Metaphor as categorisation: a connectionist implementation

Running head: CONNECTIONIST MODEL OF METAPHOR AS CATEGORISATION

Metaphor as Categorisation: A Connectionist Implementation.

Michael S. C. Thomas

Neurocognitive Development Unit,

Institute of Child Health.

Denis Mareschal

Birkbeck College, University of London.

Address for correspondence:

Dr. Michael Thomas,
Neurocognitive Development Unit,
Institute of Child Health,
30, Guilford Street,
London WC1N 1EH
UK.
Email: M.Thomas@ich.ucl.ac.uk
Tel.: 0171 905 2747
Fax: 0171 242 7177

Running head: CONNECTIONIST MODEL OF METAPHOR AS CATEGORISATION

Metaphor as Categorisation: A Connectionist Implementation.

Abstract

A key issue for models of metaphor comprehension is to explain how in some metaphorical comparison <A is B>, only some features of B are transferred to A. The features of B that are transferred to A depend both on A and on B. This is the central thrust of Black's well known interaction theory of metaphor comprehension (1979). However, this theory is somewhat abstract, and it is not obvious how it may be implemented in terms of mental representations and processes. In this paper we describe a simple computational model of on-line metaphor comprehension which combines Black's interaction theory with the idea that metaphor comprehension is a type of categorisation process (Glucksberg & Keysar, 1990, 1993). The model is based on a distributed connectionist network depicting semantic memory (McClelland & Rumelhart, 1986). The network learns feature-based information about various concepts. A metaphor is comprehended by applying a representation of the first term A to the network storing knowledge of the second term B, in an attempt to categorise it as an exemplar of B. The output of this network is a representation of A transformed by the knowledge of B. We explain how this process embodies an interaction of knowledge between the two terms of the metaphor, how it accords with the contemporary theory of metaphor stating that comprehension for literal and metaphorical comparisons is carried out by identical mechanisms (Gibbs, 1994), and how it accounts for both existing empirical evidence (Glucksberg, McGlone, & Manfredi, 1997) and generates new predictions. In this model, the distinction between literal and metaphorical language is one of degree, not of kind.

Why use metaphor in language? Gibbs (1994) summarises three kinds of answers to this question (Fainsilber & Ortony, 1987; Ortony, 1975). First, the inexpressibility hypothesis suggests that metaphors allow us to express ideas that we cannot easily express using literal language. Second, the compactness hypothesis suggests that metaphors allow the communication of complex configurations of information to capture the richness of a particular experience. The use of literal language to communicate the same meaning would be cumbersome and inefficient. Third, the vividness hypothesis suggests that the ideas communicable via a metaphor are in fact richer than those we may achieve using literal language.

When we receive information coded in the form of a metaphor (e.g., not that Richard is brave, aggressive, and so on, but that Richard is a lion), how do we process such language to extract its vivid meaning? The traditional view in philosophy and linguistics was that language comprehension and production are built around literal language, that metaphorical language is both harder to comprehend (given that it is literally false – in our example, Richard is not a lion) and requires special processing mechanisms to decode. Although it is distinguished by its communicative advantages, metaphor was seen as a purely linguistic phenomenon (Grice, 1975; Searle, 1975). More recently, this view has been challenged on two grounds (e.g. Gibbs, 1994, 1996; Lakoff, 1993). First, it is claimed that metaphor is conceptual rather than linguistic. Second, it is claimed that metaphor is not an add-on to the more primary literal language processing system, but a key aspect of language itself, sharing the same kind of processing mechanisms. In this paper, we will be focussing on the second of these claims.

The argument that metaphor comprehension does not require special processing mechanisms has two strands (Gibbs & Gerrig, 1989). The first of these is

that on-line processing studies suggest that (with appropriate contextual support) metaphors and literal statements take the same amount of time to process (e.g. Inhoff, Lima, & Carroll, 1984; Ortony, Schallert, Reynolds, & Antos, 1978). This seems to rule out the possibility that metaphors are initially processed as literal statements, found to be false, and only then processed by metaphor-specific mechanisms. It does not, however, rule out the possibility that literal and metaphorical meanings of a sentence may be computed simultaneously and in parallel by separate mechanisms. The second strand suggests that literal language processing is no easier than metaphorical processing, given that both rely on a "common ground" between speaker and listener to comprehend what a given utterance means (Gibbs, 1994). That is, an apparently literal statement may well have an implicated meaning given a certain set of shared assumptions between speaker and listener. If both types of language involve similar problems, it makes sense to see them as engaging the same sort of mechanisms.

Black (1955, 1962, 1979) outlined three views of how the metaphor comprehension process may work. In the first of these, the substitution view, to understand the metaphorical comparison Richard is a lion, this comparison must initially be replaced by a set of literal propositions that fit the same context, e.g. Richard is brave, Richard is aggressive. In the comparison view, the metaphor is taken to imply that the two terms are similar to each other in certain (communicatively relevant) respects. For example, both Richard and the lion are brave, aggressive, and so forth. The intention of the comparison is to highlight these properties in the first term Richard. In effect, the comparison is shorthand for the simile Richard is like a lion. 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. In turn, this "parallel implication complex" may cause changes in our understanding of the vehicle in the comparison.

Although the interaction view has been described as "the dominant theory in the multidisciplinary study of metaphor" (Gibbs, 1994, p. 234), it has nevertheless been criticised for the vagueness of its central terms (ibid., p. 235). One of the key issues for psycholinguistic models of metaphor comprehension is to explain the nature of the interaction between topic and vehicle that constrains the emergent meaning of the comparison. Three main models have been proposed. These are the salience imbalance model (Ortony, 1979), the structural mapping model (Gentner, 1983; Gentner & Clements, 1988), and the class inclusion model (Glucksberg & 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 (that is, Richard falls into the class of brave, aggressive things whose prototypical member is lions). 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.

The view of metaphor as a form of categorisation seems perhaps most consistent with the claim that metaphor comprehension requires no special processes over and above literal comprehension. Both the salience imbalance model and the structural mapping model imply a property matching procedure which is engaged for non-literal comparisons (Glucksberg, McGlone, & Manfredi, 1997). Moreover, Glucksberg et al. (1997) have argued that the class inclusion theory is empirically distinguishable from these property matching models. Although literal comparisons are asymmetric (in that the similarity of two terms can be rated differently depending on the order of presentation; e.g. Tversky & Gati, 1982), class inclusion statements should be more than asymmetric, they should be non-reversible. The lion is Richard should make very much less sense than Richard is a lion, unless Richard happens to be a prototypical member of a category of which lion could also be a member. Secondly, Glucksberg et al. claimed that the topic and vehicle should make very different (though interactive) contributions to the metaphor's meaning, and that these contributions are predictable. While the vehicle provides the properties that may be attributed to the topic, the listener's familiarity with the topic constrains those properties that may be attributed to it. Glucksberg et al. primed comprehension of metaphorical comparisons by pre-exposure to either topic or vehicle. They predicted that only comparisons involving topics with few potentially relevant attributes, or vehicles with few properties available as candidate attributes, should benefit from pre-exposure. In their view, neither property-matching model should predict the non-reversibility or specific interactivity effects. Nevertheless, Glucksberg et al. (1997) found empirical support for both of their predictions.

The class inclusion model contrasts with Lakoff and colleague's theory that metaphors rely on established mappings between pairs of domains in long term

memory (Lakoff, 1987, 1990, 1993; Lakoff & Johnson, 1980; Lakoff & Turner, 1989). Thus comprehension of the metaphor this relationship is going nowhere proceeds via a pre-existing system of correspondences between the conceptual domains of LOVE and JOURNEY. The class inclusion theory on the other hand posits no such pre-existing metaphorical structures. In a comparison of the class inclusion and conceptual metaphor theories, McGlone (1996) determined that it was not yet possible to find conclusive evidence for either theory. McGlone presented four experiments, employing metaphor paraphrasing, comparison, and cued recall, the results of which he took to support the class inclusion theory over the conceptual metaphor theory. However, he admitted that the use of these off-line measures may not have tapped the use of conceptual metaphors during on-line interpretation. Evidence for the class inclusion model comes from the irreversibility of metaphors and related discourse phenomena (Glucksberg, 1991), while the primary evidence for the conceptual metaphor theory comes from the observed systematicity of idiomatic expressions in certain semantic domains. Lakoff (1993) has criticised the class inclusion model for its use of 'metaphorical attributive categories' to mediate metaphor comprehension. Thus the metaphor my job is a jail must be understood via appeal to the category 'restraining' things (of which jail is a prototypical member). However, the application of the term 'restraining' to the concept job itself is itself metaphorical. Yet Lakoff's own theory incurs the same problem in his use of the Invariance Principle, by which domains are linked in long term memory. Thus containers and categories, for instance, are linked in a particular way such that 'source domain interiors correspond to target domain interiors' (1993, p. 215). But the notion of the 'interior' of a container can only be metaphorically applied to the concept

category. In sum, it is premature to reject either of these theories at the current time. In what follows, we will be concentrating on the class inclusion theory.

In this paper, our aim is to propose a computational model of metaphor comprehension based on a categorisation device, as opposed to the property-matching device which would have to lie at the heart of a salience imbalance or a structural mapping model. Since our model will be based on a previously proposed mechanism of semantic memory, it exemplifies the idea that metaphor comprehension is not a "special" function of the language processing system. Indeed we suggest that within this mechanism, literal and metaphorical comparisons are distinguished only quantitatively, not qualitatively. The implemented model demonstrates in concrete terms how topic and vehicle interact in metaphor comprehension, addressing some of the vagueness in the interaction position. Lastly, we show how the model accounts for both of the empirical findings demonstrated by Glucksberg et al., and how it generates new predictions.

First however, we lay out the assumptions of the MPC (metaphor by pattern completion) model.

Assumptions of the model

The model builds on the following assumptions:

1. The aim of comprehension is the on-going development of a semantic representation, and that representation is feature-based.
2. The on-going semantic representation is continually monitored against expectations based on a common ground between listener and speaker. Specifically with regard to metaphor comprehension, the on-going semantic

representation is monitored for degree of expected meaning change. (It will be monitored in other ways for other non-literal communication).

3. Comparisons of the form <A is B> are class inclusion statements where the intended meaning is that A is a member of category B and so should inherit its attributes (Glucksberg & Keysar, 1990, 1993).
4. The meaning produced by a metaphor is the result of using a categorisation mechanism to transfer attributes from B to A when A is not in fact a member of B. However, membership of B is not all-or-nothing but depends on degree of featural overlap.
5. The categorisation mechanism is an autoassociative neural network. Category membership is established by the accuracy of reproduction of a novel input A to a network trained to reproduce exemplars of category B. The output of such a network is a version of A transformed to make it more consistent with B.
6. Metaphorical comparisons must exceed some expected level of semantic distortion (for a given context) to be interpreted as metaphorical. When a comparison is interpreted as metaphorical, not all feature changes induced in the topic A are accepted as the communicative intent of the comparison. More specifically, the accepted features of the comparison are those initially non-zero features of the topic A that are amplified by the transformation caused by the vehicle knowledge base B.
7. Metaphorical mappings caused by the topic may be learnt in the network storing the vehicle knowledge. The topic may become a (highly atypical) member of the vehicle category, so changing that category in long term memory. Thus metaphors may either be computed on-line or retrieved from long term memory.

Before describing the details of the model, we wish to expand on two of these assumptions, and situate our model with respect to previous connectionist models of metaphor comprehension. The first is the idea that meaning can be described as a set of features, or in connectionist terms, as a vector representation. Although there is a significant debate surrounding the legitimacy of feature vectors, much research has used vector-based semantic representations. For instance, connectionist models of word recognition which employ such representations have successfully captured a great deal of empirical data in both normal and impaired language processing (Plaut, McClelland, Seidenberg, & Patternson, 1996; Plaut & Shallice, 1993). Moreover, using a semantic priming paradigm, McRae, Cree, and McNorgan (1998) generated empirical predictions for the feature-based theory of lexical semantic representation and its main competitor, the hierarchical semantic network theory. Their results supported feature-based accounts, finding no evidence that priming proceeded via intervening superordinate nodes rather than shared feature sets. McRae et al. concluded that "lexical concepts are not represented as static nodes in a hierarchical system" (p. 681). Lastly, corpus-based approaches have demonstrated that valid measures of word meaning can be generated using vector based co-occurrence statistics of the context in which words appear (Lund & Burgess, 1996). This has led to new theories of the acquisition of word meaning per se (Landauer & Dumais, 1997). While there are certainly problems with vector-based accounts and their difficulty in representing conceptual structure, they are nevertheless an active theoretical approach to the representation of meaning.

The second assumption is that connectionist networks are a valid cognitive model of categorisation. Connectionist models have tended to take two approaches to categorisation (see e.g. Small, Hart, Nguyen, & Gordon, 1996). In one approach, the

network takes object features as inputs, and maps to category names as outputs. In the other, a network is trained to simply reproduce the object features for the category it is storing (a task known as "autoassociation"). Category membership is tested depending on the accuracy with which a novel input is reproduced. An accurate reproduction indicates a high probability of category membership. It is the latter approach we will be adopting for our model. This approach has been used previously in models of the acquisition of word meaning (Plunkett, Sinha, Mueller, & Strandsby, 1992) and of semantic memory (McClelland & Rumelhart, 1986; Small et al., 1996).

A number of previous researchers have exploited the soft multiple constraint satisfaction capabilities of connectionist systems to propose models which find systematic mappings between the two concepts of a metaphor. Some of these models build in complex pre-existing structure to represent the various concepts (e.g., Holyoak & Thagard, 1989; Hummel & Holyoak, 1997; Narayanan, 1999; Veale & Keane, 1992; Weber, 1994). Others have emphasised featural representations. Thus Sun (1995) showed how a network trained on a subset of metaphors relating items in the landscape to facial features (around the core metaphor, billboards are warts), could generalise this knowledge to produce plausible meanings for metaphors it had not seen (see also Chandler, 1991). In our model we minimise the weight attributed to structural relations in metaphor, focusing on the learnability of the concepts in a distributed system. Models that build in complex pre-existing structured representations entrust much of their performance to the precise nature of these representations, limiting their generality and robustness. We build in no a priori metaphor structures other than the ability of a system to select the knowledge with which it attempts to categorise a given input. However, the concepts learned by our

model do contain structure in the form of systematic (though probabilistic) co-variation of the features which define them.

The Metaphor by Pattern Completion (MPC) model

The model below is simple and is primarily intended to illustrate the "metaphor as categorisation" approach. Figure 1 illustrates the model architecture. A three-layer connectionist network is trained to autoassociate (reproduce across the output units) semantic vector representations of exemplars from a number of different categories. Each category knowledge base is stored across a different set of hidden units [Footnote 1]. Metaphor processing is modelled by inputting a semantic vector for the topic to the part of the network storing a category of which it is not a member (i.e., the vehicle). The output of the network is a semantic representation of the topic transformed to make it more consistent with the vehicle. To understand why this transformation should occur, we need to consider a property of connectionist networks known as pattern completion.

Insert Figure 1 about here.

Pattern completion is a property of connectionist networks that derives from their non-linear processing (Rumelhart & McClelland, 1986). A network trained to respond to a given input set will still respond adequately given noisy versions of the input patterns. For example, if an autoassociator is trained to reproduce the vector <0 1 0 0> and is subsequently given the input <.2 .6 .2 .2>, its output is likely to be much closer to the vector it 'knows', perhaps <.0 .9 .0 .0>. An input is transformed so as to make it more consistent with the knowledge that the network has been

previously trained on. The connection weights store the feature correlation information in previously experienced examples. If a partial input is presented to the network, it can use that correlation information to reconstruct the missing features.

When processing metaphors, the input is not a noisy version of a pattern on which the network has previously been trained, but an exemplar of another concept. The output is then a transformed version of the topic, changed to make it more consistent with the knowledge stored about the vehicle. Metaphorical meaning emerges as a result of deliberate misclassification. As we will shortly see, the way in which a network transforms an input depends on that input. In this way, the model captures the interactivity between the terms of the metaphor.

For this simple model, we chose a small set of features with which to describe the concepts. In order to generate knowledge bases for separate concepts, the network was trained to autoassociate exemplars of each concept. For simplicity, we restricted the model to the formation of A is B metaphors between three concepts: Apples, Balls, and Forks. Two of these could plausibly be used in a metaphorical comparison (e.g. the apple is a ball), one of them much less so (e.g. the apple is a fork).

The concepts were defined by a set of prototypical tokens representing different kinds of apples, balls, and forks that could be encountered in the individual's world (see Table 1). The network was not trained on the prototypes themselves, but on exemplars clustered around these prototypes. Exemplars were generated from each prototype by adding Gaussian noise (variance 0.15) to the original.

Insert Table 1 about here.

The exemplars for each concept formed three training sets used to develop the network's three prior-knowledge bases 'about' apples, balls, and forks. The existence of a prior knowledge base is a necessary feature of metaphor comprehension. Prior knowledge bases are analogous to Black's (1979) "implicative complex" and reflect an individual's personal experience with exemplars of each concept. The apple sub-network was trained to autoassociate patterns corresponding to 10 exemplars of each of three apple kinds (e.g., red, green, and rotten) for a total of 30 patterns. Similarly, the ball sub-network was trained to autoassociate 10 exemplars from three different kinds (for a total of 30 patterns). Finally, the fork sub-network was trained to autoassociate 10 exemplars from 1 kind (for a total of 10 patterns). Because there was only 1 kind of fork (as opposed to 3 kinds of both apples and balls), a single blank training pattern (zero input and output) was added to the fork training set to inhibit over-learning of the fork exemplars. All networks were trained with Backpropagation using the following parameter values: learning-rate: 0.1, momentum: 0.0, initial weight range: ±0.5. Each sub-network (knowledge-base) was trained for 1000 epochs. All reported results are averaged over n=10 replications using different random starting weights and concept exemplars.

After training, the network demonstrated prototype effects in each knowledge base. They responded most strongly to the prototypes for each category, despite never encountering them in training (cf. human performance, Posner & Keele, 1968). This suggests that the Apple, Ball, and Fork categories had been adequately learnt. Metaphors were processed by the redirection of information flow into one knowledge base or another. The role of the 'is' in the <A is B> metaphor is to trigger that redirection.

Results

Interaction between topic and vehicle

Figure 2 shows the transformation of the semantic features of an apple concept for the metaphor the apple is a ball. The input is an exemplar of apple close to its prototype kind and is presented to the network storing knowledge about balls. The effect of this metaphor is to produce as output a representation of apple in which the suitability for throwing, the hardness, and roundness features are exaggerated, while the edibility feature is reduced and the colour features become more ambiguous. Provided the context dependent threshold for semantic distortion is exceeded, this metaphor will be interpreted to mean that the apple is question is round, hard, and likely to be thrown.

Insert Figure 2 about here.

In Glucksberg and Keysar's class inclusion theory (1990), a metaphor highlights an underlying category of which both topic and vehicle are members (but the vehicle is the prototypical member). Thus my job is a jail highlights that job is a member of the underlying category {restraining things}. In the MPC model, one could see such a new inclusive category as emerging from the juxtaposition. That is, the features of A that are emphasised by processing in the B network define the category of which apple and ball are both members (but of which ball is the prototypical member): {hard round things that can be thrown}. This is a possible response to Lakoff's (1993) criticism that in the class inclusion theory, metaphor comprehension relies on unlikely pre-existing 'metaphoric attributive categories' (e.g. 'restraining things' in the above example) – in the current model, such attributive categories are newly created by the categorisation process itself.

Alternatively, we could describe this transformation in terms of Black's parallel implication complex. Either way, these modified features are a result of the interaction of the topic and vehicle. For example, note that despite the fact that 20 of the 30 Ball exemplars are soft beachballs, apple is still made to look harder rather than softer by this metaphor. This is because the apple is closer in size to a hard baseball than it is to a soft beachball. The semantic transformation is not a default imposition of ball features onto those of an apple, but an interaction between stored ball knowledge and the nature of the apple exemplar being presented to the ball sub-network. Thus the model offers an instantiation of Black's interactive theory of metaphor comprehension.

We can now attempt to formulate a clearer answer to the question of why in a metaphor <A is B>, some features of B should be transferred to A but not others. Let us assume that features x, y, and z tend to co-occur in exemplars of B. Transfer of feature z from B to A will occur only when features x and y are present in A. Concept A can 'key in' to a strong co-variance of features in B, thus triggering the pattern completion processes to transfer the additional feature z. Pattern completion would cause the set of features x, y, z to be completed in A. Such pattern completion is even more effective if z is already present to some extent in A, so that this feature need only be exaggerated. Metaphorical comparisons are thus used to exaggerate existing features of the topic.

The transfer of features also depends on the strictness of co-variance in exemplars of B. Thus, if x, y, and z always co-occur in B, A is highly likely to inherit feature z when it already possesses x and y. However, if there are some exemplars of B which have x and y but not z, transfer is less likely. It may only occur if A shares other features of the particular exemplars of B which have x, y, and z in common.

In terms of the communicative advantage of metaphor, this model accords most closely with the compactness hypothesis. That is, vehicles embody a co-variance of features which, so long as the topic can key into them, may be transferred to the topic as a whole. Figure 2 demonstrates that the transformation of the features of the topic is a subtle one – features are not all or nothing, but enhanced or attenuated. It may also be that transformations of meaning of this sort cannot be achieved by the use of literal language alone. Thus the model may also accord with the inexpressibility hypothesis.

The reversibility of metaphors

Glucksberg, McGlone, and Manfredi (1997) have claimed that metaphors are characterised by the property of non-reversibility, a property that only the class inclusion model can explain. The authors had subjects rate the sense of literal and metaphorical comparisons in original (sermons are sleeping pills) and reversed (sleeping pills are sermons) formats. The subjects also paraphrased the two versions. The experimenters judged the forward and reverse paraphrases for how much sense they made. The results showed that literal comparisons were far more reversible than metaphors. Glucksberg et al concluded that their data "provided strong support for the claim that metaphors and similes either lose or change meaning when reversed" (1997, p. 57).

Figure 3 shows the transformation for the metaphor the ball is an apple, the reverse of the metaphor shown in Figure 2. In Figure 3, the effect of comparing the ball to an apple is to exaggerate the softness and irregularity and edibility of the ball, whilst reducing its likelihood of being thrown, its size, and its roundness. The semantic effect of this metaphor is quite different from that in the previous case

despite the fact that the feature overlap of ball exemplars and apple exemplars defining the knowledge bases is the same in each case. The change in meaning between the forward and reverse metaphors, found in the empirical data, arises in the MPC model from the non-linear nature of its transformations. These transformations are not symmetrical.

Insert Figure 3 about here.

Glucksberg et al noted that literal similarity statements are asymmetric – the rated similarity changes with the order of presentation of two terms – and that property matching models can account for this asymmetry by rating properties of the first and second term differently. However, Glucksberg et al maintained that non-reversibility is different in kind than asymmetry, and that property-matching models such as the salience imbalance and the structural mapping model cannot account for non-reversibility. We see literal and metaphorical comparisons as lying on a continuum, just as category membership can be a graded rather than binary phenomenon. We have shown elsewhere that an architecture similar to the MPC model is able to account for the asymmetry in general similarity judgements (Thomas & Mareschal, 1997). Reversibility and asymmetry are also matters of degree. Support for this is provided by Sternberg, Tourangeau, and Nigro (1979) who found an inverse relationship between the similarity of two terms in a comparison and the aesthetic impact of that comparison. Metaphors are about having just the right amount of dissimilarity. The greater the dissimilarity, the greater the asymmetry.

Predictability of interactions

Glucksberg et al (1997) manipulated the ambiguity of vehicles and the number of potentially relevant attributes of topics in metaphorical comparisons. They primed comprehension of metaphors either by prior exposure of the topic or the vehicle. The results showed that when either ambiguity or number of potentially relevant attributes was constrained, subjects benefited from the prime, in terms of the time it took them to comprehend the metaphors. It is difficult to directly relate our current model to reaction time data, since we do not believe the simple mechanism depicted in our model is the only mechanism involved in metaphorical comprehension. Other more complex mechanisms may contribute to a comprehension response time. Nevertheless, we are able in our model to systematically alter aspects of the topic or vehicle and demonstrate how the interaction is affected.

Figure 4 shows the metaphorical comparison the apple is a fork. Where there is little overlap between the concepts, the resulting output shows no strongly activated features, only a weak activation of the characteristics of a fork. Comparisons involving a narrowly defined vehicle with little similarity to the topic produce a weak and non-interactive metaphor.

Insert Figure 4 about here.

Figure 5 shows the metaphorical comparison the ball is a fork for balls of various different colours. The results again show weak imposition of the fork's characteristics, except when the ball is of the same colour of the fork. In this case, the topic can key into the narrowly defined vehicle concept and evoke a stronger transformation.

Insert Figure 5 about here.

Figure 6 shows the metaphorical comparison the ball is an apple, again for balls of various different colours. Here the vehicle, apple, is more ambiguous than fork, in that it has more widely varying prototypes. The resulting transformation is thus more interactive – that is, it depends more on the particular features of the topic. Once more, when the topic keys into a particular co-variance in the vehicle (red and green apples are firm, rotten brown apples are soft), the nature of the transformation differs – brown balls are seen as softer as a result of this metaphor in contrast to red and green balls.

Insert Figure 6 about here.

Figure 7 shows the metaphorical comparison the apple is a ball, but now supplying contextual information to further specify the type of ball referred to in the vehicle. (This is implemented by providing a label to each type of ball during training). When the apple is compared to a small hard baseball, the transformation is very different to when it is compared to a large soft beachball. Nevertheless, both types of ball knowledge are represented over the same hidden units within the network.

Insert Figure 7 about here.

These effects show that under certain circumstances, the nature of the interaction between topic and vehicle is predictable. With regard to Glucksberg et al.'s data, we might suggest the following explanation. A topic which has many potentially relevant properties (e.g. life is a _) is less able to prime subsequent metaphors than a topic with few (e.g. temper is a _) since such a topic has many keys which could engage patterns of co-variant features in the vehicle. Subsequent interactions are therefore less predictable. A vehicle with a variety of sets of co-variant features (e.g. _ is an ocean) is a less effective prime that one with few (e.g. _ is a crutch) since it has more patterns which could be keyed into by the topic. Once more, the interaction is less predictable. (Examples from Glucksberg et al., 1997).

Further predictions

Our model makes the following testable predictions.

Two phenomena can be predicted on the basis of the way autoassociative networks generalise to novel patterns given their training set and the degree of training they have undergone. (1) A lack of variance in the exemplars of the vehicle category will reduce interactivity in metaphor comparisons - that is, it will produce the same transfer of attributes across a range of topics. (2) In the same way, over-trained or highly familiar vehicles will also generate less interactivity in metaphorical comparisons.

We have proposed an explicit example of how literal and metaphorical comparisons may involve the same type of processing mechanism. However, for a metaphorical comparison, the listener does not accept the full meaning change implied by the comparison, but accepts only features that have been enhanced. This suggests that there is feature change in a metaphorical comparison that is not reported

by the subjects. For example, in the metaphor my rock is a pet, we do not conclude that the rock is alive. But we predict that (3) given a metaphorical comparison, subjects will show delayed responses to questions about features of the topic which they would nevertheless not report as aspects of the metaphorical expression (e.g., for my rock is a pet, the question 'Is a rock animate or inanimate?'). Evidence for such implicit featural change would support the idea that the reported meaning of a metaphor is the tip of the iceberg of a process of featural enhancement that has much in common with literal language processing.

Discussion

The relation of literal to metaphorical comparisons

The MPC model uses a categorisation device to transfer attributes of the category onto a novel input. Categorisation causes a transformation of the input vector to make it more consistent with the category. Metaphor occurs when the novel input is not a member of the category to which it is applied. However, category membership is a graded notion and categories themselves have internal structure, having more or less typical members (Rosch, 1975).

If we see metaphor as categorisation, it only requires a small step to see literal and metaphorical categorisation as differing in degree rather than kind. A literal comparison involves categorisation of a novel input that is a member of the vehicle category. However, the novel item may be a highly prototypical member of the category. This defines one end of a continuum. The item may be a less typical member – still falling within the category but in some sense being less similar to it. A metaphorical comparison involves an input that has some similarities to the category but falls beyond the normal limits of the category. An anomalous comparison

involves an input that falls beyond the normal limits and has few if any similarities to the category.

We propose then, that literal and metaphorical comparisons are on a continuum of reducing similarity. However, importantly, we propose that literal and metaphorical comparisons are also distinguished by how the semantic distortion caused by the categorisation process is then handled. If the change in meaning of the topic caused by the semantic distortion exceeds (context dependent) expectations, the comparison is taken to be metaphorical, and the communicative intent is taken to refer only to the features of the topic that have been amplified by the transformation. If the threshold is low and we are told that Richard (whom we thought to be a man) is a lion, we are likely to view this claim as literally false and ask for clarification. A higher threshold will cause us to focus only on enhanced features of the topic distortion, viewing the statement as a metaphor. If the threshold expectation of meaning change is very high (for instance, the listener expects the speaker to convey brand new information), then the same statement can be taken as literal and all meaning change accepted as the communicative intent. Richard is a lion can be false, or it can tell us that a man we know, Richard, is a brave and aggressive man; or it can tell us that a particular lion has been named Richard (a name usually reserved for humans). The actual meaning is not derivable from the comparison alone, but depends on context. Similarly, before context definitively disambiguates the meaning, my job is a jail could be seen as incongruous (occupations can't be buildings), or it could tell us that I feel constrained by my job, or it could tell us that my daytime occupation is to act as a physical restraint for some sentient being.

Criticisms of semantic feature explanations of metaphor

The MPC model is based on simple semantic feature representations of concepts. Such representations have been criticised on a number of grounds as insufficient to explain the processes of metaphor comprehension. In this section, we consider a number of these criticisms. Criticisms 1-4 are from Gibbs (1994). Criticism 5 considers the importance of conceptual structure in metaphor comprehension.

Criticism 1: How can feature-based representations deal with semantically non-deviant representations that are nevertheless metaphorical, i.e. those that can have a valid literal interpretation? Response: Our response to this criticism is detailed in the previous section. Simple metaphor comprehension is a two-stage process involving both semantic distortion caused by the juxtaposition and context-dependent interpretation of that distortion.

Criticism 2: Feature-based representations seem insufficient to deal with the complexities of sophisticated metaphorical expressions. Response: At the moment, this criticism is certainly valid. However, it is also true that we do not know what a more realistic feature-based representation of meaning looks like. The representations in our model are undoubtedly too simple to deal with anymore than two term metaphors involving attribute mapping. We would expect more realistic and complex feature-based representations to support richer metaphorical distortions in a system following the same principles – that metaphor relies on pattern completion processes invoked through deliberate mis-classification.

Criticism 3: 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 she is frivolous - but lollipops themselves can be described as frivolous). Furthermore, features must not themselves be metaphorical. For example, in the metaphor the

legislative program was a rocket to the moon, we might think this implies that both are fast. But legislative programs and rockets are not fast in the same way. Response: One response for a feature-based account would be that semantic features are not lexical concepts. That is, in the previous example, a cluster of semantic features defines fast for the rocket, and a different cluster, though sharing many of the same features, defines fast for legislative program. Similarly, in the girl is a lollipop, the cluster of features which is enhanced in the representation of girl by the knowledge base for lollipop would share features with the cluster that defines the lexical concept "frivolous".

The notion that lexical concepts are made up of features is the essence of sub-symbolic representation. Features only appear as lexical concepts in our own model for ease of exposition. Thus 'hard' in our model might itself correspond to a set of lower level features, different groups of which would make up different versions of hardness. (See Harris, 1994, for an example of a connectionist model exhibiting sub-symbolic, context dependent meanings of a lexical concept). Clearly such an account needs to be fleshed out, but we don't believe that this criticism is a terminal one for feature-based representations.

Criticism 4: Feature overlap accounts do not explain why metaphors have directionality. Response: In the section entitled The Reversibility of Metaphors, we show how the model accounts for the directionality of metaphorical comparisons.

Criticism 5: Feature-based or vector representations cannot deal with relational structure in concepts. Gentner (1988) has shown that adults prefer topics and vehicles to be structurally related in metaphors. Response: The current model can only address part of the metaphor story, for more complex metaphors will necessarily involve semantic distortions at various levels of conceptual structure. Structured

representations are not easily implemented in connectionist systems. However, recent work in the connectionist modelling of analogy formation has shown how feature-based attributes may be dynamically bound to relational structure in a distributed network (Hummel & Holyoak, 1997). Such a network still exploits similarity-based processing and pattern completion in forming and retrieving analogies. Moreover, Henderson and Lane (1998) have shown that such dynamically bound representations may be learnt in a neural network architecture. We would make two claims. First, we believe that the approach of the MPC model is extendable to structured representations in a connectionist system (similar to that of Hummel & Holyoak, 1997). The principles of such a model would be the same: metaphor comprehension would rely on pattern completion and subsequent semantic distortion in a system designed for categorisation, in this case of structured concepts. Second, we believe it important to embed such future accounts in neurally plausible learning systems, which minimise the proportion of the theory that relies on arbitrary decisions about the nature of pre-existing structured representations (or, indeed, postulates representations with no apparently learnability at all).

An interesting avenue of research will be to explain why children show a shift in preference from attribute mapping to relational mapping during development. Thus far, we have applied the MPC model only to the emergence of the distinction between literal and metaphorical similarity in young children, based on the maturity of their semantic representations (Thomas, Mareschal, & Hinds, under revision). In future work we hope to explore extensions of the model to include relational structure, and therefore investigate the developmental shift to more complex metaphors.

Conclusion

In this paper we have introduced a simple and predominantly illustrative model of how metaphor comprehension may be explained as a form of categorisation (Glucksberg & Keysar, 1990, 1993). We have offered the beginnings of an answer to the thorny question of why certain attributes are transferred from the vehicle to the topic in a metaphorical comparison, but not others. The answer was in terms of attributes that the topic possesses which key into co-variances of features in the vehicle, and which pattern completion processes in a neural network allow to be transferred to the topic. This is an essentially interactive account, in line with Black's favoured view of metaphor comprehension (1979). The model is able to offer accounts for recent empirical evidence on the non-reversibility of metaphorical expressions, and the nature of the interaction between topic and vehicle (Glucksberg et al., 1997), as well as generating further testable predictions.

Lastly, in wider theoretical terms, the model conforms to the notion that metaphor comprehension requires no special processes over and above literal language comprehension, by suggesting that metaphorical language and literal language are different points on a continuum of meaning change. Literal and metaphorical statements update comprehension in a different way.

References

Black, M. (1955). Metaphor. Proceedings of the Aristotelian Society, 55, 273-294.

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

Black, M. (1979). More about metaphor. In A. Ortony (Ed.), Metaphor and Thought (pp. 19-43). Cambridge, England: Cambridge University Press.

Fainsilber, L. & Ortony, A. (1987). Metaphorical uses of language in the expression of emotion. Metaphor and Symbolic Activity, 2, 239-250.

Chandler, S. R. (1991). Metaphor comprehension: A connectionist approach to implications for the mental lexicon. Metaphor and Symbolic Activity, 6, 227-258.

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. Cognitive Science, 7, 155-170.

Gentner, D. (1988). Metaphor as structure mapping: The relational shift. Child Development, 59, 47-59.

Gentner, D. & 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, pp. 307-358). Orlando, FL: Academic Press.

Gibbs, R. W. (1994). The poetics of mind. Cambridge University Press.

Gibbs, R. W. (1996). Why many concepts are metaphorical. Cognition, 61, 309-319.

Gibbs, R. W. & Gerrig, R. (1989). How context makes metaphor comprehension seem "special". Metaphor and Symbolic Activity, 4, 154-158.

Glucksberg, S. (1991). Beyond literal meanings: The psychology of allusion. Psychological Science, 2, 146-152.

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

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

Glucksberg, S., McGlone, M. S., & Manfredi, D. (1997). Property attribution in metaphor comprehension. Journal of Memory and Language, 36, 50-67.

Grice, H. P. (1975). Logic and conversation. In P. Cole and J. Morgan (Eds.), Syntax and semantics: Vol. 3. Speech acts (pp. 41-58). New York: Academic Press.

Harris, C. L. (1994). Back propagation representations for the rule-analogy continuum. In J. A. Barnden & K. J. Holyoak (Eds.), Advances in connectionist and neural computation theory, Vol. 3: Analogy, metaphor, and reminding (282-326). Norwood, NJ: Ablex.

Henderson, J. & Lane, P. (1998). A connectionist architecture for learning to parse. In Proceedings of COLING-ACL'98 (pp. 531-537).

Holyoak, K. J. & Thagard, P. (1989). Analogical mapping by constraint satisfaction. Cognitive Science, 13, 295-355.

Hummel, J. E. & Holyoak, K. J. (1997). Distributed representations of structure: A theory of analogical access and mapping. Psychological Review, 104, 427-466.

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

Lakoff, G. (1987). Women, fire, and dangerous things: What categories reveal about the mind. Chicago: University of Chicago Press.

Lakoff, G. (1990). The invariance hypothesis: Is abstract reason based on image-schemas? Cognitive Linguistics, 1, 39-74.

Lakoff, G. (1993). The contemporary theory of metaphor. In A. Ortony (Ed.) Metaphor and Thought (2nd Ed.). Cambridge: Cambridge University Press.

Lakoff, G. & Johnson, M. (1980). Metaphors we live by. Chicago: University of Chicago Press.

Lakoff, G. & Turner, M. (1989). More than cool reason: A field guide to poetic metaphor. Chicago: University of Chicago Press.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychological Review, 104, 211-240.

Lund, K. & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. Behaviour Research Methods, Instruments and Computers, 2, 203-208.

McClelland, J. L. & Rumelhart, D. E. (1986). A Distributed Model of Human Learning and Memory. In J. L. McClelland, D. E. Rumelhart, & The PDP Research Group, Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Vol. 2: Psychological and Biological Models, (pp. 170-215). Cambridge, MA: MIT Press.

McGlone, M. S. (1996). Conceptual metaphors and figurative language interpretation: Food for thought? Journal of Memory and Language, 35, 544-565.

McRae, K., Cree, G. S., & McNorgan, C. (1998). Semantic similarity priming without hierarchical category structure. In Proceedings of the 20th Annual Meeting of the Cognitive Science Society, (pp. 681-686). Erlbaum.

Narayanan, S. (1999). Moving right along: A computational model of metaphoric reasoning about events. Proceedings of the National Conference on Artificial Intelligence, Orlando Florida, July 1999.

Ortony, A. (1975). Why metaphors are necessary and not just nice. Educational Theory, 25, 45-53.

Ortony, A. (1979). Beyond literal similarity. Psychological Review, 86, 161-180.

Ortony, A. (1993). The role of similarity in similes and metaphors. In A. Ortony (Ed.), Metaphor and Thought 2nd Ed., (pp. 342-356). Cambridge, England: Cambridge University Press.

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

Plaut, D., McClelland, J. L., Seidenberg, M., & Patterson, K. (1996). Understanding normal and impaired reading: Computational principles in quasi-regular domains. Psychological Review, 103, 56-115.

Plaut, D. & Shallice, T. (1993). Deep dyslexia: A case study of connectionist neuropsychology. Cognitive Neuropsychology, 10, 377-500.

Plunkett, K., Sinha, C., Mueller, M. F., & Strandsby, O. (1992). Symbol grounding or the emergence of symbols? Vocabulary growth in children and a connectionist net. Connection Science, 4, 293-312.

Rosch, E. (1975). Cognitive representations of semantic categories. Journal of Experimental Psychology: General, 104, 192-223.

Rumelhart, D. E., & McClelland, J. L. (1986). Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Vol. 1: Foundations. Cambridge, MA: MIT Press.

Posner, M. & Keele, S. (1968). On the genesis of abstract ideas. Journal of Experiment Psychology, 77, 353-363.

Searle, J. (1975). Indirect speech acts. In P. Cole and J. Morgan (Eds.), Syntax and semantics: Vol. 3. Speech acts (pp. 59-82). New York: Academic Press.

Small, S. L., Hart, J., Nguyen, T., & Gordon, B. (1996). Distributed representations of semantic knowledge in the brain: Computational experiments using feature based codes. In J. Reggia, E. Ruppin, & R. S. Berndt, Neural modelling of brain and cognitive disorders. World Scientific.

Sternberg, R. J., Tourangeau, R., & Nigro, G. (1979). Metaphor, induction, and social policy: The convergence of macroscopic and microscopic views. In A. Ortony (Ed.), Metaphor and Thought, (pp. 277-306). Cambridge, England: Cambridge University Press.

Sun, R. (1995). A microfeature based approach towards metaphor interpretation. In Proceedings of the 14th International Joint Conference on Artificial Intelligence, Montreal, Canada, pp. 424-429.

Thomas, M. S. C. & Mareschal, D. (1997). Connectionism and psychological notions of similarity. In M. G. Shafto & P. Langley (Eds.), Proceedings of the 19th annual conference of the Cognitive Science Society (pp. 757-762). Erlbaum.

Thomas, M. S. C., Mareschal, D., & Hinds, A. C. (under revision). A connectionist account of the emergence of the Literal-Metaphorical-Anomalous distinction in young children.

Tversky, A. & Gati, I. (1982). Similarity, separability, and the triangle inequality. Psychological Review, 89, 123-154.

Veale, T. & Keane, M. T. (1992). Conceptual scaffolding: A spatially founded meaning representation for metaphor comprehension. Computational Intelligence, 8, 494-519.

Weber, S. H. (1994). A structured connectionist model of figurative adjective noun combinations. In J. A. Barnden & K. J. Holyoak (Eds.), Advances in connectionist and neural computation theory, Vol. 3: Analogy, metaphor, and reminding. Norwood, N. J.: Ablex Publishing Corp.

Footnotes

Footnote 1. The use of separate banks of hidden units is not a necessary assumption of the model. ‘Soft’ modularity of knowledge bases can be achieved by using input and output labels to index each concept during training and categorisation.

Tables

Table 1: Prototype feature sets for each category.

Feature sets	Colour				Actions		Shape			Texture		Size	
Concepts	Red	Green	Brown	White	Edible	Thrown	Round	Irregular	Pointed	Soft	Hard	Hand Sized	Lap Sized
Apples													
Red	1	0	0	0	1	.2	.8	.3	0	.3	.7	1	0
Green	0	1	0	0	1	.2	.8	.3	0	.3	.7	1	0
Rotten	0	0	1	0	0	.2	.8	.3	0	1	0	1	0
Balls													
Baseballs	0	0	0	1	0	1	1	0	0	0	1	.9	.1
Beachballs													
Red	1	0	0	0	0	1	1	0	0	1	0	.1	.9
Green	0	1	0	0	0	1	1	0	0	1	0	.1	.9
Forks													
Fork	0	0	0	.9	0	.1	0	0	1	0	1	.7	.3

Figure Captions

Figure 1. Architecture of the MPC model.

Figure 2. Transformations of semantic features by a metaphorical comparison (Topic/Input = Apple; Vehicle/Network = Ball).

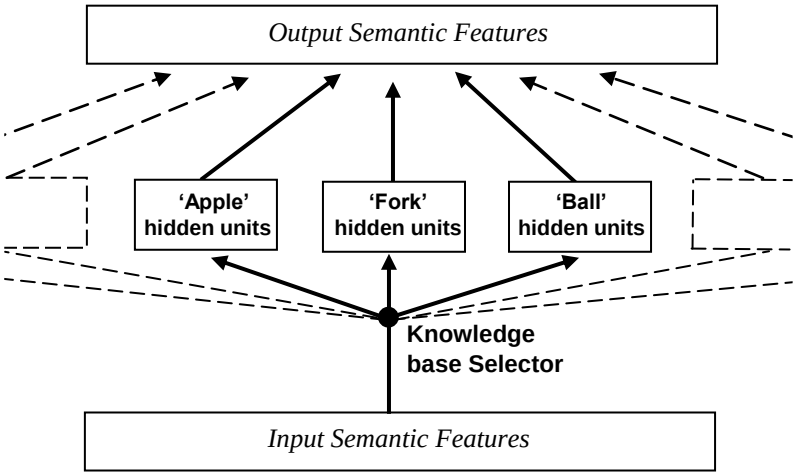
Figure 3. The non-reversibility of metaphorical comparisons (see text).

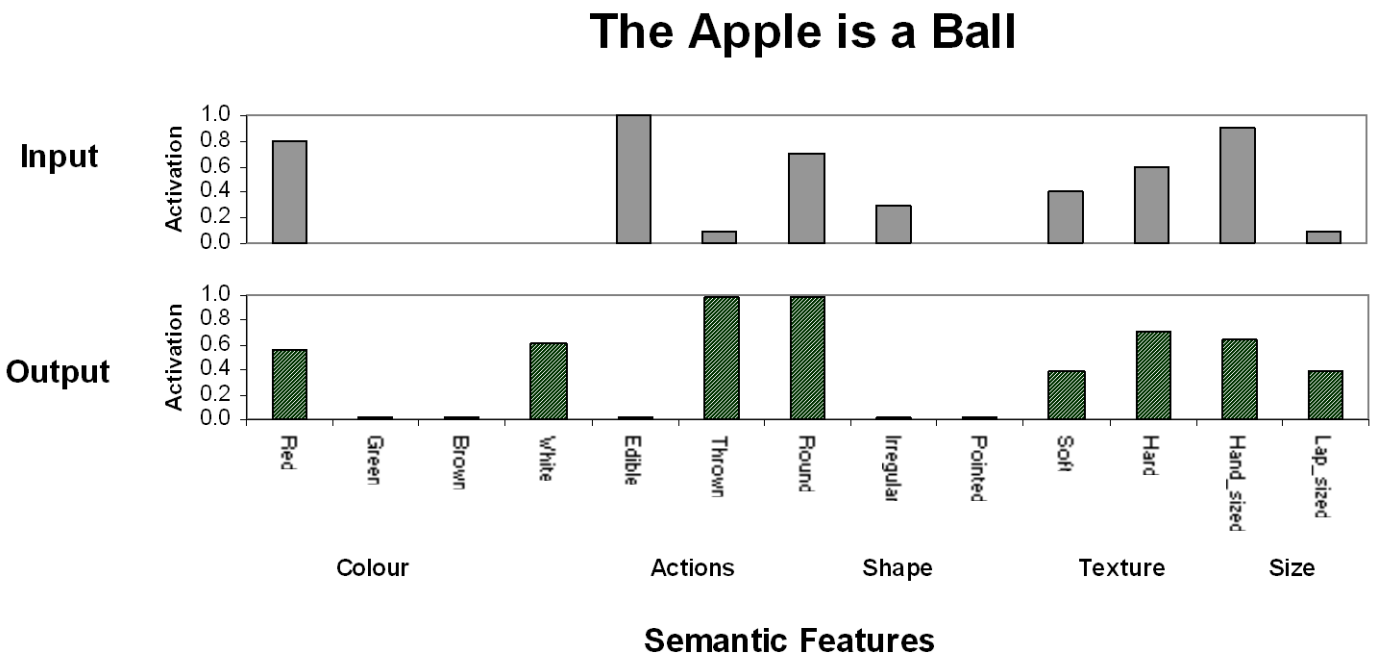
Figure 4. When metaphors fail: interactions between topic and vehicle (see text).

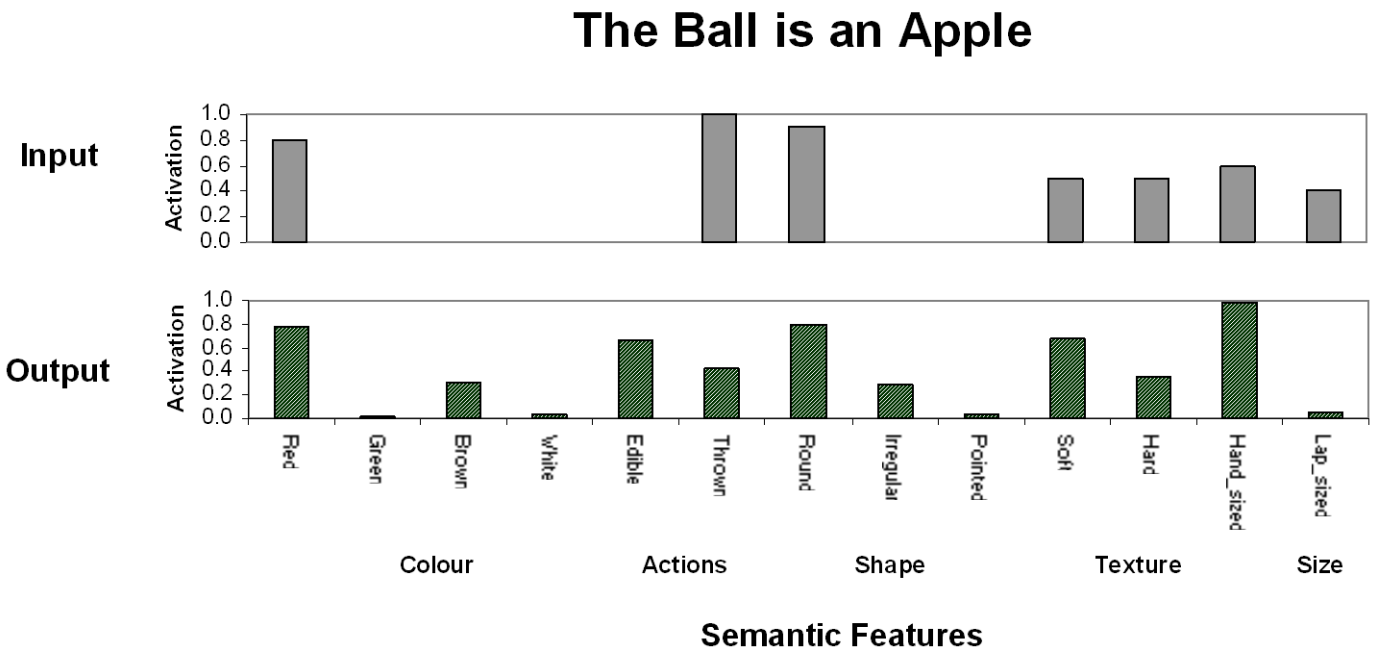
Figure 5. The role of the topic in determining the interaction between topic and vehicle (see text).

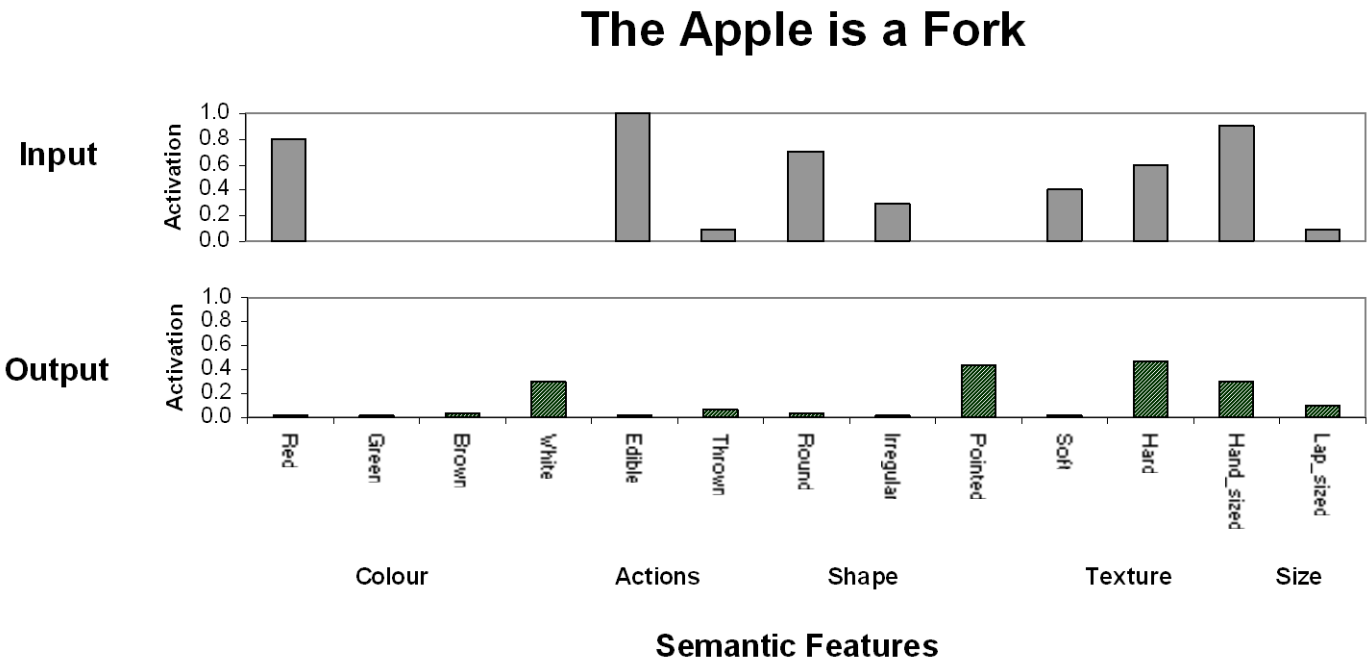
Figure 6. The role of the topic in determining the interaction between topic and vehicle (see text).

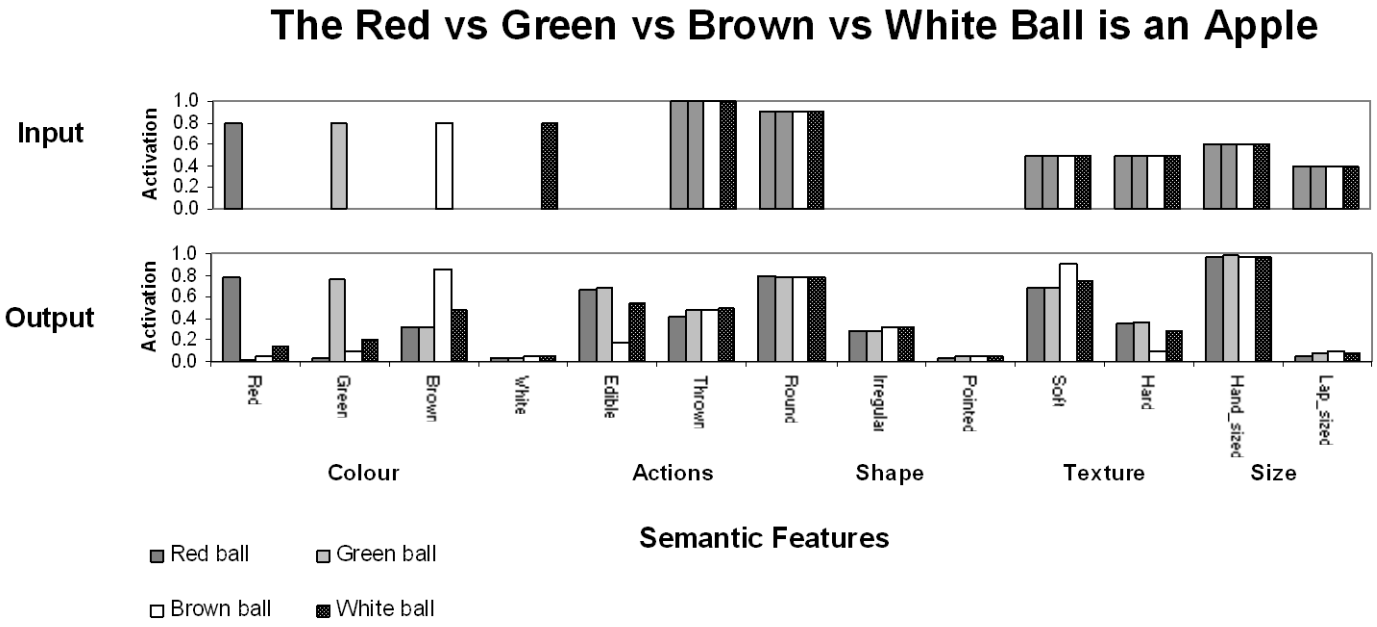
Figure 7. The role of the vehicle in determining the interaction between topic and vehicle (see text).

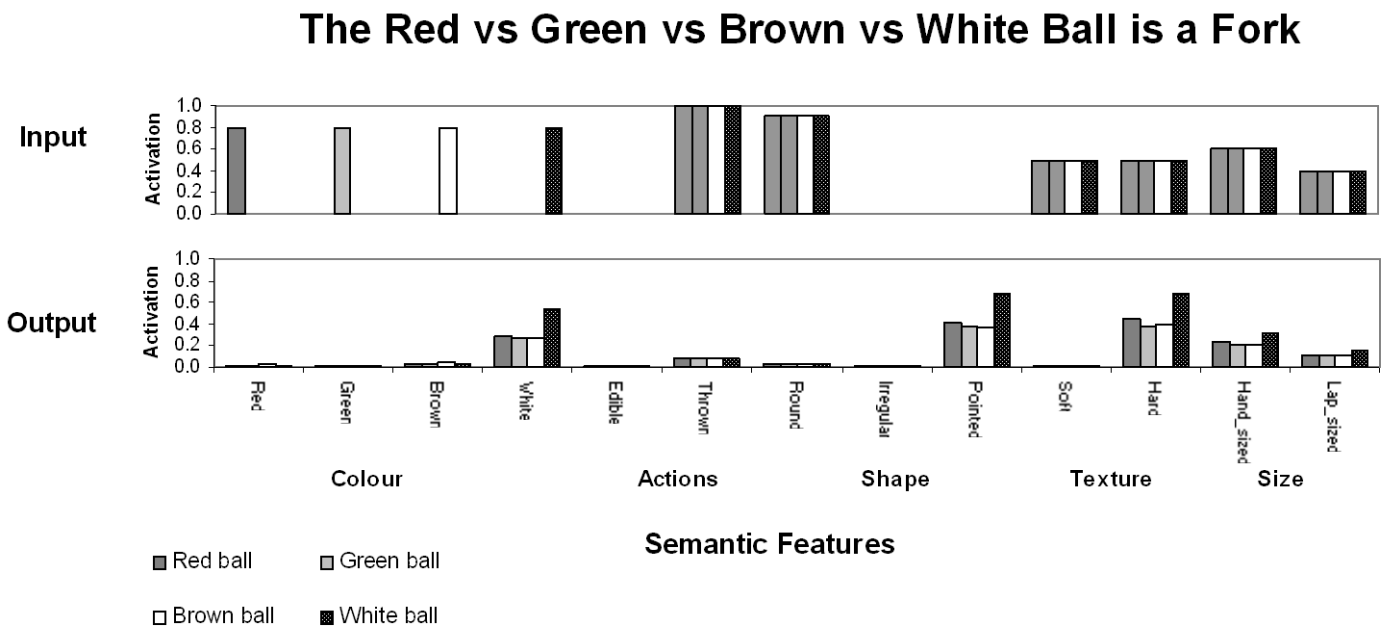

Output Semantic Features
'Apple' hidden units
'Fork' hidden units
'Ball' hidden units
Knowledge base Selector
Input Semantic Features

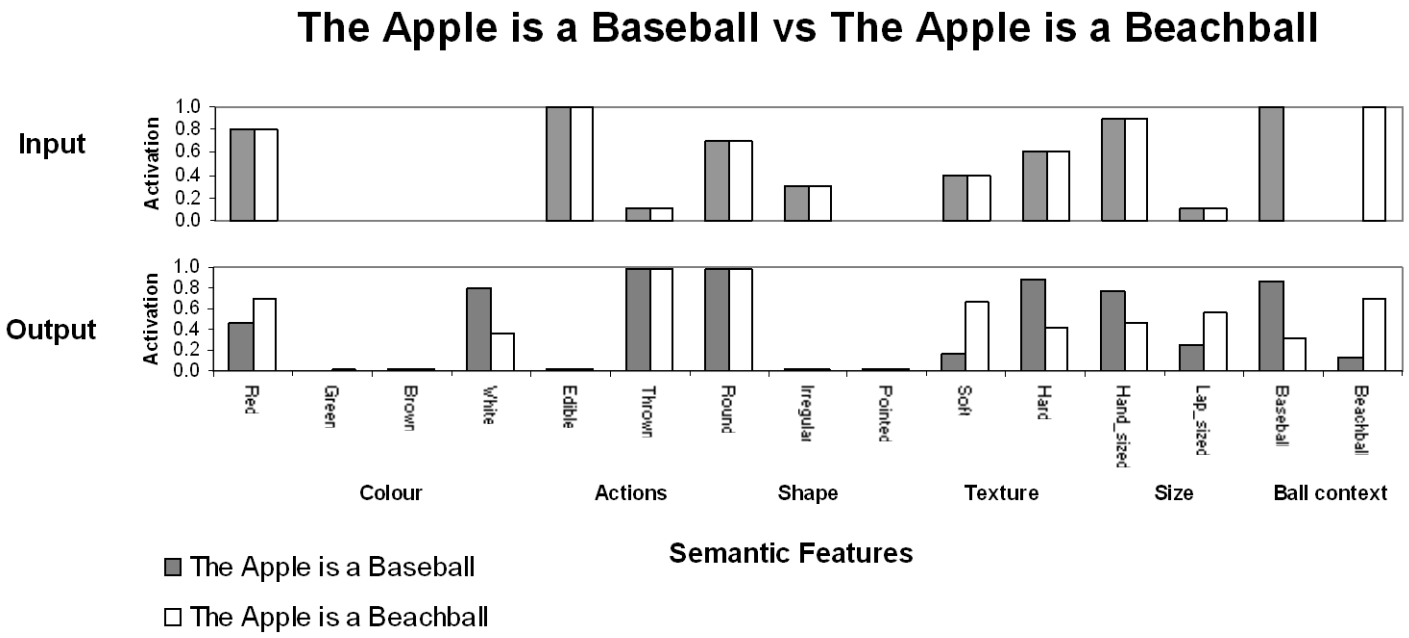

The Apple is a Ball
Input
Output
Activation
1.0
0.8
0.6
0.4
0.2
0.0
Red
Green
Brown
White
Edible
Thrown
Round
Irregular
Pointed
Soft
Hard
Hand_sized
Lap_sized
Colour
Actions
Shape
Texture
Size
Semantic Features


The Ball is an Apple
Input
Output
Activation
1.0
0.8
0.6
0.4
0.2
0.0
Red
Green
Brown
White
Edible
Thrown
Round
Irregular
Pointed
Soft
Hard
Hand_sized
Lap_sized
Colour
Actions
Shape
Texture
Size
Semantic Features


The Apple is a Fork
Input
Output
Activation
1.0
0.8
0.6
0.4
0.2
0.0
Red
Green
Brown
White
Edible
Thrown
Round
Irregular
Pointed
Soft
Hard
Hand_sized
Lap_sized
Colour
Actions
Shape
Texture
Size
Semantic Features


The Red vs Green vs Brown vs White Ball is an Apple
Input
Output
Activation
1.0
0.8
0.6
0.4
0.2
0.0
Red
Green
Brown
White
Edible
Thrown
Round
Irregular
Pointed
Soft
Hard
Hand_sized
Lap_sized
Colour
Actions
Shape
Texture
Size
Semantic Features
Red ball
Green ball
Brown ball
White ball


The Red vs Green vs Brown vs White Ball is a Fork
Input
Output
Activation
1.0
0.8
0.6
0.4
0.2
0.0
Red
Green
Brown
White
Edible
Thrown
Round
Irregular
Pointed
Soft
Hard
Hand_sized
Lap_sized
Colour
Actions
Shape
Texture
Size
Semantic Features
Red ball
Green ball
Brown ball
White ball


The Apple is a Baseball vs The Apple is a Beachball
Input
Output
Activation
1.0
0.8
0.6
0.4
0.2
0.0
Red
Green
Brown
White
Edible
Thrown
Round
Irregular
Pointed
Soft
Hard
Hand_sized
Lap_sized
Baseball
Beachball
Colour
Actions
Shape
Texture
Size
Ball context
Semantic Features
The Apple is a Baseball
The Apple is a Beachball