattern	Input 1	Input 2	Output
p1	0	0	0
p2	1	0	1
p3	0	1	1
p4	1	1	0

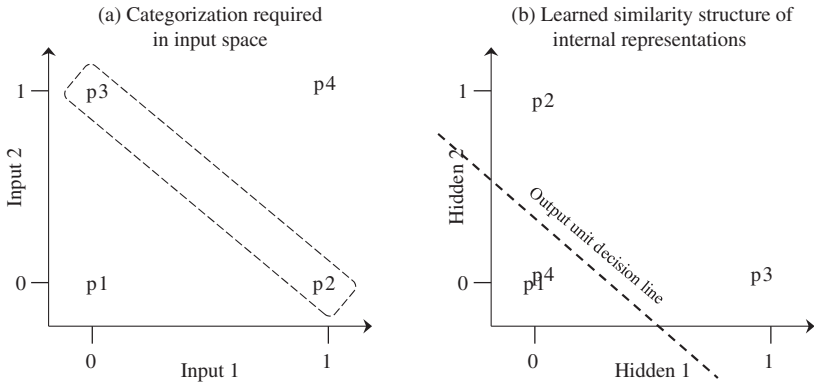**Table 1** *The XOR mapping problem.*

Figure 2 *Geometric representation of the XOR input space and a sample hidden unit space for a network that has learnt to solve the problem (p = pattern). (a) Input space: p2 and p3 must be categorised separately from p1 and p4, but this cannot be achieved by a single decision line. (b) In hidden unit space, the mapping problem is re-represented so that a single line can now achieve the categorisation at output.*

The following is an example of such a five-unit network that has learned the internal representations necessary to solve the XOR problem. It demonstrates how the similarity structure of the internal representations develops to solve the categorisation problem. The network was trained for 2500 presentations of the complete training set, using the back-propagation learning algorithm. The learning rate and momentum were set to 0.1 and 0.0, respectively. In the same way that the two input units specify two dimensions of input space, the two hidden units specify two dimensions of hidden unit space. Figure 2b shows how the similarity structure of the input space has been re-represented in hidden unit space, for one sample run of the XOR network. The figure includes the decision line employed by the output unit, which is determined by its threshold and the two weights connecting the hidden units to the output unit. It is evident that the patterns [1,0] and [0,1] now fall on one side of the decision line and [0,0] and [1,1] fall on the other, and therefore that the required categorisation can now be achieved.

We now introduce a new problem whose solution requires contextual modulation of the internal similarity structure. The *Hexagon* problem shown in Table 2

Pattern	Input 1	Input 2	Output	
			Context A	Context B
p1	.25	0	0	0
p2	.75	0	1	0
p3	1	.5	0	1
p4	.75	1	0	0
p5	.25	1	0	1
p6	0	.5	1	0

Table 2 *The Hexagon mapping problem.*

is a slightly modified version of the XOR problem. There are now six input patterns, in the shape of a hexagon in input space (Figure 3a and 3b). However, the network is now required to learn *two different categorisations* of these input patterns, depending on the context. The categorisations are both linearly inseparable and are partly mutually exclusive (that is, two of the input patterns that must be classified positively in one context must be classified negatively in the other and vice versa, while two must be classified negatively in both). The current context is provided to the network by two additional input units (Figure 4). These context units are identical in nature to the other inputs. (This raises the question of what differentiates context from input: we take up this theme in the discussion.)

Again, we can examine the similarity structure of internal representations that are developed when the network is trained on the Hexagon problem. The network was trained for 8000 presentations of the training set, with back–propagation, a learning rate of 0.1, and a momentum of 0.0. Figure 3c and 3d show the similarity structure of the internal representations under the two contexts, for a sample network. Both cases resemble the solution for the XOR network: the input space has been represented over the two hidden units in such a way that patterns to be classified positively lie on one side of the output unit's decision line, while those to be classified negatively lie on the other side. The crucial point to note here is that, although the decision line learned by the output unit is itself insensitive to context, *the similarity structure of the internal representations shifts dynamically underneath this line in a manner that depends on context*. Some patterns that fall on one side of the decision line in one context, fall on the other side of the decision line in the other context. Note that the network has a fixed architecture (the connection weights and thresholds are the same for each context). Recall that Fodor (2000) argues that it is the architecture that is the causally efficacious property of connectionist networks. How, then, does the network manage to alter the similarity structure of its internal representations depending on the context? Figure 5 shows sample solutions adopted by the XOR and Hexagon networks, in terms of their connection weights and unit thresholds. The principle of functioning is that a receiving unit sums the products
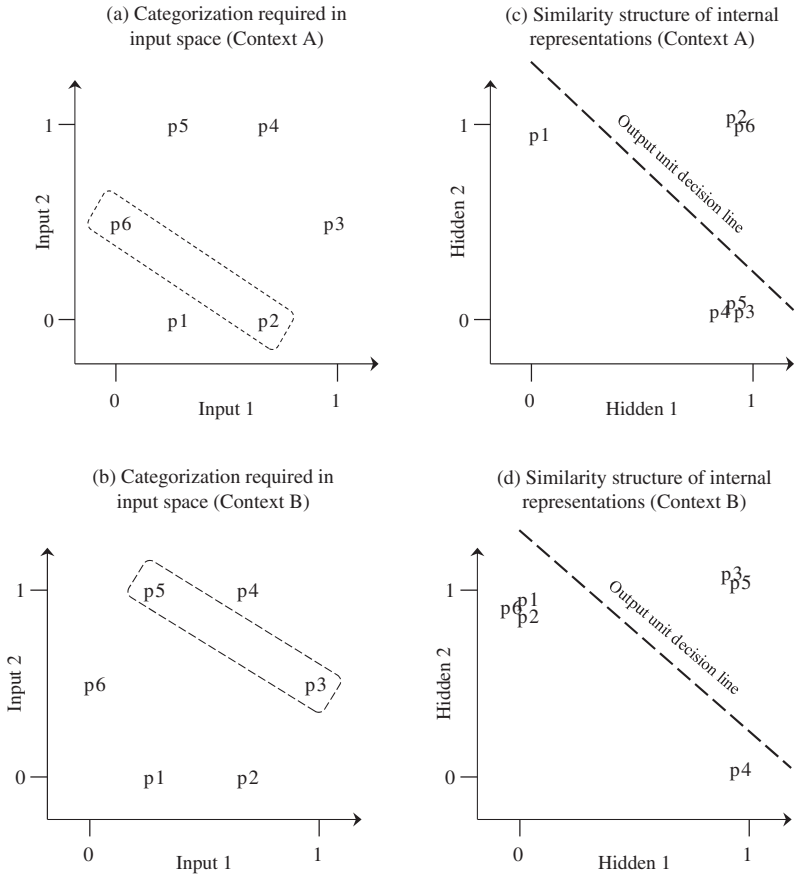
(a) Categorization required in
input space (Context A)

(c) Similarity structure of internal
representations (Context A)

(b) Categorization required in
input space (Context B)

(d) Similarity structure of internal
representations (Context B)

Figure 3 *Input and sample hidden unit spaces for the Hexagon network, for categorisations in two different contexts.*

of the activation of each sending unit with the connection strength between the two units; if the sum exceeds the receiving unit's threshold, it becomes active.

Note that in the Hexagon diagram in Figure 5, we have included the contribution of each context unit in terms of the *effective* thresholds that they produce in the respective hidden units. This means that if a context unit serves to excite a hidden unit, it is lowering the hidden unit's effective threshold, because less excitation is now necessary for the input units to push the hidden unit over its actual threshold. Conversely, if a context unit inhibits the hidden unit, it is raising the hidden unit's effective threshold. For example, if the *actual* threshold of a hidden unit is 5 (i.e. the value of the incoming activation that must be exceeded for the hidden unit to turn itself on), context unit A connects to this unit with a weight of −4, and context unit B connects to the unit with a weight of +1, then the effective threshold of the
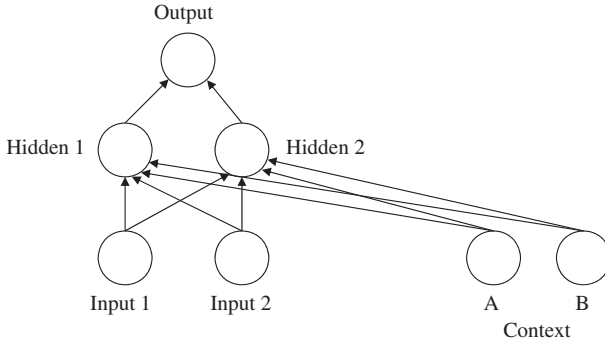
Output

Hidden 1   Hidden 2

Input 1   Input 2   A   B

Context

Figure 4 *Hexagon network.*

>6

12   12

Thresholds: >4   >4

-7   7   6   -7

Thresholds

>17

12   12

Context A: >4   >-9
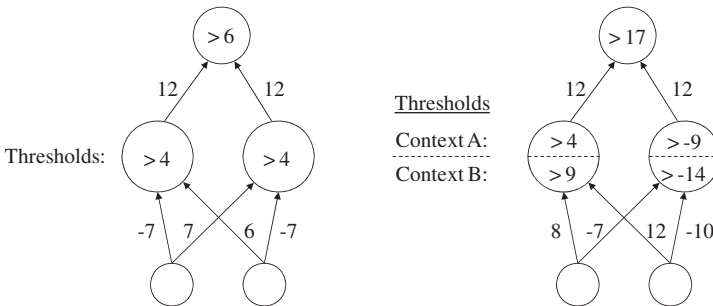Context B: >9   >-14

8   -7   12   -10

Figure 5 *Network solutions for XOR and Hexagon problems (numbers inside units show effective thresholds).*

hidden unit is [5 − (−4) = 9] in context A and [5 − (+1) = 4] in context B. In other words, input from A makes the hidden unit less likely to turn on, while input from B makes it more likely to turn on.

The notion of threshold used in this example is a slight simplification, since the activation of a processing unit in a typical connectionist network is, in fact, determined by passing the summed input through a smoother sigmoid activation function, rather than through a binary step function. Nevertheless, it should be clear here how context succeeds in modulating the similarity structure of the internal representations: it does so by producing different effective thresholds in the hidden units. The activation arriving from the input units is the same in each case, because the weights between inputs and hidden units are fixed. The decision line of the output unit is the same in each case because, again, its connections to the hidden units are fixed and it receives no direct input from the context units. In contrast, the computational properties of the internal representations with respect to the input are defined with respect to the activity of the context units.

Importantly, then, this simple model demonstrates that it is quite feasible for context to radically alter the similarity structure of internal representations—sufficient

for the output units to achieve different categorisations of the input. The fluidity of the internal representations occurs by virtue of the activation dynamics in the network, even though the weight matrix of the network is fixed. *Contra* Fodor, then, it is the permissible activation dynamics of a connectionist network that define its causal properties, not its connection weights alone (although it should be noted that the two are, of course, very closely linked). A context-invariant connectivity matrix can support context-sensitive internal representations because the activation dynamics are an emergent global property of the network. A straightforward demonstration of this point constitutes the central message of this article.

Of course, the example is very limited. How might this mechanism for producing context-sensitive categorisation be applied to more complex models that address aspects of high-level human cognition? In the next two sections, we discuss two such models. Both represent variants on a general architecture in which a contextual source of information is used to modulate the similarity structure of internal representations of semantic knowledge. The first example comes from the work of Rogers and McClelland (2004), investigating the development of semantic knowledge in children. The second is a model of the comprehension of simple metaphorical comparisons (Thomas and Mareschal, 2001). The second model is of particular relevance, here, because it addresses the idea of non-literal similarity, which is so problematic for approaches to cognition that rely on literal categories.

3. Two Models of Context–Dependent Categorisation in High–Level Cognition

3.1 The Development of Semantic Knowledge

Rogers and McClelland (2004) explored a model of the development of semantic knowledge. Extending initial work by Hinton (1981) and Rumelhart and Todd (1993), the authors construed semantic knowledge in terms of sets of propositions linking items and features (e.g. *a robin is a bird*, *a robin can fly*, *a robin has wings*). The architecture of the model is shown in Figure 6. The individual nodes in the network's input and output layers correspond to the constituents of these propositions: items (e.g. *pine*, *rose*, *robin*, *salmon*), relations (*IS A*, *is*, *can*, *has*), and attributes (e.g. *living thing*, *plant*, *animal*, *bird*, *red*, *grow*, *fly*, *wings*, *leaves*, *skin*). When presented with a particular pair of items and relations at input, the network attempts to switch on the attribute units in the output layer that correspond to valid completions of the proposition. For example, when the units corresponding to *salmon* and *can* are activated at input, the network must learn to activate the nodes that represent *grow*, *move* and *swim*. Although localist representations are used at the model's input and output, the learning process allows the model to derive distributed internal representations that do not have this atomic character and that instead exhibit similarity structure. The conceptual knowledge, stored
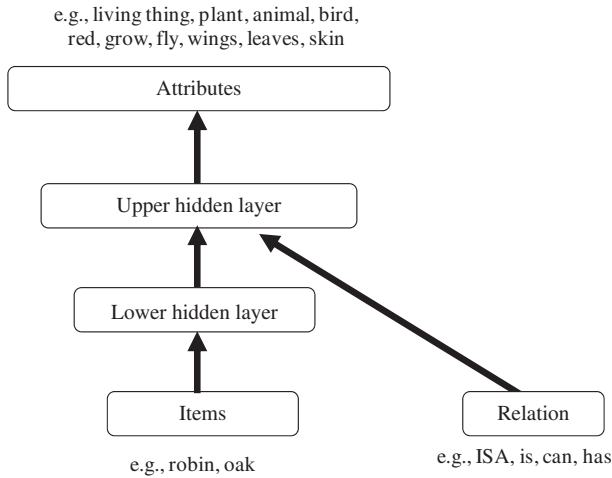
e.g., living thing, plant, animal, bird,
red, grow, fly, wings, leaves, skin

Figure 6 *Model of the development of semantic knowledge (Rogers and McClelland, 2004).*

across distributed representations, gradually differentiates across development, as the network learns the full set of propositions.

This model is important because the authors argued that the model exhibits many of the behaviours that other researchers had taken to indicate the presence of naïve, domain–specific theories guiding children's semantic cognition (e.g. one might have a theory about the differences between plants and animals, involving facts such as that the latter tend to move around a lot more.) For example, in the *theory* theory, knowledge of a concept consists not in a static list of features, but in its relation to a set of theories of how entities of various types tend to behave (e.g. this object is a living thing, it is an animal, and it is a bird; it therefore inherits a series of properties of living things, a more restricted set of properties for animals, and more restricted still for birds, and so forth).

One behaviour used to measure the structure of semantic knowledge is *inductive projection*. Children and adults are told that a given item has a novel property (e.g. an oak can *queem*, or an oak has a *queem*, or an oak is a *queem*). They are then asked which other items (objects, animals, etc.) might also have this novel property. In a series of experiments, Carey (1985) showed that children's answers to these kinds of questions change in systematic ways over development. Because abstraction and induction are key functions of the semantic system, these patterns provide important evidence about developmental change in the structure of semantic representations. Rogers and McClelland (2004) presented a series of simulations aimed at explaining two of these empirical effects: patterns of inductive property attribution can be different for different kinds of object properties; and patterns of inductive projection change over development, generally becoming more specific.

In order to simulate inductive projection, Rogers and McClelland took models at different stages of training and added a new attribute feature. The model was
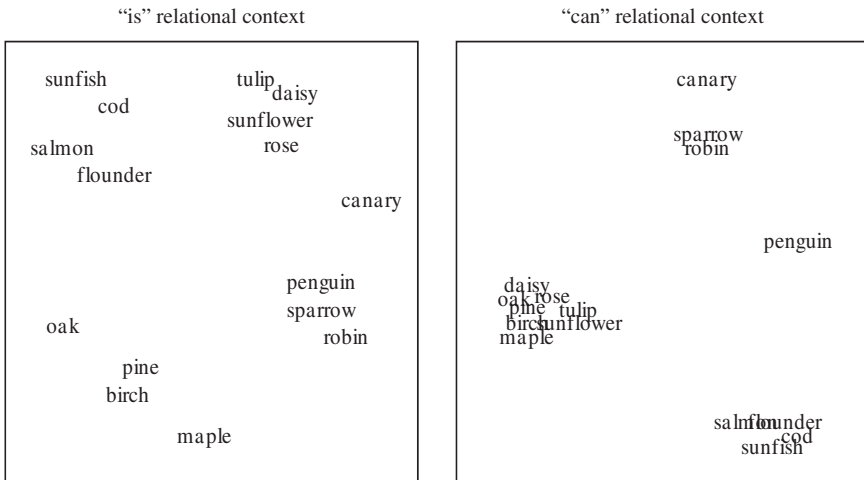
"is" relational context   "can" relational context

sunfish   tulip
        cod   daisy
                  sunflower
salmon   rose
    flounder

                        canary

                    penguin
                    sparrow
    oak   robin

        pine
        birch

            maple

canary

sparrow
robin

penguin

daisy
oak rose
    pine tulip
    birch sunflower
maple

salmflounder
        cod
    sunfish

Figure 7 *Similarity structure of hidden unit representations in the upper layer using multi-dimensional scaling, under two different 'relational' contexts.*

then trained to associate this attribute to the existing representation in the upper hidden layer, in the context of a particular relation (e.g. learning that an *oak can queem*). The authors then explored which other items also activated the new attribute, as a measure of inductive generalisation. Could *pines* also queem? What about *tulips*, or *canaries*?

Importantly, Rogers and McClelland viewed the representations in the upper hidden layer as being context-dependent, exhibiting different similarity structure depending on the relation that was specified, and as a consequence, exhibiting different generalisation properties. Figure 7 depicts the similarity structure of the representations in the upper hidden layer for two different contexts, the *is* relation and the *can* relation (adapted from Rogers and McClelland, 2004, Figure 8.2). Items that share many *is* or *can* properties generate similar patterns of activity across units in the upper hidden layer when that relation unit is activated. The model's behaviour reflects the acquisition of knowledge that different kinds of properties extend across different sets of objects.

Similar to the results of Carey's (1985) studies, this knowledge undergoes a gradual developmental change, whereby the model learns that different kinds of properties should be extended in different ways. The *is* context produces representations that are more delineated, because in the network's world, there are few properties shared among objects of the same kind. It therefore differentiates items in this context and as a result shows less of a tendency to generalise newly learned *is* context properties across categories. By contrast, in the *can* context, the items show less differentiated representations. For example, plants are collapsed into a single clump. This is because in the *can* context, all plants are associated with very similar upper hidden layer representations, because they all share exactly the same behaviours:
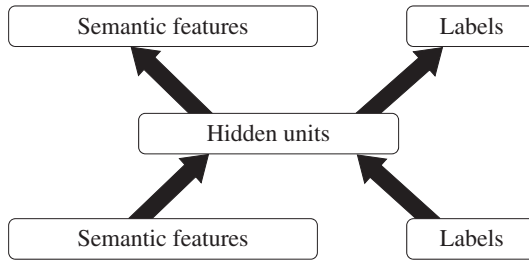
Figure 8 *Model of the comprehension of simple metaphorical comparisons (Thomas and Mareschal, 2001); labels of the B term in the metaphor 'an A is a B' serve as the context for reproducing the features of A.*

in the network's world, the only thing a plant can do is grow. Novel properties associated to any given plant in the *can* context are therefore more likely to be generalised to other plants.

In this model, then, the context of the relation fluidly shifted the similarity structure within semantic knowledge. The shift altered inductive behaviour in such a way that the network's behaviour seemed to be shaped by implicit conceptual theories. In fact, these theories consisted of statistical regularities learned in a given context. When the context changed, so did the statistical regularities that were brought to bear in processing the input. Rogers and McClelland's model shows us that the computational principle of context-sensitivity demonstrated in the previous section (arising from activation dynamics) scales to a larger and more complex connectionist network, in which it altered the 'meaning' of semantic tokens.

3.2 A Model of the Comprehension of Simple Metaphorical Comparisons

As we saw in the Introduction, the existence of non-literal similarity exemplified in metaphor and analogy is problematic for a theory of cognition based on literal categories. At least, if cognition were based on literal categories, and by extension, literal similarity, our ability to produce and comprehend instances of non-literal similarity would presumably require special purpose processing mechanisms, whose operation would (again presumably) produce some markers in behaviour or in brain activity. In fact, there is little evidence of either type to suggest that the distinction between literal and non-literal similarity has any psychological validity.

It may be unsurprising that the figurative meanings of well-known idioms, such as 'chew the fat' and 'kick the bucket', are comprehended more quickly than their literal interpretations (Gibbs, Nayak and Cutting, 1989): it seems plausible that they are lexicalised as specialist vocabulary. However, given enough context, people are no slower at reading familiar metaphorical sentences than comparable literal ones (Gibbs and Nagaoka, 1985; Ortony *et al.*, 1978). Inhoff, Lima and Carroll (1984) confirmed this finding with an eye-tracking study and also replicated it

with shorter contexts. This implies either that the metaphors were not interpreted figuratively, or, if they were, then the process required no additional computation to that required by literal processing. Furthermore, even *novel* metaphors may be comprehended as rapidly as comparable literal sentences, provided that the metaphors are apt (Blasko and Connine, 1993).

Similarly, neuroimaging studies also support the idea that literal/non-literal may be an uninformative dichotomy (see Giora, 2007, for discussion). Rapp and colleagues (Rapp *et al.*, 2007) failed to find differences in laterality between metaphorical and non-metaphorical sentences, either when the task involved judging a statement's metaphoricity, or whether it had positive or negative connotations. One exception is an fMRI study by Stringaris and colleagues (Stringaris *et al.*, 2007). These authors found that the left inferior frontal gyrus (LIFG) was more activated when judging metaphorical and anomalous sentences than when judging comparable literal statements. The LIFG has been hypothesised to mediate retrieval of semantic knowledge (e.g. Fiez *et al.*, 1992; Thompson-Schill *et al.*, 1997) and the authors suggested that additional semantic processing capacities were required for metaphorical processing. However, their task involved explicit judgement of the meaningfulness of statements, so it is not clear whether this recruitment of additional resources would take place in passive comprehension. Furthermore, the LIFG was also more active when judging anomalous statements, so whatever the region was doing, there was no suggestion that it was specific to non-literality.

Other imaging techniques such as electrophysiology have also failed to find evidence in favour of a literal/non-literal distinction. Pynte and colleagues (Pynte *et al.*, 1996) recorded event-related potentials (ERPs) and found that the terminal word of metaphors elicited larger N400 voltage components than did the terminal word of literal sentences, suggesting that the (incongruous) literal meaning of the metaphors was accessed during metaphor comprehension. However, the stimuli in that experiment were unfamiliar metaphors out of context. In a further experiment, it was found that preceding the metaphorical statement with a sentence that provided relevant context for the metaphor strongly reduced the N400 component, consistent with the notion that, when contextually relevant, the metaphorical meaning is the only one accessed. In other words, with sufficient context, metaphors appear to be processed in the same way as literal statements. (Across the various studies outlined above, 'sufficient context' appears to be that which constrains the discourse content such that the dimension(s) of similarity highlighted by the metaphor are expected or at least highly consistent with the discourse.)

Turning to metaphor theory, Black (1955, 1962, 1979) outlined three views of how the metaphor comprehension process may work. In the first of these, the substitution view, a metaphorical comparison must initially be replaced by a set of literal propositions that fit the same context. In the comparison view, the metaphor is taken to imply that the two terms are similar to each other in certain (communicatively relevant) respects. The intention of the comparison is to highlight these properties in the first term. In the interactive view, the comparison of the two terms in the metaphor is not taken to emphasise pre-existing similarities

between them, but itself plays a role in creating that similarity. The topic (first term) and vehicle (second term) interact such that the topic itself causes the selection of certain of the features of the vehicle, which may then be used in the comparison with the topic. The interactive view has been described as the dominant theory in the study of metaphor but also criticised for the vagueness of its central terms (Gibbs, 1994). One of the key issues for psycholinguistic models of metaphor comprehension has been to explain the nature of the interaction between topic and vehicle that constrains the emergent meaning of the comparison.

Three main models of the process have been proposed. These are the salience imbalance model (Giora, 1997, 2003, 2007; Ortony, 1979), the structural mapping model (Gentner, 1983, 1989; Gentner and Clements, 1988), and the class inclusion model (Glucksberg and Keysar, 1990, 1993). The salience imbalance model proposes that metaphors are similarity statements whose two terms share attributes. However, the salience of these attributes is much higher in the second term than the first. The comparison serves to emphasise these attributes in the first term. The structural mapping model suggests that topic and vehicle can be matched in three ways: in terms of their relational structure (that is, in the hierarchical organisation of their properties and attributes); in terms of those properties themselves; or in terms of both relational structure and properties. People tend to show a preference for relational mappings in metaphors. Lastly, the class inclusion model proposes that metaphors are understood as categorical assertions. In a metaphor *A is B*, *A* is assigned to a category denoted by *B*. Only those categories of which *B* is a member that could also plausibly contain *A* are considered as the intended meaning of the categorical assertion. That is, when I say *my job is a jail*, I am indicating that my job falls within the abstract category of jails, i.e. the category of constraining things (Glucksberg and Keysar, 1990).

Thomas and Mareschal (2001) investigated the third of these proposals, that simple metaphorical comparisons may be viewed as a form of categorisation. Thomas and Mareschal (2001; see also Purser *et al.*, 2009) used an auto-associative model of semantic memory to explore the hypothesis that metaphor comprehension may involve a form of strategic misclassification (see McClelland and Rumelhart, 1986, on the use of auto-associator networks as a model of semantic memory). It is the process of classification that transfers certain attributes from the *B* term (e.g. constraining things) to the *A* term (my job). In order to test whether *A* is a member of *B*, *A* is transformed by *B* knowledge. If it is little changed, it is likely a member of *B*. Reproduction as a means of assessing category membership is a widely used mechanism in connectionist models of memory (see Mareschal and Thomas, 2007). Under this view, psychological similarity itself is a transformational process rather than a comparison of static representations, which explains properties such as the asymmetry of comparisons, where *A* may be judged more similar to *B* than *B* is to *A* (Thomas and Mareschal, 1997).

One version of the metaphor model is shown in Figure 8. The network has distributed representations at all layers. For an illustrative example, the model was given a restricted semantic knowledge base covering just three concepts: apples,

balls, and forks. Training involved learning to reproduce the semantic features for individual exemplars of each category in the presence of the labels for that category (see Purser *et al.*, 2009). Once trained, a token is presented to the network, let us say an instance of a particular green apple. The system is now required to assess literal, metaphorical, or anomalous comparisons relating to this token. The sentence *this apple is an apple* would be viewed as a literal comparison; the sentence *this apple is a ball* would be viewed as a metaphorical comparison, perhaps emphasising that this apple is particularly round and that you are more likely to hit, kick or throw it than eat it; and the sentence *this apple is a fork* would be viewed as an anomalous comparison.

Each sentence is applied to the model in the following manner. The semantic features for the *A* term, the green apple, are applied to the input units across a semantic feature set, while the label for the *B* term (apple, ball, or fork) is also activated. The semantic output represents a version of the *A* term transformed by the comparison, while the activation of the output label tests membership of the category. Figure 9 shows the inputs and outputs for these comparisons over a set of semantic features. The literal comparison reproduces the apple features accurately and indicates high confidence that the token is indeed an apple. The metaphorical comparison produces lower confidence that the apple is a member of the category ball, but produces a transformed representation of the apple that attenuates the 'eaten' feature, and exaggerates both the 'roundness' of the apple and that it will be 'kicked' or 'hit'. The anomalous comparison produces the lowest confidence that the apple is a member of the category fork, and imposes properties of the central features of the fork category on the transformed representation: 'white', 'irregular', and 'large'. Importantly, the processing within semantic memory is identical in kind for the three types of comparison.

The model functions by using the context of the label (APPLE, BALL, FORK) to alter the similarity structure of the internal representations. The similarity structure serves to apply a different transformation to the semantic feature input, in a way that partly depends on the identity of that input. Figure 10 depicts the similarity structure of the internal representations under four contexts: (a) with each training exemplar for apples, balls, and forks presented in the context of its correct category label; (b) each exemplar presented in the context of the apple label; (c) each exemplar presented in the context of the ball label; (d) each label presented in the context of the fork label. The figure indicates the extent to which the similarity structure is warped by each label. This model can also be viewed as exploiting the *globality* of knowledge characterised by Fodor (2000). For example, one may view the output labels as testing the respective *simplicity* of the theory that the *A* term is a member of the *B* category: here, the simplest theory is that the green apple is indeed a member of the category apple. And the semantic transformation caused by activating different labels may be seen as exaggerating the *central* features of the *B* category when they are present in the *A* term. The globality of knowledge, in this case, is achieved by the full connectivity between features, internal representations, and labels.
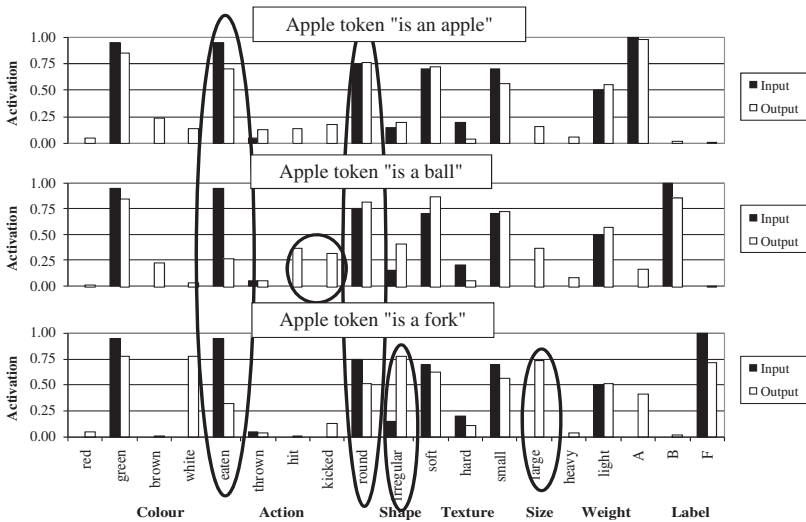
Figure 9 *Transformations of the meaning of the A term (a particular token of apple) by comparison to three B domains for the metaphor an A is a B. Ellipses indicate semantic features showing particular modulation (see text).*

The model constitutes the following theory of metaphor: all semantic knowledge is stored across a global representational system (as in Rogers and McClelland's model). Language labels are used as part of a strategic mechanism to manipulate context, bringing to bear different knowledge in the processing of a given semantic token than would normally be available when that token is met (e.g. ball knowledge would not normally be brought to mind when presented with apple tokens). This altered context serves to exaggerate or attenuate particular features of the token (depending on whether they are covariant with those same features in the 'ball' knowledge base, in this example), in the service of facilitating a particular communicative goal appropriate to the current discourse context (e.g. that this token of an apple is markedly round, or it may be thrown). However, within this framework, there is no principled difference between literal, metaphorical, or anomalous comparisons: they are just different forms of contextual modulation of semantic knowledge (see Leech, Mareschal and Cooper, 2008, for a related model applied to analogy).

One might argue that the view of metaphor as a form of categorisation is most consistent with the claim that metaphor comprehension requires no special processes over and above literal comprehension, since both the salience imbalance model and the structural mapping model imply a property matching procedure that is engaged for non-literal comparisons (Glucksberg, McGlone and Manfredi, 1997). Of course, the implemented version of the model only demonstrates its viability with simple comparisons. There are a number of criticisms that might be levelled at the model: Feature-based representations seem insufficient to deal with
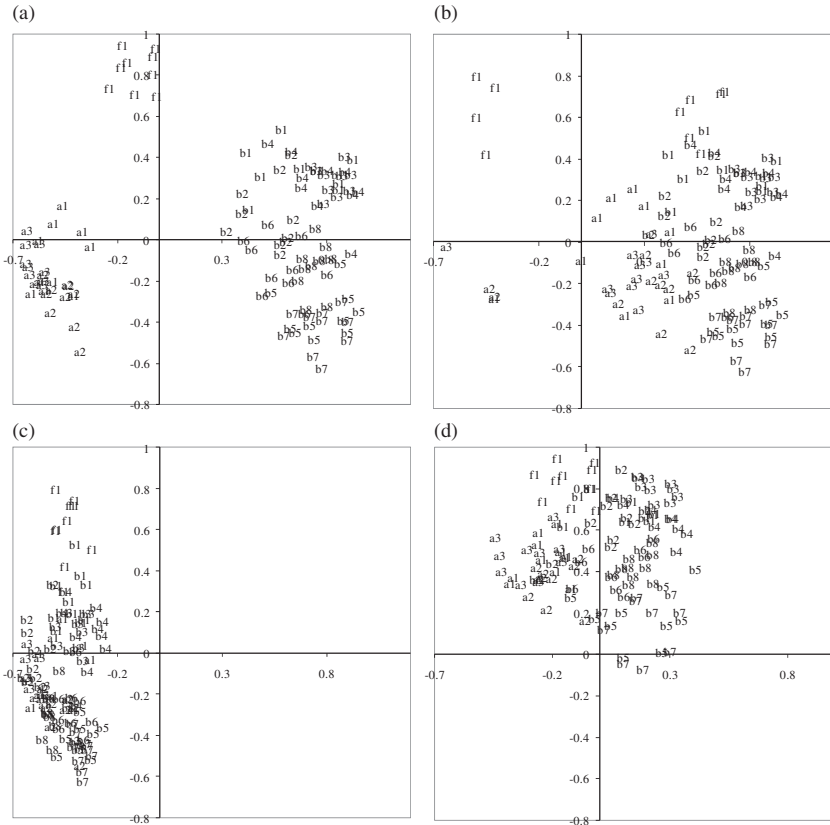
Figure 10 *The similarity structure of the internal representations (1st and 2nd principal components) under four contexts: (a) semantic feature vectors accompanied by their correct category labels; (b) all vectors labelled as balls; (c) all vectors labelled as apples; (d) all vectors labelled as forks.*

the complexities of sophisticated metaphorical expressions, and cannot deal with relational structure in concepts; in reality, the property transferred from vehicle to topic may not be a property of the vehicle itself (e.g. *the girl is a lollipop* may be taken to imply that the girl is frivolous—but lollipops themselves cannot be described as frivolous). Responses to these criticisms are discussed in Thomas and Mareschal (2001). Nevertheless, the model has a number of advantages: it is an implemented demonstration of the viability of context-sensitive representations as a means of explaining the interaction process in metaphorical comprehension; it explains the predictability of these interactions, based on properties of connectionist auto-associators; the model is developmental and its representations are learnable; it explains the asymmetry (and in some cases, non-reversibility) of metaphorical comparisons; and it generates testable predictions, which have subsequently received empirical support (Purser *et al.*, 2009).

4. Discussion

We have demonstrated a mechanism for producing fluid changes in the similarity structure of internal representations depending on context, using three different models. In the first two, the context altered the categorisations that the models performed; in the third, the context altered the transformations applied to the input. If these processing architectures are analogous to human cognition, they imply that categories of human knowledge are not fixed; instead, they represent temporary coalescences of dimensions of similarity, which are brought together by context in order to create the similarity structure in mental representations appropriate for the task at hand—a trait increasingly recognised as also characteristic of early cognition (Deák, 2003; Mareschal and Tan, 2007). What we view as literal categories are merely the most frequent or canonical contexts in which we process knowledge; the alterations in categorical structure produced by different and potentially novel contexts testify that literal knowledge is but one identity of a flexible underlying similarity structure.

We have argued that the view that categories are context dependent is not new. Our aim here was to demonstrate a computational mechanism by which context-dependent categorisation can be implemented in a connectionist network, one of the leading approaches to the computational theory of mind (McClelland *et al.*, 2010; Thomas and McClelland, 2008). Implementation demonstrates the viability of the theoretical proposal *contra*, for example, arguments by Fodor (2000) that context-dependent processing in connectionist networks is not possible since the causal property that drives processing—the connectivity matrix—is itself not dependent on context. This view is erroneous because it omits alterations in the effective thresholds of processing units. These change the computations that a layer of units can perform, even while the connectivity matrix is fixed. (It should be noted that Fodor has other reasons for not preferring connectionist architectures; see Fodor, 2000). Effective thresholds are not the only mechanism that can allow networks to behave in a context-sensitive manner. For instance, if the activity of one processing unit is permitted to alter the strength of a connection between two other units, much more radical variations in the causal properties of the network are possible (so-called *sigma-pi units*; see Feldman and Ballard, 1982; Rumelhart and McClelland, 1986).

Implementation also clarifies the assumptions of a theoretical proposal. In this case, the assumptions are that: (1) categorisation behaviour nevertheless relies on feature-based representations that are meaningful to the task at hand (even if these features may in practice be sub-lexical; see Thomas and Mareschal, 2001). These features are flexibly combined in different ways according to context; and (2) globality where it occurs is achieved by multiple connectivity. In other words, all bits of information can in principle influence the processing of all other bits of information because their representations are physically connected (directly or indirectly). High levels of connectivity in neural networks provide the opportunity for the operation of globality, but it is the details of the activation dynamics that demonstrate how it may be realised in a way that corresponds to human behaviour.

Last, implementation demonstrates that the representations required for context-dependent categorisation are learnable—all the models acquired their processing properties via exposure to a structured training environment.

The model of metaphorical comparisons we have presented here, although simple in itself, has wider implications that pertain to the relationship between language and thought. By way of background, there are broadly speaking two ways language is traditionally thought of as relating to thought: first, language may simply offer a read-off of the contents of thought; second, language may actively shape or constrain thought. The notion of 'verbal report' in psychology exemplifies the simplest version of the read-off idea. A more formal account can be found in Karmiloff-Smith's (1992) Representational Redescription model, which also holds that the language system can render explicit mental representations (thoughts) in a direct manner. The Sapir-Whorf hypothesis (Sapir, 1929/1958; Whorf, 1940) exemplifies the idea of an interaction between language and thought, the strong form of the hypothesis stating that our thinking is determined by language, and that linguistic form and meaning are inseparable.

In the metaphor model, language (in the form of labels) offers strategic control over context-sensitive representations of knowledge. In processing the semantic token of apple, the sentence 'The apple is a ball' brings to bear knowledge that shapes the representation of apple to enhance certain properties: the apple is thought of as rounder and more likely to be thrown. Consider that the normal function of context-sensitive representations, in humans and likely other animals, is to bring to bear knowledge that is relevant to the current situation—activated by certain perceptual features of that situation. The only knowledge that is activated is the relevant knowledge, which facilitates efficient action.

Language provides a set of labels and structures that are in principle independent of the properties of the current situation. In the form of metaphor and analogy, they allow the manipulation of context-sensitive representations to bring to bear a different context, which may enhance only certain features of the individual's representation (that is, thoughts about) the current situation. For example, even though an encounter with a ripe apple might normally evoke thoughts of edibility, its linguistic comparison to a ball would suppress this property and enhance others. The creative properties of metaphor and analogy stem from the fact that the enhanced features may prompt actions or responses to the situation that were not prompted by its initial, context appropriate representation in thought.

The combination of language and fluid, context-sensitive representations of knowledge, then, provides a third way to construe the relationship between language and thought: neither reading off nor shaping the contents of thought, but offering a tool to strategically bring to bear knowledge independent of the constraints of the current context. Needless to say, without language, animal cognitive systems (while perhaps equally powerful) would be locked into activating only the knowledge that is relevant to the current context.

Finally, the core of our argument has been the contention that categories are context sensitive, but this raises the following question: What is context? In the

models we presented, context took different guises, but in each case, it represented an additional input to the model. One could therefore argue that 'context-dependent processing' is an artefact of our definitions. We call one part of the input layer 'The Input' and another part 'The Context' and show how the activity of one part of the input layer influences computations carried out over another part of the input layer. But in reality, there is only a pattern of activation over a single input layer. Context is just another form of knowledge. The response to this argument is simply to ask, what else could context be but another source of information? The challenge is to identify experimentally the information sources that drive contextual effects in human categorisation. Of course, the division of input layers into Input and Context is, to some extent, arbitrary. In reality, all inputs serve as the context for all other inputs. This is an intrinsic property of the processing within connectionist networks, which makes them advantageous architectures for capturing the fluidity with which humans apply their knowledge to guiding their behaviour.

M. Thomas
Developmental Neurocognition Lab, Birkbeck College
H. Purser
Department of Psychological Sciences, Institute of Education London
D. Mareschal
Birkbeck College, University of London

References

Barsalou, L. 1983: Ad hoc categories. *Memory and Cognition*, 11, 211−27.

Barsalou, L. 1993: Flexibility, structure, and linguistic vagary in concepts. In A. Collins, S. Gathercole and M. Conway (eds), *Theories of Memory*. London: LEA, 29−101.

Black, M. 1955: Metaphor. *Proceedings of the Aristotelian Society*, 55, 273−94.

Black, M. 1962: *Models and Metaphors*. Ithaca, NY: Cornell University Press.

Black, M. 1979: More about metaphor. In A. Ortony (ed.), *Metaphor and Thought*. Cambridge: Cambridge University Press, 19−43.

Blasko, D. M. and Connine, C. M. 1993: Effects of familiarity and aptness on the comprehension of metaphor. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 19, 295−308.

Carey, S. 1985: *Conceptual Change in Childhood*. Cambridge, MA: MIT Press.

Chalmers, D., French, R. and Hofstadter, D. 1992: High-level perception, representation, and analogy. *Journal of Experimental & Theoretical AI*, 4, 185−211.

Deák, G. O. 2003: The development of cognitive flexibility and language abilities. In R. Kail (ed.), *Advances in Child Development and Behavior*, vol. 31. San Diego, CA: Academic Press, 271−327.

Feldman, J. A. and Ballard, D. H. 1982: Connectionist models and their properties. *Cognitive Science*, 6, 205–54.

Fiez, J. A., Peterson, S. E., Cheney, M. K. and Raichle, M. E. 1992: Impaired non-motor learning and error detection associated with cerebellar damage. A single-case study. *Brain*, 115, 155–78.

Fodor, J. 2000: *The Mind Doesn't Work That Way*. Cambridge, MA: MIT Press.

Gentner, D. 1983: Structure-mapping: a theoretical framework for analogy. *Cognitive Science*, 7, 155–70.

Gentner, D. 1989: The mechanisms of analogical learning. In S. Vosniadou and A. Ortony (eds.), *Similarity and Analogical Reasoning*. Cambridge: Cambridge University Press.

Gentner, D. and Clements, C. 1988: Evidence for relational selectivity in the interpretation of analogy and metaphor. In G. Bower (ed.), *The Psychology of Learning and Motivation*, vol. 22. Orlando, FL: Academic Press, 307–58.

Gibbs, R. W. 1994: *The Poetics of Mind*. Cambridge: Cambridge University Press.

Gibbs, R. W. and Nagaoka, N. 1985: Getting the hang of American slang: studies on understanding and remembering slang metaphors. *Language & Speech*, 28, 177–94.

Gibbs, R. W., Nayak, N. P. and Cutting. C. 1989: How to kick the bucket and not decompose: analyzability and idiom processing. *Journal of Memory and Language*, 28, 576–93.

Giora, R. 1997: Understanding figurative and literal language: the graded salience hypothesis. *Cognitive Linguistics*, 7, 183–206.

Giora, R. 2003: *On Our Mind: Salience, Context, and Figurative Language*. Oxford: Oxford University Press.

Giora, R. 2007: Is metaphor special? *Brain and Language*, 100, 111–4.

Glucksberg, S. 2000: *Understanding Figurative Language: From Metaphor to Idioms*. Oxford: Oxford University Press.

Glucksberg, S. and Estes, Z. 2000: Feature accessibility in conceptual combination: effects of context-induced relevance. *Psychonomic Bulletin & Review*, 7, 510–15.

Glucksberg, S. and Keysar, B. 1990: Understanding metaphorical comparisons: beyond similarity. *Psychological Review*, 97, 3–18.

Glucksberg, S. and Keysar, B. 1993: How metaphors work. In A. Ortony (ed.), *Metaphor and Thought*, 2nd edn. Cambridge: Cambridge University Press.

Glucksberg, S., McGlone, M. S. and Manfredi, D. 1997: Property attribution in metaphor comprehension. *Journal of Memory and Language*, 36, 50–67.

Hinton, G. E. 1981: Implementing semantic networks in parallel hardware. In G. E. Hinton and J. A. Anderson (eds), *Parallel Models of Associative Memory*. Hillsdale, NJ: Erlbaum, 161–87.

Inhoff, A. W., Lima, S. D. and Carroll, P. J. 1984: Contextual effects on metaphor comprehension. *Memory and Cognition*, 12, 558–67.

James, W. 1890: *The Principles of Psychology*, 2 vols. New York: Henry Holt; Reprinted Bristol: Thoemmes Press, 1999.

Karmiloff-Smith, A. 1992: *Beyond Modularity: A Developmental Perspective on Cognitive Science*. Cambridge, MA: MIT Press/Bradford Books.

Leech, R., Mareschal, D. and Cooper, R. 2008: Analogy as relational priming: a developmental and computational perspective on the origins of a complex cognitive skill. *Behavioral & Brain Sciences*, 31, 357–414.

Mareschal, D. and Tan, S. 2007: Flexible and context-dependent categorisation by eighteen-month-olds. *Child Development*, 78, 19–37.

Mareschal, D. and Thomas, M. S. C. 2007: Computational modeling in developmental psychology. *IEEE Transactions on Evolutionary Computation (Special Issue on Autonomous Mental Development)*, 11, 137–50.

Mareschal, D., Johnson, M. H., Sirois, S., Spratling, M., Thomas, M. S. C. and Westermann, G. 2007: *Neuroconstructivism, Vol. I: How the Brain Constructs Cognition*. Oxford: Oxford University Press.

McClelland, J. L., Botvinick, M. M., Noelle, D. C., Plaut, D. C., Rogers, T.T., Seidenberg, M. S. and Smith, L. B. 2010: Letting structure emerge: connectionist and dynamical systems approaches to understanding cognition. *Trends in Cognitive Sciences*, 14, 348–56.

McClelland, J. L. and Rumelhart, D. E. 1986: A distributed model of human learning and memory. In J. L. McClelland, D. E. Rumelhart and The PDP Research Group (eds), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Vol. 2: Psychological and Biological Models*. Cambridge, MA: MIT Press, 170–215.

Murphy, G. L. 1988: Comprehending complex concepts. *Cognitive Science*, 12, 529–62.

Murphy, G. L. 2003: *The Big Book of Concepts*. Cambridge, MA: MIT Press.

Murphy, G. L. and Medin, D. L. 1985: The role of theories in conceptual coherence. *Psychological Review*, 92, 289–316.

Oakes, L., Newcombe, N. and Plumert, J. 2009: Are dynamic systems and connectionist approaches an alternative to good old fashioned cognitive development? In J. Spencer, M. S. C. Thomas and J. McClelland (eds), *Toward a New Unified Theory of Development*. Oxford: Oxford University Press, 279–94.

Ortony, A. 1979: Beyond literal similarity. *Psychological Review*, 86, 161–80.

Ortony, A., Schallert, D. L., Reynolds, R. E. and Antos, S. J. 1978: Interpreting metaphors and idioms: some effects of context on comprehension. *Journal of Verbal Learning & Verbal Behaviour*, 17, 465–77.

Purser, H. R. M., Thomas, M. S. C., Snoxall, S. and Mareschal, D. 2009: The development of similarity: testing the prediction of a computational model of metaphor comprehension. *Language and Cognitive Processes*, 24, 1406–30.

Pynte, J., Besson, M., Robichon, F. H. and Poli, J. 1996: The time-course of metaphor comprehension: an event-related potential study. *Brain & Language*, 55, 293–316.

Quine, W. V. O. 1977: Natural kinds. In S. P. Schwartz (ed.), *Naming, Necessity, and Natural Kinds*. Ithaca, NY: Cornell University Press, 155–75.

Rapp, A. M., Leube, D. T., Erb, M., Grodd, W. and Kircher, T. T. J. 2007: Laterality in metaphor processing: Lack of evidence from functional magnetic resonance imaging for the right hemisphere theory. *Brain & Language*, 100, 142–49.

Rogers, T. and McClelland, J. 2004: *Semantic Cognition*. Cambridge, MA: MIT Press.

Roth, E. M. and Shoben, E. J. 1983: The effect of context on the structure of categories. *Cognitive Psychology*, 15, 346–78.

Rumelhart, D. E. and McClelland, J. L. 1986: *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Volumes I, II*. Cambridge, MA: MIT Press.

Rumelhart, D. and Todd, P. 1993: Learning and connectionist representations. In D. Meyer and S. Kornblum (eds.), *Attention and Performance XIV*. Cambridge, MA: MIT Press, 3–30.

Rumelhart, D., Hinton, G. and Williams, R. 1986: Learning internal representations by error propagation. In D. Rumelhart, J. McClelland and the PDP Research Group (eds), *Parallel Distributed Processing Vol.* 1. Cambridge, MA: MIT Press, 318–62.

Sapir, E. 1929/1958: The status of linguistics as a science. In E. Sapir, *Culture, Language and Personality*, D. G. Mandelbaum, ed. Berkeley, CA: University of California Press.

Sirois, S., Spratling, M., Thomas, M. S. C., Westermann, G., Mareschal, D. and Johnson, M. H. 2008: Précis of neuroconstructivism: how the brain constructs cognition. *Behavioral and Brain Sciences*, 31, 321–56.

Stringaris, A. K., Medford, N. C., Giampietro, V. C., Brammer, M. J. and David, A. S. 2007: Deriving meaning: distinct neural mechanisms for metaphoric, literal, and non-meaningful sentences. *Brain & Language*, 100, 150–62.

Thomas, M. S. C. and Mareschal, D. 1997: Connectionism and psychological notions of similarity. *Proceedings of the 19th Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Lawrence Erlbaum.

Thomas, M. S. C. and Mareschal, D. 2001: Metaphor as categorisation: a connectionist implementation. *Metaphor & Symbol*, 16, 5–27.

Thomas, M. S. C. and McClelland, J. L. 2008: Connectionist models of cognition. In R. Sun (ed.), *Cambridge Handbook of Computational Psychology*. Cambridge: Cambridge University Press, 23–58.

Thompson-Schill, S. L., D'Esposito, M., Aguirre, G. K. and Farah, M. J. 1997: Role of left inferior prefrontal cortex in retrieval of semantic knowledge: a reevaluation. *Proceedings of the National Academy of Sciences, USA*, 94, 14792–7.

Whorf, B. L. 1940: Science and linguistics. *Technology Review*, 42, 229–31, 247–8.

Wittgenstein, L. 1953: *Philosophical Investigations*. Oxford: Blackwell.