Connectionism and Psychological Notions of Similarity

Michael S. C. Thomas
Department of Psychology
King Alfred's College
Sparkford Road, Winchester
SO22 4NR, UK
michael.thomas@psy.ox.ac.uk

Denis Mareschal
Department of Psychology
Exeter University
Perry Rd., Exeter
EX4 4QG, UK
d.mareschal@exeter.ac.uk

Abstract

Kitcher (1996) offers a critique of connectionism based on the belief that connectionist information processing relies inherently on metric similarity relations. Metric similarity measures are independent of the order of comparison (they are symmetrical) whereas human similarity judgments are asymmetrical. We answer this challenge by describing how connectionist systems naturally produce asymmetric similarity effects. Similarity is viewed as an implicit by-product of information processing (in particular categorization) whereas the reporting of similarity judgments is a separate and explicit meta-cognitive process. The view of similarity as a process rather than the product of an explicit comparison is discussed in relation to the spatial, feature, and structural theories of similarity.

Introduction

Connectionist models of cognitive processing have been criticized for their apparent reliance on a notion of psychological similarity that empirical evidence has demonstrated to be flawed (Kitcher, 1996). This argument is based on the belief that connectionist information processing relies on metric distance measures of similarity. Whether the similarity occurs at the level of the input representation, the hidden unit representation, or the output representation, proximal tokens in a multi-dimensional space (defined by the characteristics of the task) are processed similarly. This, argues Kitcher, is necessarily wrong since metric distance measures have been ruled out as plausible models of psychological similarity.

In this paper, we argue that non-metric similarity measures do arise naturally out of connectionist information processing. These measures are based on functional transformations and are not constrained to obey the metric axioms of Minimality, Symmetry, and the Triangle Inequality. Therefore, they are immune to the objection that psychological similarity does not itself appear to obey the metric axioms (Kitcher, 1996; Tversky, 1977). We will suggest that there are two similarity processes in the cognitive system. One is non-metric and arises naturally from the functional transformation properties of non-linear connectionist information processing. The other can be metric and is constructed from the outcome of a prior and inevitable non-metric phase of processing. The non-metric component is implicit and not accessible to meta-cognitive processes. The metric component is only engaged when the evaluation of similarity has to be made explicit (e.g., it has to be communicated) such as in a similarity judgment task. That is, the initial comparison is implemented by a non-metric transformation; the requirement to make a similarity judgment introduces a further metric comparison process. We will argue that under certain conditions, functional transformation measures can generate behavior similar to metric distance measures, and hence that metric distance measures can offer an approximate description of the processes underlying similarity judgments.

The rest of this paper proceeds as follows. First, we present Kitcher's (1996) argument in more detail and discuss asymmetry as a counter example to metric distance measures of similarity. Then we present the Transformational Function Similarity (TFS) measure and discuss how it overcomes the asymmetry problems of metric measures. Finally, we discuss how a metric comparison measure can be constructed from the products of prior TFS stage.

Connectionism and metric similarity

A recent attack on connectionist information processing as a model of cognition has focused on the question of how information is processed in a network (Kitcher, 1996). Kitcher begins by unpacking Churchland and Sejnowski's (1992) characterization of activation patterns in a network in terms of vectors. The implication of this characterization is that the activation patterns define a multidimensional vector space that naturally supports metric distance measures. The similarities between objects are reflected by the distance between the positions their representations occupy in activation space. However, Tversky (1977) has identified a number of ways in which psychological notions of similarity do not appear to accord with predictions of a metric model of similarity. Although his efforts to show that Minimality and the Triangle Inequality do not hold for human similarity judgments may be inconclusive, Symmetry certainly does not hold in human similarity judgments (Hahn & Chater, 1996). As a result, we will focus our discussion on the notion of symmetry in psychological similarity.

Symmetry in this context is taken to mean that similarity judgments are commutative. In other words:

$$aSb = bSa \qquad \text{where } xSy \text{ is the similarity of x to y} \quad (1)$$

Let *a* and *b* be two tokens that can be described as occupying positions in a metric space. Then, the similarity relation is the same whatever the order of comparison. Studies requiring subjects to rate the similarity of a pair of items suggests that this is not the case with psychological

notions of similarity (e.g., Tversky & Gati, 1978). For example, when asked to compare pairs of concepts, subjects readily rated North Korea as being more similar to Red China than Red China was to North Korea. In short, reported similarity seems to change according to the order of the comparison.

Solutions to the asymmetry problem

Most theories of similarity have grappled with the asymmetry problem. The spatial theory of similarity (e.g., Rips, Shoben, and Smith, 1973; Rumelhart and Abrahamson, 1973) envisages concepts as points in a multi-dimensional space. The similarity between two concepts corresponds to their distance (e.g., Euclidean distance) apart in this space. This theory can account for asymmetric comparisons provided each concept is given a *bias* (Nosofsky, 1991). The bias relates to how easy it is to process a given concept. The direction of travel between the concepts in similarity space (corresponding to the order of the comparison) interacts with their respective biases. If the two concepts have different biases, the similarity will be different depending on the direction of travel. The feature theory of similarity (Tversky, 1977) measures the similarity between two concepts as some function of the number of features they have in common and the number on which they differ. This theory can account for the asymmetry by proposing that concepts have features with different *salience*. The concepts will be judged more similar if the features that the concepts have in common have a higher salience in the second term of the comparison (Ortony, Vondruska, Foss, and Jones, 1985). The structural alignment theory of similarity (Markman and Gentner, 1993; Medin, Goldstone, and Gentner, 1993) measures the similarity between two concepts depending on how well the structures of each concept can map onto one another. This theory can account for the asymmetry as long as the *coherence* of the structures of the concepts is taken into account (Gentner and Bowdle, 1994). Coherence is defined as the degree of systematicity a concept possesses. A coherent concept will have many "causal or explanatory connections" (Gentner and Bowdle, 1994, p. 352). Similarity judgments are higher if the more coherent concept is the second term in a comparison.

In all these theories, the basic measure of similarity is symmetrical. Asymmetries are derived by introducing additional factors, such as *bias*, *salience*, or *coherence*. Tversky introduces the idea that some features are more distinctive than others. Ortony et al and Gentner and Bowdle seek to capture the notion that similarity comparisons are psychologically informative. Nevertheless, the mechanisms for asymmetry lie in extensions of the basic symmetrical comparison procedures. A more parsimonious solution would derive the asymmetry as a consequence of the basic mechanism by which similarity was computed. As we shall see below, passing an input vector through a connectionist autoassociator does just that.

With this debate in mind, we can begin to reassess Kitcher's (1996) critique. Generally, there appears to be no entirely satisfactory solution to the asymmetry problem. The first point we might make then, is that the problem is not unique to connectionist information processing and therefore should not be used to single out connectionist approaches in particular for criticism. However, the ultimate answer to Kitcher's argument would be to produce a connectionist model that employed vector-coding and yet showed realistic, non-metric similarity judgments. This is exactly what we propose to do. The key is to move away from the idea of similarity as the outcome of a direct comparison procedure and to move towards the idea of similarity as a process whose outcome can only be reported in a post hoc fashion. The similarity process does not rely on placing the comparative elements in some metric relation to each other. Only the post hoc reporting (or explicit access) of similarity requires the establishment of a metric relation.

One way to understand this distinction is to think of the mind as a modular information processing system. As information passes through a module it is processed (or transformed) and passed on to the next module. This next module takes the transformed information as input and continues to process the information further. Note that the second module does not need to know anything about the nature of the previous transformation. The system as a whole continues to function without any need to relate explicitly the outcome of a process (the transformed information) with the initial state of the information prior to processing by the first module. However, some meta-process or control-process wishing to evaluate the functioning of the first module can do so by sampling and comparing the input information to the resulting output information. We want to suggest that similarity is related to the way in which information is processed (transformed) whereas the reporting of similarity judgments is a meta-cognitive process requiring the explicit comparison of information prior and subsequent to processing by the cognitive system.

The rest of this paper will describe how such similarity arises naturally from connectionist information processing through a process of selective dimensional distortion of the input vectors. The degree to which distortion occurs is inversely related to the similarity between the input vector and the contextualized knowledge stored in the network.

Transformational Function Similarity

Feedforward connectionist networks implement a transformation function from an input space to an output space. The dimensionality of the input and output spaces are defined by the task domain. Consider the set of networks for which the input and output space are of the same dimensionality. Such networks can be seen as engines that twist and distort the metric relations of the input space. Autoassociators are a subset of this set of networks for which a number of input vectors are exactly reproduced by the network. These vectors are invariant under the transformation that the network performs.

The training set of a fully trained autoassociator constitute the invariant vectors. In standard matrix algebra, vectors which are invariant under matrix multiplication (modulo multiplication by a constant) are described as *eigenvectors*. By analogy, we might define the trained inputs to an autoassociator network as the quasi-eigenvectors or *q-eigenvectors* of the network's transformation function. Parts

of the input space in the neighborhood of these q-eigenvectors will act as attractor basins and map onto the invariant vector at the output. This is what gives the network the power to deal with noisy input and to perform pattern completion. Other input vectors will be distorted according to how much they lie within the attractor basins of the network's q-eigenvectors. If they lie completely within an attractor basin, they will be mapped onto a particular eigenvector. However, in most cases, an unrelated input vector will fall across several attractor basins, each of which will attempt to map that segment of the input vector onto its appropriate q-eigenvector. What results will be a significant distortion of the input vector, with different parts being mapped onto different q-eigenvectors.

More formally, the transformation function implemented by this network is:

$$\text{OUTPUT} = f(\text{INPUT}) = \mathbf{M}_2(g(\mathbf{M}_1(\text{INPUT}))) \quad (2)$$

where OUTPUT is the output vector, INPUT is the input vector, $\mathbf{M}_1$ is the matrix of weights between the input and hidden units, $\mathbf{M}_2$ is the matrix of weights between the hidden units and the output units, and g(x) is a non-linear monotonic function (such as the logistic function) applied to each component of x.

The degree to which an input is distorted will depend on how close INPUT is to an eigenvector of $\mathbf{M}_1$, how strongly non-linear g(x) is, and how close $g(\mathbf{M}_1(\text{INPUT}))$ is to the eigenvectors of $\mathbf{M}_2$. Note that because g(x) is non-linear, it is not metric preserving and hence *f* itself is not a metric invariant transformation. The function implemented by a feedforward network does not preserve metric relations.

We can then define Transformational Function Similarity (TFS) as the inverse of the distance between the original input and the transformed output of the network (i.e. 1/√error score). Comparing A to B involves presenting A to a network able to autoassociate B and evaluating how much A has been transformed by the q-eigenvector encoding the B representation. Patterns that experience a small degree of transformation are very similar to B, whereas patterns that are transformed to a high degree are dissimilar to B. The transformation occurs naturally as part of connectionist information processing. The evaluation of the transformation (measuring the degree of distortion of A) is a post hoc process that can involve metric comparisons.

An example of TFS measurements

In order to generate notional knowledge bases to illustrate TFS measurements, we will define 3 concepts, {a}, {b}, and {c}. Each concept will comprise 3 prototypes, defined over a vector of 15 features. We generate 10 exemplars from each prototype, by adding Gaussian noise (sd=0.2). The network will gain its knowledge of each concept by training on each set of 30 exemplars. A network is trained to autoassociate each knowledge base over separate representational resources. The network is shown in Figure 1.[1]

For each knowledge base, we derive a mean vector from the set of exemplars. This represents the central tendency of that knowledge base and will provide us with a characterization of the knowledge stored in the network, for comparative purposes in the analysis provided below. The mean vectors are as follows.

a' = .1 .6 .6 .1 .6 .6 .1 .1 .1 .1 .3 .3 .3 .3 .2
b' = .1 .1 .1 .1 .6 .6 .1 .7 .6 .1 .3 .2 .3 .3 .2
c' = .1 .6 .6 .1 .4 .4 .1 .6 .6 .1 .3 .3 .3 .3 .2

If we define the metric similarity between these vectors as the inverse of the Euclidean distance between them, then their metric similarities are as follows.

aSb = 0.9, bSc = 1.1, aSc = 1.1

Note that these values are symmetrical:

bSa = 0.9, cSb = 1.1, cSa = 1.1

Now consider the TFS values (where TFS is defined as 1/√SSE of the autoassociator):

aSb = 1.0, bSc = 1.3, cSa = 1.1

These are not symmetrical measures:

bSa = 1.0, cSb = 1.1, aSc = 1.3

The TFS measure is at a maximum when the mean vector for a given knowledge base is transformed by that knowledge base. Thus

aSa = 2.5, bSb = 2.5, cSc = 2.5

These figures are the average of 12 runs of the network. This averaged result demonstrates that in principle, transformation based comparisons do not have to be symmetrical. However, it also masks individual cases where there are greater asymmetries in the comparisons (see Figure 2, cases 1 & 2). This demonstrates that comparisons will be sensitive to prior network states and the nature of the exemplar set to which the network is exposed.

In this example, we have used the mean of the training exemplars, a', to represent the input vector in the comparison *A is like B*. This is a simplification. Subjects will use their own conceptual store (reflecting their personal history of encounters with the exemplars) to internally generate the most prototypical representation of a concept in the given context, rather than a simple average of all their encounters with exemplars of that concept. It is this representation that will be transformed in the comparison. Nevertheless, the current example demonstrates that TFS

[1] 6 hidden units were used to represent each knowledge base, and the sub-networks were trained for 1000 epochs with a learning rate of 0.1 and a momentum of 0. The network also employed sigmoidal output units. Network weights were initially randomized between ±1.0.

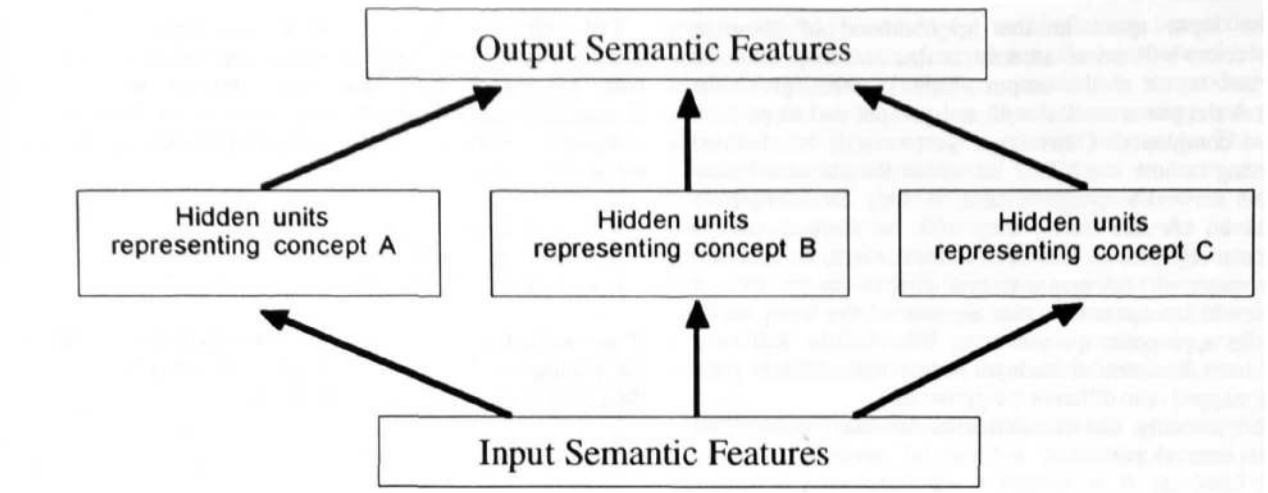

Figure 1. Example neural network architecture for performing similarity judgments using Transformational Function Similarity.

similarity may generate comparisons that are not symmetrical.

Concepts and classification

The psychological story behind this form of processing is as follows. Similarity judgments *per se* are not a primary function of the cognitive system. Similarity arises as a consequence of classification. It is of crucial importance for an organism to be able to classify new situations and objects in its environment so that it may bring to bear appropriate knowledge in dealing with them. Given a set of features that describe a new situation/object, the cognitive system's task is one of pattern recognition. This is a task that connectionist networks are well suited to perform. A frequently proposed architecture for connectionist pattern recognition is autoassociation. To establish whether X is an instance of A, we find out whether a network trained to autoassociate the various instances of A can accurately reproduce X. The similarity judgment is a reflection of the accuracy of that reproduction. In this view, similarity judgments do not require a special purpose mechanism. Similarity judgments are an adjunct to our ability to classify.

In general, the representation of a concept is developed through experience with a range of exemplars. A network that autoassociates knowledge about these exemplars will extract a prototype of the concept, to which it will respond preferentially. This reflects the typicality effects demonstrated by humans in classification tasks (Rosch, 1973). A network that represents a concept will thus tend to generate q-eigenvectors for the prototype or prototypes of that concept.

The representation of concepts and the role of context in comparisons.

In Figure 1, we have split the representations of the concepts {a}, {b}, and {c}, into separate sub-networks. In fact, it is more likely that concepts would share representations as a function of their similarity. One avenue of future work would be to determine how this organization might emerge by virtue of the learning procedure.

We envisage that specific instances of a general concept would be represented as specific mappings across the general area of the network responsible for representing the concept. Thus the concept {Michael Jordan} would be a mapping across the sub-network responsible for representing basketball players. This has the following implication with regards to similarity comparisons. Asking whether X is similar to a basketball player enforces a given transformation on X. Asking whether X is similar to Michael Jordan, however, would involve using this same basketball network with the Michael Jordan label activated. This would modify the transformation performed by the network. In this example, the Michael Jordan label plays the role of contextual information, that mediates the transformation performed by the basketball player network. This illustrates the more general point that Transformational Function Similarity is context sensitive. See Thomas and Mareschal (1996) for a more detailed discussion of context effects.

Bias, Salience, and Coherence revisited

We have claimed that asymmetries in comparisons fall naturally out of the comparison process itself. In section 3, we reported a number of additional factors proposed as explanations for how symmetrical comparison procedures could generate asymmetries. The notions of bias and salience were ascribed to individual concepts. For example, when a concept with a high salience formed the second term of a comparison, then similarity was greater than if it formed the first term. The TFS approach also allows for effects stemming from individual concepts. These will relate to the normal factors which determine how well networks perform transformations in general. Thus an autoassociative network with more training will tend to produce more accurate reproductions than an equivalent network with less training. And within a given network, exemplars appearing more often in the training set will tend to be reproduced

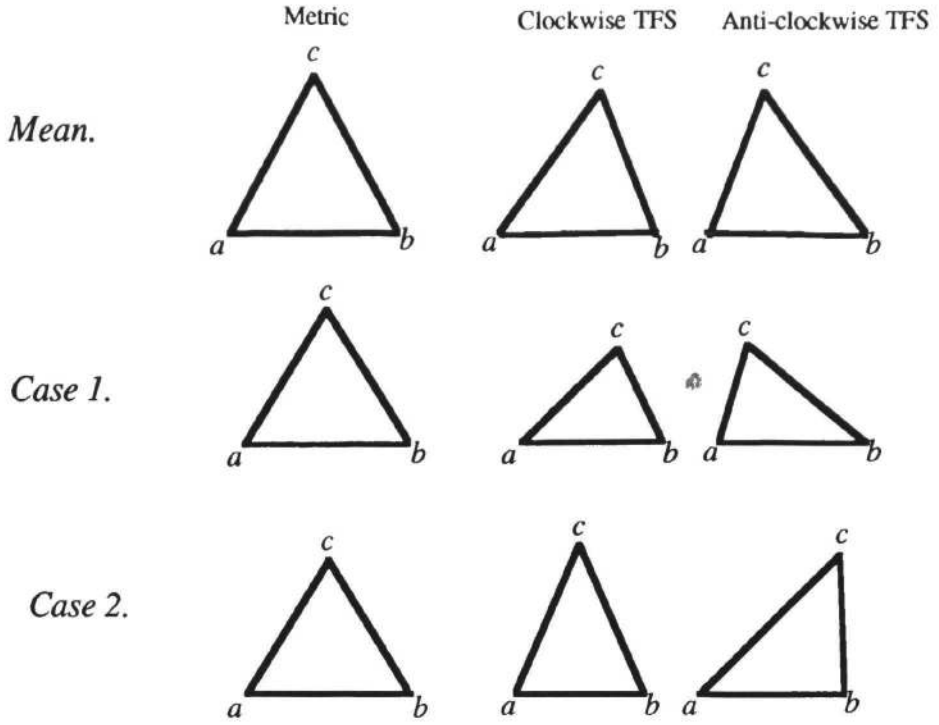

Figure 2. These triangles illustrate the similarity distance between three concepts {a}, {b}, and {c}, generated by a metric similarity measure such as Euclidean distance (column 1) or by the metric comparison phase of the Transformational Function procedure (columns 2 and 3). The metric comparison is symetrical. The TF similarity is asymmetrical, varying according to the direction of comparison (Clockwise = *a* to *c*, *c* to *b*, and *b* to *a*; Anti-clockwise = *a* to *b*, *b* to *c*, and *c* to *a*). The length of the side between two vertices corresponds to the inverse of the similarity between the corresponding concepts.

more accurately than those presented less often. The network performing the transformation corresponds to the second term in the comparison. If we see the salience of a concept as equivalent to the amount of training a network has received on that concept, then the claim that more salient concepts produce greater similarity judgments when they form the second term of a comparison corresponds to the idea that a better trained autoassociator with more training, autoassociates better.

The second term is also privileged when it is a more general or prototypical case of a given concept. Thus subjects prefer Red China to come second in the comparison of Red China and North Korea because China is the more general case of a communist country (Ortony, Vondruska, Foss, and Jones, 1985). In network terms, this preference reflects that fact that an autoassociator trained on a wider range of patterns will tend to produce more accurate autoassociations of any given pattern. Thus a network trained on every possible autoassociation would reproduce every pattern very accurately. Every pattern would have a high similarity to that knowledge base.

Gentner and Bowdle (1994) put forward the notion of coherence to explain how a symmetrical mapping procedure could generate asymmetrical comparisons. Similarity will be judged greater when the more coherent concept comes second in a comparison. The notion of coherence is tied to theories concerning the relation between linguistically structured representations - that is, those constructed along assumptions of compositionality and systematicity. Connectionist networks are not currently at a stage to give robust accounts of this kind of conceptual representation. If a solution to this problem can be found then the TFS theory may similarly be extended. For example, a structural comparison of the concepts A and B, might involve a transformation of the structure of A using the network representing the structure of B.

In short, the TFS measure is consistent with ideas of bias and salience previously proposed to account for asymmetry effects. Both fall naturally out of the training procedures used with connectionist autoassociators. For the moment, it is difficult to see how the idea of coherence could be extended to the TFS measure.

Analogy: Static Mapping or High Level Perception?

Previous computational models of analogy have broadly fallen into two camps. The first of these sees analogical comparisons as involving mappings or links between two static representations (e.g. ACME: Holyoak and Thagard, 1989). Some kind of mapping "engine" sees how well one representation fits over another: whether they have the same shape, which parts of one correspond to which parts of the other, and so on. The second view sees analogical comparisons as involving the formation of new, dynamically configured representations, created by the comparison process itself (e.g. Copycat: Hofstadter, 1984;

Mitchell, 1993; Tabletop: Hofstadter and French, 1994). These researchers describe analogy as a process of "high level perception". In the comparison "A is like B", the process really is one of seeing *A as if it were B*.

Theories of analogy must be based on an underlying notion of similarity. A theory of analogy based on the TFS view would have a foot in both of the above camps. A comparison initially involves a transformation, which generates a new representation. For "A is like B", the B knowledge transforms the A representation to create a new representation, B(A). Thomas and Mareschal (1996) proposed that this new representation might be seen as a metaphorical comprehension of A, transformed by seeing A as B. To measure the similarity of A to B (for example, in order that one might respond in a similarity judgment task), one must evaluate how well B knowledge has reproduced the A representation. That is, a procedure must compare the static representations for A and B(A). To derive a list of features which A and B have in common, one notes the features of A that have been strongly reproduced in B(A). Under a TFS view, first there is a transformation, then there is a comparison. In other words, analogy involves *both* processes of high level perception and of the comparison of static representations.

Conclusion

In this paper we have outlined Kitcher's (1996) criticism of connectionist processing; namely, that Connectionism employs similarity based processing, but that its basis of similarity is not supported in human similarity judgments. We have sketched out an approach based on connectionist processing, in which similarity is conceptualized as a transformation. Transformational Function Similarity (TFS) naturally exhibits asymmetry in comparisons, so that the similarity of A to B is not always equal to the similarity of B to A. This asymmetry emerges directly from the non-linear processing of connectionist networks. Connectionist processing is thus consistent with psychological notions of similarity, and Kitcher's criticism is unwarranted.

Other theories of similarity have accounted for the asymmetric nature of comparisons by extending basically symmetrical comparison procedures. In the TFS account, the asymmetry is a property of the comparison procedure itself. The notion of similarity as a transformation offers a basis from which to explain effects such as asymmetry that arise from highly constrained empirical situations, such as asking subjects to compare countries. Such tasks are thought to be simple and to reveal the basic processes of similarity judgments. That asymmetries exist even in apparently straightforward examples could be taken to imply that the basic mechanisms underlying comparisons must themselves generate the asymmetry. However, these considerations may obscure the fact that there are many more complex types of analogical problem solving, which involve the extended comparisons of previously unrelated domains. We suggest that the explanations for simple, rapid judgments of similarity between concepts may differ from those required to account for slower, reasoning based comparisons. The TFS account lies very much with the class of simple, rapid mechanisms.

References.

Churchland, P. S. and Sejnowski, T. J. (1992). *The Computational Brain.* Cambridge, Mass.: MIT Press.

Gentner, D. and Bowdle, B. F. (1994). The coherence imbalance hypothesis: a functional approach to asymmetry in comparison. *In Proceedings of the 16th Annual Conference of the Cognitive Science Society.* Erlbaum. 351-356.

Hahn, U. and Chater, N. (1996) Concepts and similarity: The chicken or the egg. In K. Lamberts and D. Shanks (Eds.) *Knowledge, Concepts, and Categories.* London, UK: UCL Press.

Hofstadter, D. R. (1984). *The Copycat project: An experiment in non-deterministic and creative analogies.* Cambridge, MA: MIT A. I. Laboratory Memo 755.

Hofstadter, D. R. and French, R. M. (1994). Probing the emergent behavior of Tabletop: an architecture uniting High-level Perception with Analogy-making. *Proceedings of the 16th Annual Meeting of the Cognitive Science Society.* Erlbaum. 528-533.

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

Kitcher, P. (1996). From Neurophilosophy to Neurocomputation: Searching the computational forest. In: R. N. McCauley (Ed.) *The Churchlands and their critics.* Blackwells.

Markman, A. B. and Gentner, D. (1993). Splitting the differences: A structural alignment view of similarity. *Journal of Memory and Language, 32,* 517-535.

Mitchell, M. (1993). *Analogy making as perception.* Cambridge, MA: MIT Press.

Nosofsky, R. M. (1991). Stimulus bias, asymmetric similarity, and classification. *Cognitive Psychology, 23,* 94-140.

Ortony, A, Vondruska, R. J., Foss, M. A., and Lawrence, E. J. (1985). Salience, similes, and the asymmetry of simil-arity. *Journal of Memory and Language, 24,* 569-594.

Rips, L. J., Shoben, E. J., and Smith, E. E. (1973). Semantic distance and the verification of semantic relations. *Journal of Verbal Learning and Verbal Behavior, 12,* 1-20.

Rosch, E. (1973). On the internal structure of perceptual and semantic categories. In T. E. Moore (Ed.), *Cognitive Development and the Acquisition of Language.* New York: Academic Press.

Rumelhart, D. E. and Abrahamson, A. A. (1973). A model for analogical reasoning. *Cognitive Psychology, 5,* 1-28.

Thomas, M. S. C. and Mareschal, D. (1996). A connectionist model of metaphor by pattern completion. In *Proceedings of the 18th Annual Conference of the Cognitive Science Society* Erlbaum. Pp. 696-701.

Tversky, A. (1977). Features of similarity. *Psychological Review, 84,* 327-352.

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