Representational distortion, similarity and the Universal Law of Generalization

Nick Chater and **Ulrike Hahn**
Department of Psychology
University of Warwick
Coventry, CV4 7AL, UK.
{n.chater,u.hahn}@warwick.ac.uk

Abstract

Psychological theories of similarity are typically defined over very limited classes of representations. Geometric models represent objects as points in a multidimensional space. Set-theoretic models represent objects as sets of features. We have recently developed an account of similarity, representational distortion, which can deal with arbitrary representations, and which includes geometric and set-theoretic accounts as special cases. The similarity between two representations is defined by the amount of distortion required to transform one representation into the other. This can be quantified by using the mathematical theory of Kolmogorov complexity (Li & Vitanyi, 1993). This theory has a range of psychologically interesting properties. In particular, we show that the Universal Law of Generalization can be derived.

Introduction

Similarity is a central notion throughout cognitive science. In perception, the similarity between sets of visual or auditory stimuli influences the way in which they are grouped. In speech recognition, the similarity between different phonemes determines how easily confused they are. In classification, the category assigned to a new instance may be influenced by the similarity of an new instance to past instances or a stored prototype. In memory, it has been suggested that retrieval from a cue depends on the similarity of past memory traces to the representation of the cue. Similarity also appears fundamental to learning and development: Because no situation, object or event is the same in all respects to any previously encountered situation, object or event, using past experience to guide future behavior requires generalizing from previous to new instances. It is widely assumed, in behaviorist as well as cognitive theories of learning, that this generalization is based, to some degree at least, on similarity. Similarity appears also to have an important role to play in problem solving, inference and scientific reasoning, especially if analogy is viewed as a special case of similarity.

Current theories treat similarity as a relation between mental representations. The two leading accounts, the geometric and featural views, differ on the nature of the representations, and the nature of the relation. The geometric view (Shepard, 1987) assumes that objects are represented as points in an internal space. The similarity between two objects is inversely related to the distance between their representations in this space. By contrast, the featural view (Tversky, 1977) assumes that objects are represented as sets of features. The similarity between two objects depends on the amount of overlap between their sets of features.

Both of these theories are limited in scope in that they define similarity over very specific kinds of representation: Points in space or feature sets. It would be attractive to have, instead, an account which applied to any representation whatever, whether a structural description of perceptual input, a parsed sentence, a schema encoding general knowledge, a pictorial represention or a sequence of motor commands. We might hope that such a framework would include geometric and set-theoretic accounts of similarity as special cases. We shall see that these goals are met by the account outlined below.

In this paper, we discuss a new theory of similarity which does apply to any representation whatever, based on the following intuition: Two representations are similar to the extent that there is a simple transformation which *distorts* one representation into the other. We call this measure of (dis)similarity *representational distortion.*

Making this intuitive notion precise requires (1) specifying what transformations the cognitive system can apply and (2) defining a simplicity measure over these transformation. We shall see that relatively general specifications of (1) and (2) suffice to develop a rich theory of similarity.

Transforming representations

Representational distortion assumes that similarity assessments involve estimating the complexity of the transformation between two representations. Exactly this process of assessing transformation complexity is at the core of an influential approach in a different area of cognitive science, perceptual organization (e.g., Palmer, 1983). We therefore begin our discussion with examples from perception, and explain how these can be related to similarity. We then consider how the approach may be applied more generally.

Figure 1 shows a pattern in which simple transformations are evident. The left and right parts of the figure are related by the geometric transformation of translation; moreover, this transformation also relates the left and right halves of these two parts of the figure.

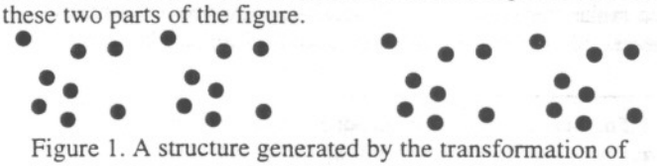

Figure 1. A structure generated by the transformation of repeating the same pattern.

In the study of perceptual organization, the presence of these transformational relations is said to give the stimulus a sense of cohesion or "goodness[1]." But according to the theory of representational distortion, these transformations can also be viewed as establishing the similarity between parts of the stimulus.

Figure 2 illustrates a case where the goodness of a figure derives from the composition of two transformations: symmetry and black-white inversion. The two halves of the stimulus are perceived as highly similar.

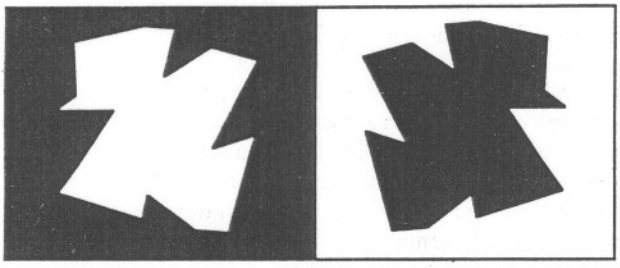

Figure 2. Symmetry and black-white inversion.

According to the transformational viewpoint concerning perceptual organization, the transformations that can be found in a stimulus determine its goodness. We have noted that simple transformations also correspond to strong similarity relations within the stimuli. This supports our suggestion that the representational distortion measures (dis)similarity.

A theory of similarity based on transformations depends, of course, on the transformations that the cognitive system can find. These are relatively easy to study in perception, where transformations correspond to manipulations of an external stimulus. But both representations and transformations over them are more difficult to study in higher level aspects of cognition—we do not have access to a person's representation of, say, "robin" or "blackbird," or the putative transformations (e.g., adding/deleting a red breast, complex deformations of shape, adding/deleting facts about behavior, and so on). This difficulty is, of course, analogous to that faced by geometric and featural views: There is no direct access to the location of objects in a putative mental space, or to the putative set of features associated with an object.

The representational distortion account has the following advantage with respect to other theories, however: It can be developed in very general mathematical terms, providing an account to which the operation of the cognitive system will approximate. This general formulation provides the basis of a powerful psychological theory of similarity.

Framework for representational distortion

We make the most general possible assumption concerning the set of allowable transformations (compatible with the computational view of mind): that it is the set of computable functions. We also define the *complexity* of a computable function in a very general and standard way, by its Kolmogorov complexity[2] (see Li & Vitanyi, 1993). Informally, the Kolmogorov complexity of a function is the length of the shortest computer program that computes that function. Thus, the intuition is that complex transformation are those that can only be expressed by long programs; simple transformation can be expressed by short programs (see Li & Vitanyi's excellent textbook for proofs that Kolmogorov complexity is language-independent and generally well-defined).

Putting these ideas together, the representational distortion between two representations, A and B, is determined by the length of the shortest program which distorts A into B[3]. This is expressed symbolically by the notation: $K(B|A)$. This formulation is both straightforward and technically attractive—it allows the rich theory of Kolmogorov complexity developed within mathematics and computer science to be exploited; and more specifically, it relates directly to specific proposals for a mathematical account of similarity advanced by Li & Vitanyi (1993), who define a natural family of measures of "information distance" between representations[4] (also see Chater, 1996 for a related application of Kolmogorov complexity).

We stress again that representational distortion, like the geometric and featural views, is defined over mental representations. To see why this is crucial, consider the psychological similarity of two unrelated bursts of white noise. At an acoustic level of description, where the bursts are considered as amplitudes varying over time, a very long set of instructions would be required to transform one of these bursts into the other. But the two noises may, nonetheless, be judged to be similar, even to the point that the auditory system cannot distinguish the two. According to our account, this is because the mental representation of the two bursts does not include minute detail of each aspect of the noise. Instead, they are concerned with a more general description, perhaps concerning the duration, loudness, location and so on of the burst. These properties may be largely or completely matched between stimuli, so that the mental representations of the two sounds are identical, or differ only slightly. We may assume that the information distance between these representations is small and this is reflected in the high psychological similarity between the two noises.

We stress also that the representational distortion found by the cognitive system will not correspond exactly to

[1] The intuitive notion of goodness may be operationalized in terms of a number of empirical measures: detectability, discriminability, and resistence to noise.

[2] Related ideas are discussed under the heads minimum message length, minimum description length, and algorithmic complexity theory.

[3] This length of program transforming from A to B is not necessarily the same as that transforming between B to A. We discuss this asymmetry further below.

[4] We do not have space here to summarize Li & Vitanyi's elegant analysis, but it provides a promising general framework for a psychological theory of representational distortion.

information distance. Discovering a short transformation between one representation and another may require arbitrary amounts of computation. For example, the sequences 1 5 3 7 2 3 9 0 6 and 3 0 7 4 4 7 8 1 2 are very simply related—if they are interpreted as base 10 numbers, the second is double the first. Hence the representational distortion between the two sequences is small; however, the cognitive system may not find this short transformation. Hence the system's estimate of representational distortion will be higher than the true representational distortion; and the similarity between the two representations may be underestimated. We assume therefore only that the cognitive system can approximate representational distortion to some degree. Indeed, finding the representational distortion between arbitrary representations is known to be an uncomputable function, and hence must necessarily be approximated (Li & Vitanyi, 1993). Moreover, note that this approach does not rely on a symbolic model of cognition—indeed, a recent connectionist model of metaphor and similarity (Thomas & Mareschal, 1996) can be viewed as a partial implementation of this approach.

As we have noted, this account applies to representations of all kinds, whether they are spatial, feature-based or, crucially, structured representations. We now note that spatial and featural models can be seen as special cases of representational distortion. The mathematical details have been omitted for brevity (see Chater & Hahn, in preparation).

Spatial model. Representations are limited to vectors of numbers. Transformations are limited to sequence of "nudges" of unit length (this length can be thought of as a limit of resolution in the space) and a "program" consists of a sequence of such nudges. If nudges can be in any direction, then the simplest transformation between two points is given by the distance of the straight line path between the points (this is the length of the "program" of concatenated nudges)[5]. This gives the Euclidean version of the spatial model. Restrictions to nudge direction to the axes gives a city-block version; allowing non-orthogonal axes derives the general Euclidean scaling model (Ashby & Townsend, 1986).

Featural model. Representations are limited to sets of features. Transformations are limited to the deletion and addition of features one by one. Thus a program consists of a sequence of deletions and additions. Assuming differential length for deletion and addition (specifically, deletion has the shorter code, because additions require specifying *what* is to be added), program length is then determined a weighted sum of the number of features that object A has and object B does not (which must be deleted) and that B has but A does not (which must be added). The length of this program is a close variant of Tversky's (1977) theory of similarity.

Aside from including existing accounts as special cases, the current approach applies quite generally, to mental representations and transformations of any kind. This general idea has been little explored in cognitive psychology. A rare exceptions is Franks and Bransford (1975), whose experiments suggest that category typicality may be related to the number of steps required to transform representations from the prototype.

We now briefly consider some basic properties of representational distortion that imply that it is a promising starting point for a psychological theory of similarity, before showing how it can be used to derive the Universal Law of Generalization.

Flexibility. The fact that similarity is defined over general representations takes account of the great flexibility of human similarity judgements (e.g., Medin, Goldstone & Gentner, 1993), because similarity is defined over representations of objects, and the goals and knowledge of the subject may affect the representations which are formed. As with the feature-based models (Tversky, 1977), this flexibility has both advantages, in terms of accounting for the flexibility of people's similarity judgements, and disadvantages, from the point of view of deriving testable empirical predictions.

Self-similarity is maximal, because no program at all is required to transform an object into itself.

Asymmetry. Representational distortion allows for asymmetry in similarity judgements: $K(x|y)$ is not in general equal to $K(y|x)$. This asymmetry is particularly apparent when the representations being transformed differ substantially in complexity. Suppose that a subject knows a reasonable amount about China, but rather little about Korea, except that it is ``rather like" China in certain ways. Then transforming the representation of China into the representation of Korea will require a reasonably short program (which simply deletes large amounts of information concerning China which is not relevant to Korea), while the program transforming in the reverse direction will be complex, since the minimal information known about Korea will be almost no help in constructing the complex representation of China. Thus, we would predict that K(China|Korea) should be greater than K(Korea|China). This is observed experimentally (Tversky, 1977). In some contexts, such as Shepard's Universal Law of Generalization below, similarity judgements are required to be symmetrical. This can also be modelled naturally by the average of the distances in either direction: $D(x,y) = 1/2(K(x|y)+K(y|x))$.

Background knowledge can be taken into account by assuming that this forms an additional input to the program which must transform one object into another. For example, if the arabic number system is part of your background knowledge, then you may perceive similarities between otherwise dissimilar patterns (i.e., dissimilar as mere patterns of dots), because numerical transformations will be available.

The Universal Law of Generalization

Shepard (1987) observes what he suggests may be a "Universal Law of Generalization," which applies across a wide range of stimuli, and applies to both people and animals. Shepard suggests that this is one of the most important regularities in the study of cognition, and that explaining it is therefore a central theoretical goal for psychological research. He proposes a derivation from the

[5] We ignore the cost of specifying the *direction* of a nudge for simplicity.

perspective of the spatial model of similarity, in terms of a set of assumptions about the "shape" and "size" of categories in an internal mental space. Here we aim to provide a simpler and more general explanation of the Universal Law, which can apply to mental representations of all types.

First, what is the Universal Law of Generalization? Suppose that a subject is trained to identify a set of stimuli. They are then tested and their errors used to compile a *confusability matrix* representing the probability that each stimulus is misidentified as each of the others. These confusion probabilities can be viewed as measures of similarity between mental representations of the stimuli, on the assumption that similar representations are more likely to be confused with each other. These probabilities are asymmetrical: i.e., the probability that stimulus x is idenified as stimulus y, written $P(\text{"}y\text{"}|x)$, is not, in general, equal to the probaility that y is identified as x, $P(\text{"}x\text{"}|y)$. Symmetry is imposed by using the measure:

$$gen(x,y)=\left[\frac{\Pr(\text{"}x\text{"}|y)\Pr(\text{"}y\text{"}|x)}{\Pr(\text{"}x\text{"}|x)\Pr(\text{"}y\text{"}|y)}\right]^{\frac{1}{2}} \qquad (1)$$

This measure is used as a proximity measure between representations of the stimuli, and non-metric multidimensional scaling (MDS) can then be applied. MDS locates the stimuli in a multidimensional space, so that distances between pairs of stimuli, $d(x,y)$ in the space preserves the rank order of the proximity measure between pairs of stimuli as far as possible. Shepard's Universal Law of Generalization is the remarkable empirical generalization that confusability is an exponentially decaying function of distance in the underlying "internal" space:

$$gen(x,y)=2^{-d(x,y)} \qquad (2)$$

We now consider how this can be derived using representational distortion. Let us assume that confusability arises because internal "noise" N degrades the representation of the remembered stimuli. We label the n training stimuli $S_1 \ldots S_i \ldots S_n$ and the noise degraded representations of these stimuli $D_1 \ldots D_i \ldots D_n$ (Figure 3). The cognitive system must compare the fresh presentation T_j (which is actually identical to the jth stimulus) to the degraded stimuli, and assess the probability that T_j is identical to each of the n stimuli, $P(T_j=i)$ (Figure 4).

First, we assume that subjects respond using "matching", as in standard models of choice (e.g., Luce, 1963) used to explain confusion data; that is, they choose responses in proportion to their probabilities of being correct, rather than always choosing the most probable response. This means that the probability, $p(\text{"}i\text{"}|j)$ that a subject *responds* "i" to a test stimulus T_j is subject's estimate of probability that that the stimulus *has* that identity, $P(T_j=i)$.

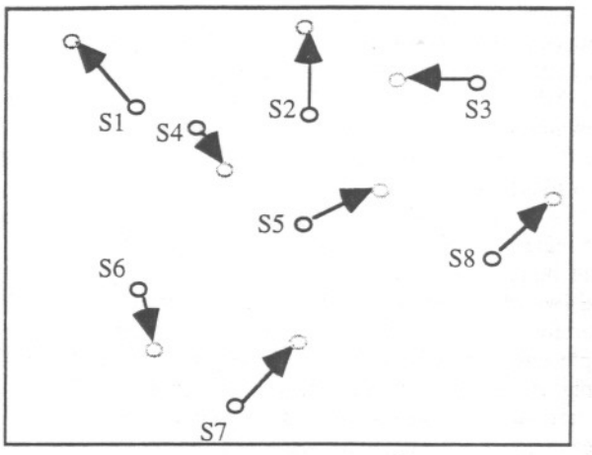

Figure 3. Stimuli, S_i, degraded by noise.

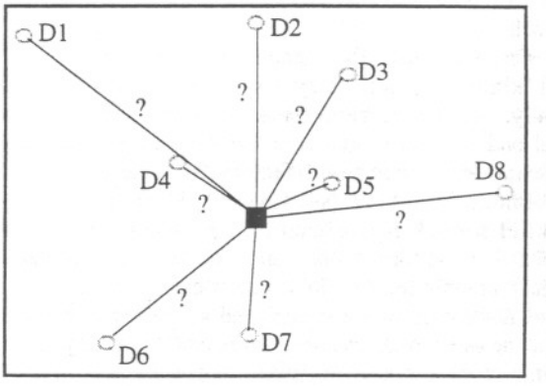

Figure 4. The target is presented (the square blob). The cognitive system must decide how likely it is to have generated each of the degraded representations D_i.

Conditional on the locations of the Di, the probability that each of the D_i is the degraded version of the target T_j is given by.

$$P(T_j=i|D_1 \ldots D_n)=\frac{P(N(T_j)=D_i)}{\sum_k P(N(T_j)=D_k)} \qquad (3)$$

The D_i will be randomly generated from the original S_i by the application of noise N. Therefore, the probability of responses given by (3) must be weighted by the probability that the D_i have these particular values (i.e., that the S_i degraded into these D_i). Calculating this quantity exactly is difficult, because the denominator involves all the D_k, and therefore the probability of each possible configuration of all the D_k must be considered. But we can approximate the denominator by assuming that $P(T_j$ degraded to $S_k)$ is a reasonable approximation to $P(T_j$ degraded to $D_k)$. Sometimes noise will move the S_k towards T_j, so that it is more probable; sometimes it will move the S_k away from

T_j, so that it is less probable. We assume that, to a reasonable approximation, these cases balance out[6]. This approximation to the denominator is independent of i, so that it can replaced by a constant. Therefore, (3) therefore can be rewritten:

$$P(T_j = i | D_i) = C \times P(N(T_j) = D_i) \quad (4)$$

where $C^{-1} = \sum_k P(N(T_j) = S_k)$. The expression (4) gives the probability of identifying the target as stimulus S_i, given that the noisy version of S_i, i.e., D_i, is given. But the D_i is not given—instead, it has a distribution (over the entire "space," Sp, of computable objects) which is determined by the noise operating on S_i. Specifically,

$$P(T_j = i) = C \times \sum_{D_i \in Sp} P(N(S_i) = D_i) P(N(T_j) = D_i) \quad (5)$$

To evaluate these probabilities we need to specify the nature of the noise. Rather than make specific assumptions about the character of this noise (e.g., that representations consist of binary strings, whose "bits" are flipped with some probability), we make the much more general assumption that representations are distorted by the action of an arbitrary "noise" program which acts on the representation as input, and produces a corrupted representation as output. For convenience, we make the psychologically unrealistic assumption that programs are written as binary strings on an arbitrary Universal Turing Machine, U. Fortunately, the theory of Kolmogorov complexity guarentees that the calculations below are independent of the choice of programming language or machine, so that no such unrealistic assumption concerning the computational processes which can corrupt internal representations in the cognitive system need be made. We assume that the probability that noise has degraded the original presentation of T_j into the degraded stimulus D_i is given by the probability that the noise program, *prog*, which is a randomly generated binary string, when used as a program for U will take T_j as input and produce D_i as output[7]. We also assume that, at each flip, there is a probability q that the flipping process is stopped[8]. If q is high, the noise programs will typically be short, and hence have relatively little effect; if q is low, programs will typically be long, and hence has a greater effect. Thus q is a parameter controlling the noise level.

The probability of generating any particular program *prog* as a random binary string is clearly:

$$P(prog) = \left(\frac{1-q}{2}\right)^{l(prog)} \quad (6)$$

where $l(prog)$ is the length of *prog*. The probability that a particular *prog* will transform T_j into D_j is:

$$P(N(T_j) = D_i) = \sum_{prog: U(prog, T_j) = D_i} \Pr(prog) = \sum_{prog: U(prog, T_j) = D_i} \left(\frac{2}{1-q}\right)^{-l(prog)} \quad (7)$$

A fundamental theorem of Kolmogorov complexity, the conditonal coding theorem (Li & Vitanyi, 1993), implies that, to an approximation, only the most probable (i.e., the shortest) program which transforms T_j into D_j , whose length is defined $K(D_j|T_j)$, must be taken into account:

$$\sum_{prog: U(prog, T_j) = D_i} 2^{-l(p)} \equiv 2^{-K(D_i|T_j)} \quad (8)$$

and we can adapt this to the present case, where the exponent is not 2, but has the larger value $2/(1-q)$, because here the smaller terms corresponding to longer programs fall off even more rapidly. Thus, we can replace (7) by:

$$P(N(T_j) = D_i) = \left(\frac{2}{1-q}\right)^{-K(D_i|T_j)} \quad (9)$$

Using this trick for the two noise terms in (5), we obtain:

$$P(T_j = i) = C \times \sum_{D_i \in Sp} \left(\frac{2}{1-q}\right)^{-K(D_i|S_i) - K(D_i|T_j)} \quad (10)$$

We further assume that $K(D_i|S_j)$ is approximately equal to $K(S_j|D_i)$, so that the exponent becomes $K(S_j|D_i) + K(D_i|T_j)$ (this will be true on the reasonable assumption that D_i and S_j) have approximately equal complexity). Thus (10) is now a weighted sum of ways of distorting T_j into S_j), via D_i, summed over all possible intermediate points D_i:

$$C \times \sum_{D_i \in Sp} \left(\frac{2}{1-q}\right)^{-\left(K(S_i|D_i) + K(D_i|T_j)\right)} \quad (11)$$

This can be simplified by the following observation (see Figure 5). Notice that we can define a universal programming language which operates in two phases, corresponding to two separate pieces of the program: first it turns its input into an intermediate form; and then it turns the intermediate form into the output. We assume that programs are generated by a sequence of random bit flips, with a probability $1-q$ of each flip being the last, as described for the noise process above. The probability of randomly

[6] These will not balance perfectly, because random displacements will more often move the S_i away from the T_j, since we are in a high dimensional spaee.

[7] Normalization is required to deal with the binary strings which are not well-formed programs. We ignore this complexity.

[8] The metaphor of generating programs by random coin flips is one way of motivating Universal A Priori Probability, see Li & Vitanyi, (1993).

generating a program that distorts T_j into Si can be expressed

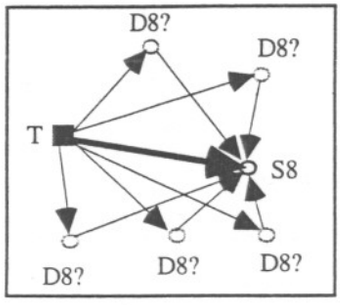

Figure 5. The target (the square blob) can be distorted into the stimulus S8 by an intermediate distortion to D8—the location of which is unknown. Summing probability over all possible routes (the thin arrows) turns out to be the effectively the same as considering only the shortest direct route, shown by the thick arrow.

in two ways. First, we can sum over all the possible intermediate representations D_i, which yields (11)[9]; second we can note that the only path which need be considered is the shortest path, which has length $K(S_i|T_j)$. This means that (11) is equivalent to:

$$\left(\frac{2}{1-q}\right)^{-(K(S_i|T_j))} \tag{12}$$

Putting our analysis together, this means that the probability of confusing S_j with a previously encountered S_i, $P(T_j=i)$, is given by:

$$P(T_j = i) \propto \left(\frac{2}{1-q}\right)^{-(K(S_i|T_j))} \tag{13}$$

Substituting into (1) gives

$$gen(i,j) = \left[\frac{\left(\frac{2}{1-q}\right)^{-K(S_i|S_j)} \cdot \left(\frac{2}{1-q}\right)^{-K(S_j|S_i)}}{\left(\frac{2}{1-q}\right)^{-K(S_i|S_i)} \cdot \left(\frac{2}{1-q}\right)^{-K(S_j|S_j)}}\right]^{\frac{1}{2}} \tag{14}$$

Note that $K(x|x)$ is 0, because the null program is sufficient to "distort" a representation into itself. Therefore $2^{-K(x|x)} = 2^0 = 1$ for all x, and hence we can simplify:

$$gen(i,j) = \left[\left(\frac{2}{1-q}\right)^{-K(S_i|S_j)} \cdot \left(\frac{2}{1-q}\right)^{-K(S_j|S_i)}\right]^{\frac{1}{2}}$$

$$= \left(\frac{2}{1-q}\right)^{-\frac{1}{2}(K(S_i|S_j)+K(S_j|S_i))}$$

$$= \left(\frac{2}{1-q}\right)^{-D(S_i,S_j)} \tag{15}$$

where D is the symmetric measure of representational distortion introduced above. Thus generalisation is an exponentially decaying function of distance in an internal space, if distance is measured in terms of representational distortion. That is, Shepard's Universal Law of Generalization can be viewed as naturally following from the representational distortion theory of similarity.

Conclusions

We have outlined a new psychological theory of similarity, based on the distortion between representations. This distortion can be quantified by the length of the shortest program which converts one representations into the other. This theory has a number of interesting psychological properties, and provides a derivation of Shepard's (1987) Universal Law of Generalization.

References

Chater, N. (1996). Reconciling simplicity and likelihood principles in perceptual organization. *Psychological Review*, 103, 566-581.

Chater, N. & Hahn, U. (in preparation). A rational analysis of similarity as representational distortion.

Franks, J. J. & Bransford, J. D. (1975). Abstraction of visual patterns. *Journal of Experimental Psychology: Human Perception and Performance*, 4, 65-74.

Li, M. & Vitanyi, P. (1993). *An introduction to Kolmogorov complexity and its applications*. New York: Springer-Verlag.

Luce, R. D. (1963). Detection and recognition. In R. D. Luce, R. R. Bush & E. Galanter (Eds.) *Handbook of Mathematical Psychology*, (pp. 103—189), New York: Wiley.

Medin, D. L., Goldstone, R. & Gentner, D. (1993). Respects for similarity. *Psychological Review*, 100, 254-278.

Palmer, S. E. (1983). The psychology of perceptual organization: A transformational approach. In J. Beck, B. Hope & J. Rosenfeld (Eds.) *Human and Machine Vision* (pp. 269—339). New York: Academic Press.

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, 237, 1317-1323.

Thomas, M. S. C. & Mareschal, D. (1996). A connectionist model of metaphor by pattern completion. *Proceedings of the Eighteenth Annual Conference of the Cognitive Science Society*. Mahwah, NJ: LEA.

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

[9] Note that in considering the probability of generating a subprogram which makes either of the two steps, we need only consider the length of the shortest such program, i.e., the Kolmogorov complexity of the step.