The mystery of thought: demystified by context-dependent categorisation?

Michael S. C. Thomas[1], Harry R. M. Purser[1], & Denis Mareschal[2]

[1]Developmental Neurocognition Lab, Birkbeck, University of London, UK

[2]Centre for Brain and Child Development, School of Psychology, Birkbeck College, London, UK

Running head: Metaphorical and literal comprehension

Word count: 7,000

Address for correspondence:

Dr. Harry Purser
Developmental Neurocognition Lab
School of Psychology
Birkbeck College
Malet Street
London WC1E 7HX, UK
Email: h.purser@bbk.ac.uk
Web: http://www.psyc.bbk.ac.uk/research/DNL/
Tel.: +44 (0)20 7631 6386
Fax: +44 (0)20 7631 6312

Abstract

Although much of cognitive science assumes that literal categories are central to cognition, there is scant evidence for a psychologically meaningful distinction between the literal and nonliteral. We advance the idea that there is, in fact, no principled difference between literal and nonliteral comparisons; each is a different type of contextual modulation of semantic knowledge. Generally, the context specificity of human thought is viewed by some as 'mysterious' under current models of cognition (e.g. Fodor, 2000). Using connectionist modelling, we demonstrate a simple computational principle by which this contextual modulation might be achieved, beginning with a simple five-unit network, then showing how the principle scales to more complex models of semantic memory and metaphor comprehension, establishing that the similarity structure of semantic knowledge can be fluidly manipulated by context within a structurally fixed processing structure. The theory of metaphor suggested by this computational view prompts a particular view of the relation between language and thought, namely that 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. The implications of such a view for creativity and the nature of categories are discussed.

Introduction

Much of cognitive science is premised on the assumption that the discovery of literal categories (i.e., pre-existing groups of related kinds in the world) is crucial to human cognition (e.g., Murphy, 2003). Intrinsic to the idea of literal categories is literal similarity, most obviously used to identify the category to which a given entity belongs. However, high-level human cognition is also characterized 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 fundamental divide between literal and figurative similarity, the production and comprehension of metaphorical and analogical comparisons would seem to imply the existence of additional, special cognitive mechanisms.

In fact, there is very little evidence to suggest that the distinction between literal and nonliteral 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, & 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 & Nagaoka, 1985; Ortony, Schallert, Reynolds, & Antos, 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 & Connine, 1993).

Any difference between literal and nonliteral processing might be expected to be seen in differential activation of brain areas for each type of processing. However, neuroimaging studies also support the idea that literal-nonliteral may be an uninformative dichotomy. Rapp and colleagues (Rapp et al., 2007) failed to find differences in laterality between metaphorical and nonmetaphorical sentences, either when the task involved judging a statement's metaphoricity, or whether it had positive or negative connotations. In another fMRI study, Stringaris and colleagues (Stringaris et al., 2007) 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 nonliterality.

Other imaging techniques have also failed to find evidence in favour of a literal-nonliteral distinction. Pynte and colleagues (Pynte et al., 1996) recorded ERPs and found that the terminal word of metaphors elicited larger N400 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. 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.

In this article, we advance the idea that there is, in fact, no principled difference between literal and nonliteral comparisons; each is a different type of contextual modulation of semantic knowledge. Metaphor, on this view, is a process of utilising language as a strategic control mechanism to manipulate context, bringing to bear particular knowledge in processing a given semantic token. This knowledge may not have been available without this manipulation of context.

Contextual effects on semantic knowledge are well established. Barsalou (1993) noted that when participants are 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 & Estes, 2000; Murphy, 1988; Roth & Shoben, 1983). For example, from the imagined perspective of a Chinese person, *swan* and *peacock* may be highly representative, 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*'s 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" (pp.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 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”, “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 characterized 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 (cf. Barsalou, 1982); 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; *simplicity* is a property at the level of the explanation, but not of the constituent representations. 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, the causal properties of reasoning systems are driven by local rather than global properties of representations (the syntactic structure and the connectivity matrix, respectively).

Thus, Fodor's comments highlight that context-sensitivity appears mysterious under current conceptions of the mind. How could context-sensitivity operate in real representational systems with fixed causal structures? Can context sensitivity emerge from 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 very 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 to show how semantic memory could demonstrate context dependence and then 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 very simple computational architectures can nevertheless show complex patterns of context-sensitive processing, and also to show that this property enables similar models to account for high-level human behaviours.

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, & Williams, 1986) (i.e. it cannot be solved by a two-layer network). The network traditionally used to solve the problem has only five units: two input units, two 'hidden' (i.e. internal) units, and one output unit (Figure 2). This network is both well-known and very 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 1a shows the four patterns comprising the problem represented in a two-dimensional 'input space'. The computational complexity arises because the output unit of a network can only make a single categorization 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.

======== *insert Table 1 about here* =========

The following is an example of such a five-unit network learning the internal representations necessary to solve the XOR problem, which 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. Figure 1b shows how the similarity structure of the input space has been re-represented over the hidden units, 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.

======== *insert Figure 1 about here* =========

======== *insert Figure 2 about here* =========

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

======== *insert Table 2 about here* =========

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 backpropagation, 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), which is the attribute that Fodor, for example, considers to be 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.

======== *insert Figure 3 about here* =========

======== *insert Figure 4 about here* =========

======== *insert Figure 5 about here* =========

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. If it inhibits the hidden unit, it is raising the effective threshold, because less excitation is now necessary for the input units to push the hidden unit over its 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 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 more likely to turn on, while input from B makes it less 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 categorizations 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).

With this demonstration in place, we can now move on to more complex models that illustrate the contextual modulation of semantic knowledge. These models will demonstrate how the similarity structure of semantic knowledge can be fluidly manipulated by context.

Two models of context-dependent categorization

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. This conceptual knowledge, stored across distributed representations, gradually differentiates across development.

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.) 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., it can *queem,* or it has a *queem,* or it 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.

======== *insert Figure 6 about here* =========

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 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 generalization. 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, different generalization 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 & McClelland, 2004, Fig. 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 generalize 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: 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 generalized 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. Rogers and McClelland's model shows us that the computational principle of context-sensitivity expostulated here (arising from activation dynamics) scales to a larger and more complex connectionist network, altering the 'meaning' of semantic tokens. Having demonstrated this principle in inductive projection, we will now do the same in a model of metaphor comprehension. Metaphor comprehension is sensitive to context (e.g. Gibbs & Nagaoka, 1985; Inhoff, Lima, & Carroll, 1984), rendering it a particularly appropriate area to illustrate the computational principle under consideration.

======== *insert Figure 7 about here* =========

2. Metaphor comprehension

Thomas and Mareschal (2001) investigated the proposal that metaphor may be viewed as a form of categorization (Glucksberg & Keysar, 1990). 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. Thomas and Mareschal (2001; see also Purser et al., in press) used an autoassociative model of semantic memory to explore the hypothesis that metaphor comprehension may involve a form of strategic misclassification (see McClelland & Rumelhart, 1986, on the use of autoassociator 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 & Thomas, 2007).

======== *insert Figure 8 about here* =========

One version of this 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., in press). 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 emphasizing 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'.

======== *insert Figure 9 about here* =========

The model functions by using the context of the label 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 characterized 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.

======== *insert Figure 10 about here* =========

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, & Cooper, 2008, for a related model applied to analogy).

Discussion

The aim of this article was to demonstrate a computational mechanism by which both metaphorical and literal comparisons can be achieved, showing that context-dependent cognitive flexibility can be implemented in a connectionist network. Implementation demonstrates the viability of the theoretical proposal *contra*, for example, arguments by Fodor that context-dependent processing in connectionist networks is not possible because 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). Thus, context sensitivity may not be a mysterious aspect of cognition, since it can be instantiated with a simple computational principle.

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 & 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). Furthermore, implementation demonstrates that the representations required for context-dependent categorisation are learnable – all three models considered acquired their processing properties via exposure to a training set.

This computational principle was initially demonstrated using a simple five-unit connectionist network, affording a clear example. Following this, it was shown that the principle scaled up to larger and more complex models of semantic memory. In Rogers and McClelland's (2004) model, context specified the 'theory' (i.e. *the thing can* or *the thing is*) that was brought to bear by the network in its inductive projection behaviour. Finally, a model of metaphor comprehension was outlined, in which the context of verbal labels served to alter the similarity structure of internal representations.

This last model demonstrated a new view of metaphor: namely that metaphorical statements are, at bottom, no different from literal ones. Each achieves its communicative purpose by strategically modulating context, in order to bring to bear particular information in the processing of a given semantic token. In the case of metaphor, this information would not normally be active in the current discourse context; if the statement is literal, this information would typically be active. If there is no meaningful difference between literal and nonliteral similarity, then metaphor may be said not to exist, because all comparisons are simply context-specific coalescences of particular semantic features and dimensions. It is worth pointing out that 'nonliteral comparison' is a potentially confusing notion: the features that are highlighted by a nonliteral (metaphorical) comparison literally are shared (e.g. "The apple is a ball": the apple really is small, round and throwable). Of course, this reflects the central message of the current article: literal and nonliteral comparisons are really the same kind of thing; the literal version of a category is its similarity structure in the most commonly encountered context of usage.

This theory of metaphor supports a particular view of the relation between language and thought, namely that language affords the strategic control of context on semantic knowledge, allowing information to be brought to bear that might otherwise not be available for processing. In other words, language is a strategic tool to manipulate the context of thought. Within this framework, the similarity structure of language representations is orthogonal to the function that they perform. Hence, in the models described in this article, language consists of atomic labels, whereas it is semantic features that have a similarity structure. In the absence of language, on this view, the availability of semantic knowledge would be situationally determined. One prediction that arises naturally from this hypothesis is that

animals that do not have language will be unable to bring to bear knowledge that is not determined by the current situation.

In order to put this view of language in context, it is worth briefly describing two well-known alternatives. One is the Sapir-Whorf hypothesis (Sapir, 1929/1958; Whorf, 1940), the strong form of which states that our thinking is determined by language, and that linguistic form and meaning are inseparable. In contrast, the notion of 'verbal report' in psychology, which entails that language is (or, at least, can be) determined by thought. A more formal account is Karmiloff-Smith's (1992) Representational Redescription model, which also holds that the language system can 'read off' mental representations (i.e. thoughts) in a direct manner. This view seems uncontroversial, in essence. We suggest that both language and thought influence each other: if we want to say that an apple is throwable, this thought of *throwing* might bring to mind the word *ball*, which, in turn, would influence the features of *apple* that were brought to mind. On this view, then, language allows us to *control* thought and this control is fluid, goal-directed and adaptive. In contrast, according to the Sapir-Whorf hypothesis, language influences thought through some deterministic inseparability of language and meaning; a critical difference between these conceptions, then, is that the similarity structure of language representations is *not* orthogonal to the function that they perform in the Sapir-Whorf view.

If language affords the strategic control of thought, then something that is generally considered rather mysterious might 'come for free', namely creativity (at least according to some definitions): while some other, verbal, accounts of metaphor allude vaguely to notions of conceptual recombination and the like, the account of metaphor expounded here demonstrates how particular features or dimensions of a metaphor topic may be exaggerated and attenuated, depending on context, allowing concepts to be modified online. Thus, after comprehending "The apple is a ball", one's online concept of *ball* will have been modified in

such a way that the *small*, *round* and *throwable* aspects of the concept will have been exaggerated. Mareschal and colleagues have argued that human categorization behavior 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; Mareschal & Tan, 2007). Synthesising Mareschal's position with our view that metaphor is not principally different from the literal, we suggest that creativity is not some esoteric notion requiring esoteric mechanisms: it is just a consequence of how the mind works, because cognition is intrinsically context-dependent and therefore is creative to the extent that context changes (and can be manipulated). Creativity, then, may be considered a tool to manipulate the salience of different features of a given object or situation, where exaggerated features trigger associations or task schemas that were previously unnoticed.

Murphy and Medin (1985) make the point that mental chemistry is a more apt metaphor for understanding concepts than mental composition, emphasising relations, operations and transformations, as opposed to viewing features as inert and independent entities. John Stuart Mill (1843/1965) had the following to say on the matter:

> ...when the seven prismatic colors are presented to the eye in rapid succession, the sensation produced is that of white. But in this last case it is correct to say that the seven colors when they rapidly follow one another *generate* white, but not that they actually *are* white; so it appears to me that the Complex Idea, formed by the blending together of several simpler ones, should, when it really appears simple, (that is when the separate elements are not consciously distinguishable in it) be said to *result from*, or be *generated by* the simple ideas, not to *consist* of them. . . . These are cases of mental chemistry: in which it is possible to say that the simple ideas generate, rather than that they compose, the complex ones. (p. 29)

The idea of partial representations interacting to produce new concepts, then, is long-established (see Barsalou, 1993; Chalmers et al., 1992; Mareschal et al., 2007; Smolensky,

1988; for interim development of the idea). To reiterate, the contribution of the current article is to demonstrate a *mechanism* by which this can be achieved.

What exactly are categories, if they are not lists of features stored in long-term memory? Harnad (2005) has argued that "categorization is any systematic differential interaction between an autonomous, adaptive sensorimotor system and its world (p.21)". In his view, categories only exist to the extent that we behave differently to different kinds of entities: a hard-line commitment to context-specificity and goal-direction. The other side of the coin of context-specific categories is *incomplete content* of definitions of those categories: it is almost always possible to think of additional ways of describing and defining a category, e.g. birds do not have fur, they tend to drink water, they are descended from dinosaurs. Linguistic descriptions may be recursive, or nested, such that any given level of description can be further explicated in terms of another. This potentially limitless aspect of description renders the possibility of complete content rather remote. Part of this problem of incomplete content stems from the difference between stored and inferred knowledge (Barsalou, 1993), because many of the linguistic descriptions that people offer for categories are likely to be formulated spontaneously rather than retrieved from long-term memory. Semioticians have used the problem of incomplete content to argue that linguistic labels (i.e. words) have no ultimate determinable meaning: Derrida (1976, 1978) coined the term 'différance' to allude to the way in which (in his view) meaning is endlessly deferred, a notion earlier offered by Peirce (1931-1958): "The meaning of a representation can be nothing but a representation ... the interpretant is nothing but another representation ... and as representation, it has its interpretant again. Lo, another infinite series". However, the view that categories are context-specific deftly avoids this problem: the meaning of a label is constrained by context and communicative intent. 'Complete' content, then, would be the integrated descriptions of a

category in every possible context. In neuroconstructivist terms, Sirois and colleagues (Sirois et al., 2008) suggest that "For a representation to become full, the individual must integrate the partial representations across the entire range of contexts in which the concept is used."

What is context? In the various models demonstrated, 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 an input layer. 'Context' is therefore 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 categorization. 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 connectionist networks, which makes them advantageous architectures for capturing the fluidity with which humans apply their knowledge to guiding their behaviour.

Acknowledgements

This work was supported by European Commission grant NEST-029088(ANALOGY) and MRC grant G0300188.

References

Barsalou, L. (1983) Ad hoc categories. *Memory and Cognition, 11*, 211-227.

Barsalou, L. (1993). Flexibility, structure, and linguistic vagary in concepts. In A. Collins, S. Gathercole, & M. Conway (Eds.), *Theories of memory* (pp. 29-101). London: LEA.

Blasko, D. M. & 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.

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

Derrida, J. (1976). *Of Grammatology* (trans. Gayatri Chakravorty Spivak). Baltimore, MD: Johns Hopkins University Press.

Derrida, J. (1978). *Writing and Difference* (trans. Alan Bass). London: Routledge & Kegan Paul.

Fiez, J. A., Peterson, S. E., Cheney, M. K., & Raichle, M. E. (1992). Impaired non-motor learning and error detection associated with cerebellar damage. A single-case study. *Brain, 115*, 155–178.

Fodor, J. (2000). *The mind doesn't work that way*. Cambridge, MA: MIT Press.

Gibbs, R. W. & Nagaoka, N. (1985). Getting the hang of American slang: Studies on understanding and remembering slang metaphors. *Language & Speech, 28*, 177-194.

Gibbs, R. W., Nayak, N. P., & Cutting. C. (1989). How to kick the bucket and not decompose: Analyzability and idiom processing. *Journal of Memory and Language, 28*, 576-593.

Glucksberg, S. & Estes, Z. (2000). Feature accessibility in conceptual combination: Effects of context-induced relevance. *Psychonomic Bulletin & Review, 7*, 510-515.

Glucksberg, S. & Keysar, B. (1990). Understanding metaphorical comparisons: Beyond similarity. *Psychological Review, 97*, 3-18.

Harnad, S. (2005) To cognize is to categorize: Cognition is categorization. In C. Lefebvre & H. Cohen (Eds.). *Handbook of categorization in cognitive science* (pp. 20-42). London: Elsevier.

Hinton, G. E. (1981). Implementing semantic networks in parallel hardware. In G. E. Hinton & J. A. Anderson (Eds.) *Parallel models of associative memory* (p. 161-187). Hillsdale, NJ: Erlbaum.

Inhoff, A. W., Lima, S. D., & Carroll, P. J. (1984). Contextual effects on metaphor comprehension. *Memory and Cognition*, 12, 558-567.

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

Leech, R., Mareschal, D. & 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., Johnson, M. H., Sirois, S., Spratling, M., Thomas, M. S. C., & Westermann, G. (2007). *Neuroconstructivism, Vol. I: How the brain constructs cognition*. Oxford, UK: Oxford University Press.

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

Mareschal, D. & Thomas, M. S. C. (2007). Computational modeling in developmental psychology. *IEEE Transactions on Evolutionary Computation (Special Issue on Autonomous Mental Development), 11*, 137-150.

Mill, J. S. (1965). *On the logic of the moral sciences*. New York: Bobbs-Merrill. (Originally published 1843).

Murphy, G. L. (2003) *The big book of concepts*. Cambridge, MA: MIT Press.

Murphy, G. L. & Medin, D. L. (1985): The role of theories in conceptual coherence. *Psychological Review, 92*, 289-316

Oakes, L., Newcombe, N., & Plumert, J. (2009). Are dynamic systems and connectionist approaches an alternative to Good Old Fashioned Cognitive Development? In J. Spencer, M. S. C. Thomas, & J. McClelland (Eds.), *Toward a new unified theory of development* (pp. 279-294). Oxford: OUP.

Ortony, A., Schallert, D. L., Reynolds, R. E., & Antos, S. J. (1978). Interpreting metaphors and idioms: Some effects of context on comprehension. *Journal of Verbal Learning & Verbal Behavior, 17*, 465-477.

Peirce, C. S. (1931-58). *Collected Writings* (8 Vols.). Cambridge, MA: Harvard University Press.

Purser, H. R. M., Thomas, M. S. C., Snoxall, S., & Mareschal, D. (in press). The development of similarity: Testing the prediction of a computational model of metaphor comprehension. *Language & Cognitive Processes*.

Pynte, J., Besson, M., Robichon, F. H., & 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* (pp. 155-175). Ithaca, NY: Cornell University Press.

Rapp, A. M., Leube, D. T., Erb, M., Grodd, W., & 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,* 142-149.

Rogers, T., & McClelland, J. (2004). *Semantic cognition*. Cambridge, MA: MIT Press.

Rumelhart, D., Hinton, G., & Williams, R. (1986). Learning internal representations by error propagation. In D. Rumelhart, J. McClelland & the PDP research group (Eds.) *Parallel Distributed Processing Vol. 1*, (pp.318-362). Cambridge, MA: MIT Press.

Rumelhart, D., & Todd, P. (1993). Learning and connectionist representations. In D. Meyer & S. Kornblum (Eds.), *Attention and performance XIV* (pp. 3-30). Cambridge, MA: MIT Press.

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

Sirois, S., Spratling, M., Thomas, M. S. C., Westermann, G., Mareschal, D., & Johnson, M. H. (2008). Precis of Neuroconstructivism: How the Brain Constructs Cognition. *Behavioral and Brain Sciences, 31,* 321-356.

Smolensky, P. (1988). On the proper treatment of connectionism. *Behavioral and Brain Sciences, 11,* 1-74.

Stringaris, A. K., Medford, N. C., Giampietro, V. C., Brammer, M. J., & David, A. S. (2007). Deriving meaning: Distinct neural mechanisms for metaphoric, literal, and non-meaningful sentences. *Brain & language, 100,* 150-162.

Thomas, M. S. C. & Mareschal, D. (2001). Metaphor as categorisation: A connectionist implementation. *Metaphor & Symbol, 16,* 5-27.

Thompson-Schill, S. L., D'Esposito, M., Aguirre, G. K., & 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,* 14792–14797.

Whorf, B. L. (1940). Science and linguistics. *Technology Review 42*, 229-31, 247-8.

Wittgenstein, L. (1953). *Philosophical investigations*. Oxford, UK: Blackwell.

Tables

Table 1: The XOR mapping problem.

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

p3	0	1	1
p4	1	1	0

Table 2: The Hexagon mapping problem.

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

Figure Captions

Figure 1: Geometric representation of the XOR input space and a sample hidden unit space for a network.

Figure 2: Exclusive-or (XOR) network.

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

Figure 4: Hexagon network.

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

Figure 6: Model of the development of semantic knowledge (Rogers & McClelland, 2004).

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

Figure 8: Model of metaphor comprehension (Thomas & 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.

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).

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 label; (b) all vectors labelled as balls; (c) all vectors labelled as apples; (d) all vectors labelled as forks.

Figure 1

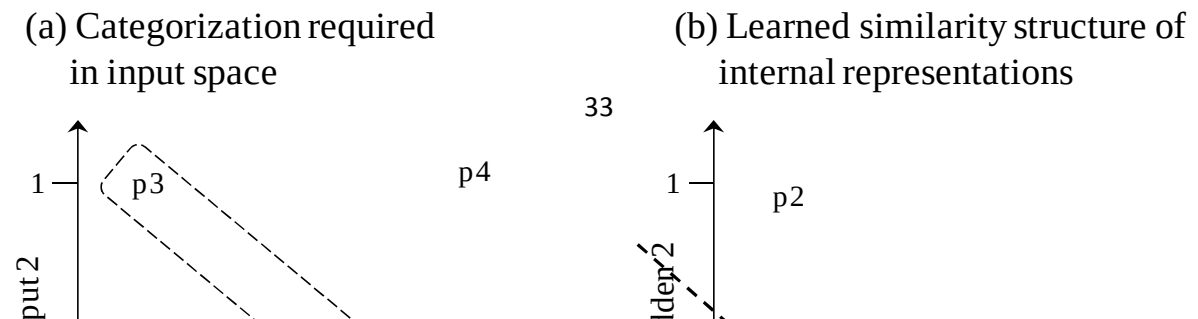


Figure 2

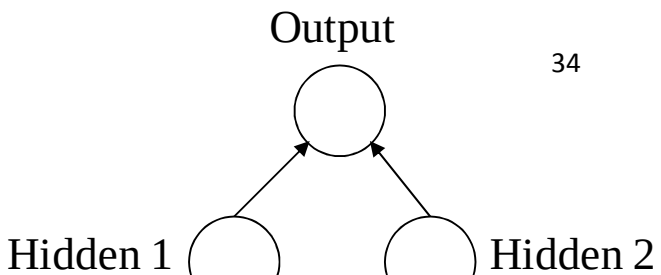
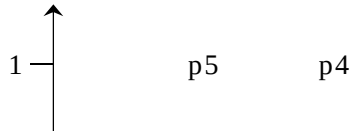
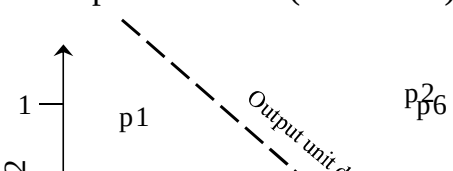


Figure 3

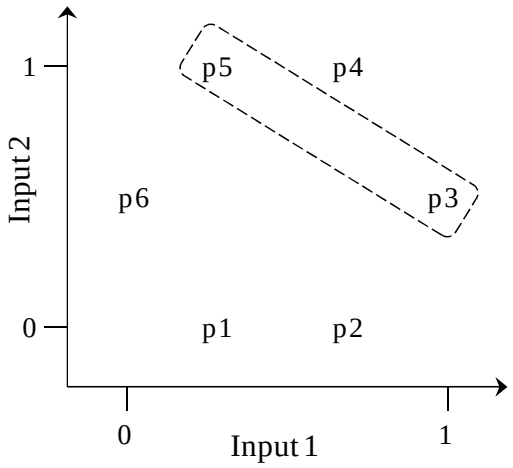
(a) Categorization required in input space (Context A)



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



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



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

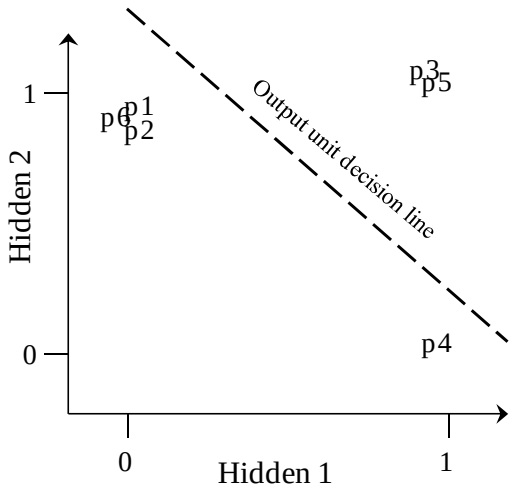


Figure 4

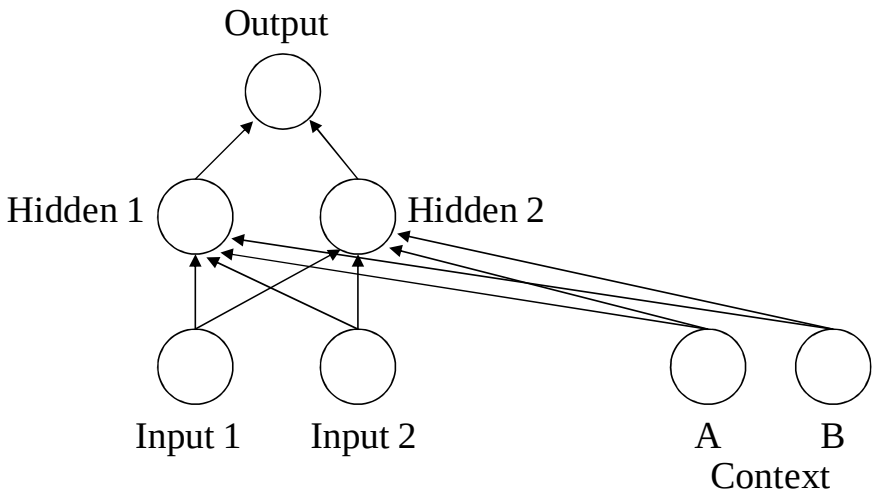


Figure 5

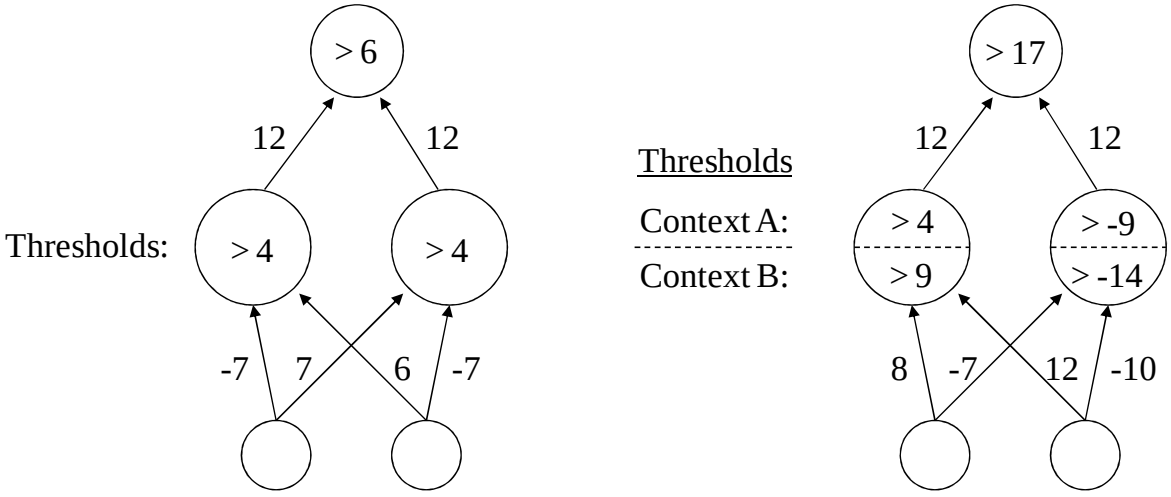

Thresholds:
> 6
12
12
> 4
> 4
-7
7
6
-7
Thresholds
Context A:
Context B:
> 17
12
12
> 4
> 9
> -9
> -14
8
-7
12
-10

Figure 6

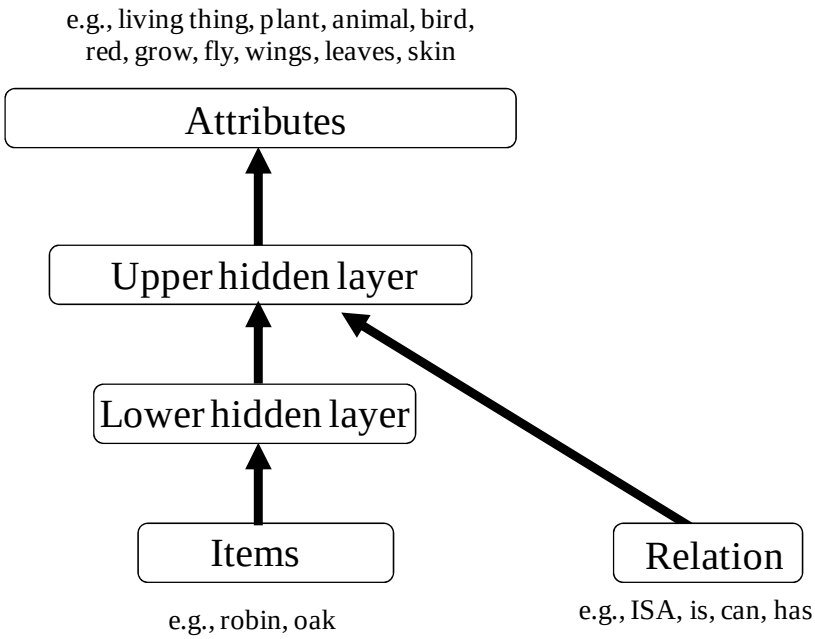

e.g., living thing, plant, animal, bird,
red, grow, fly, wings, leaves, skin
Attributes
Upper hidden layer
Lower hidden layer
Items
Relation
e.g., robin, oak
e.g., ISA, is, can, has

Figure 7

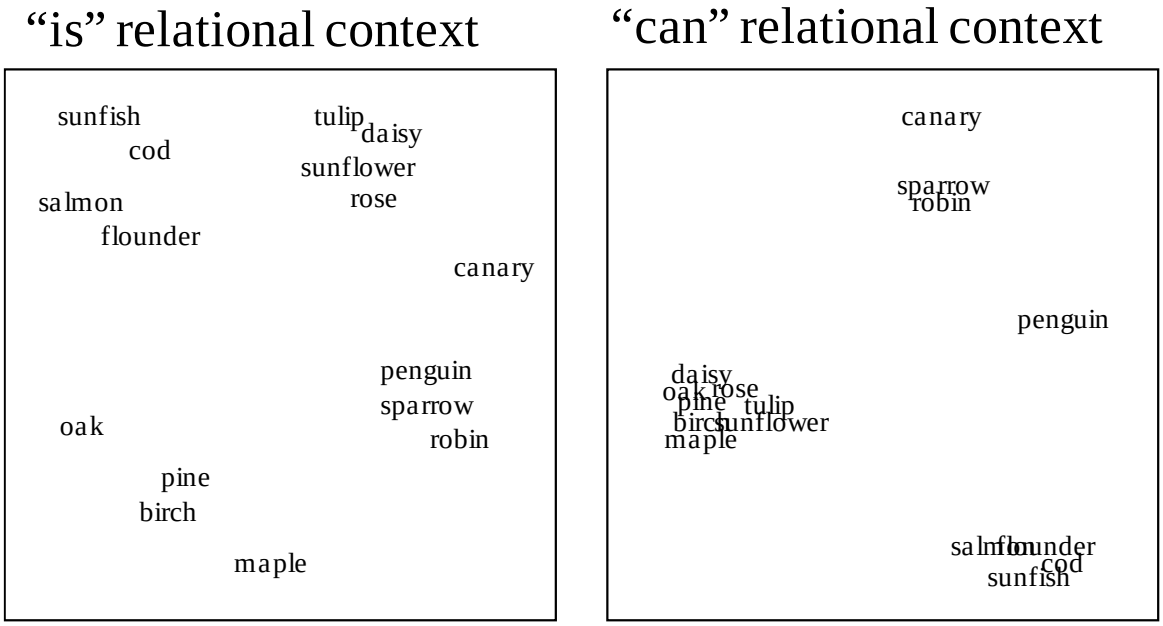

"is" relational context
sunfish
cod
salmon
flounder
tulip
daisy
sunflower
rose
canary
penguin
sparrow
robin
oak
pine
birch
maple
"can" relational context
canary
sparrow
robin
penguin
daisy
oak
rose
pine
tulip
birch
sunflower
maple
salmon
flounder
cod
sunfish

Figure 8

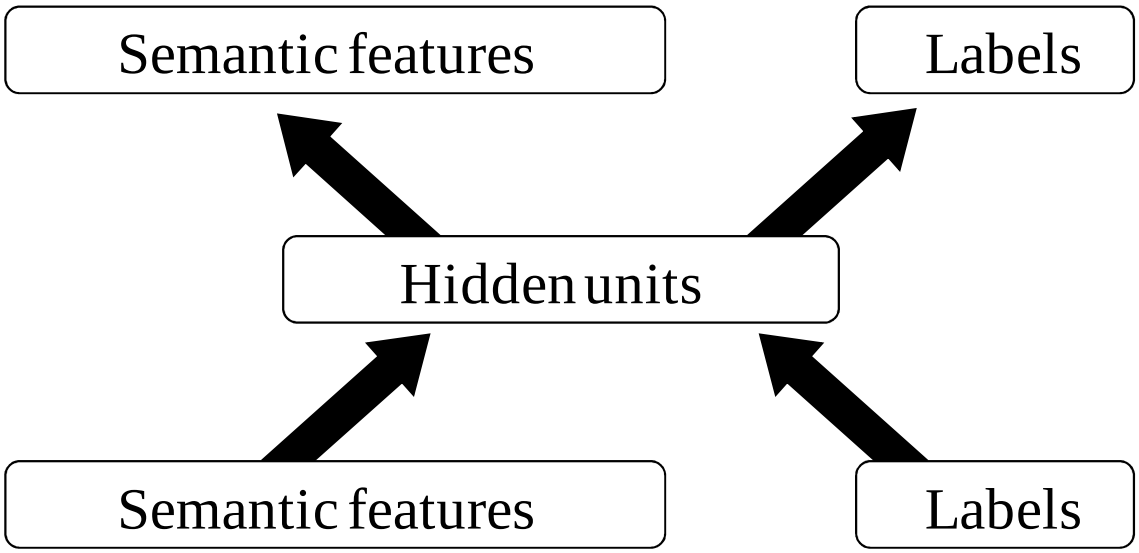

Semantic features
Labels
Hidden units
Semantic features
Labels

Figure 9

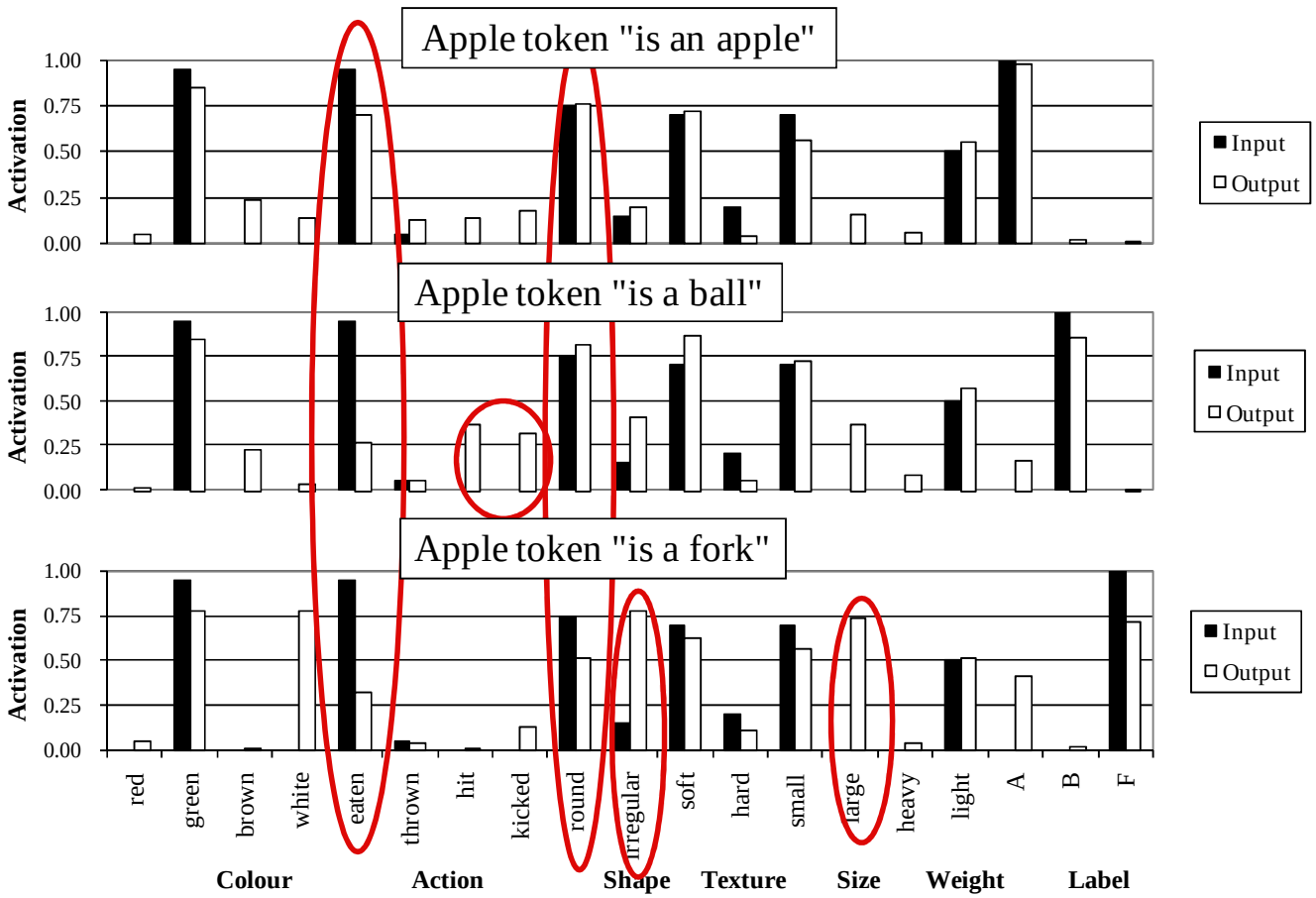

Apple token "is an apple"
Apple token "is a ball"
Apple token "is a fork"
Activation
1.00
0.75
0.50
0.25
0.00
Input
Output
red
green
brown
white
eaten
thrown
hit
kicked
round
irregular
soft
hard
small
large
heavy
light
A
B
F
Colour
Action
Shape
Texture
Size
Weight
Label

Figure 10