Available online at www.sciencedirect.com
ScienceDirect
Cognitive Systems Research 10 (2009) 216–228
Cognitive Systems RESEARCH
www.elsevier.com/locate/cogsys

Using analogical mapping to simulate time-course phenomena in perceptual similarity

Action editor: Angela Schwering

Andrew Lovett *, Dedre Gentner, Kenneth Forbus, Eyal Sagi

Northwestern University, 2133 Sheridan Road, Evanston, IL 60208-3118, United States

Received 12 March 2008; accepted 19 March 2008
Available online 10 January 2009

Abstract

We present a computational model of visual similarity. The model is based upon the idea that perceptual comparisons may utilize the same mapping processes as are used in analogy. We use the Structure Mapping Engine (SME), a model of Gentner's structure-mapping theory of analogy, to perform comparison on representations that are automatically generated from visual input. By encoding visual scenes incrementally and sampling the output of SME at multiple stages in its processing, we are able to model not only the output of similarity judgments, but the time course of the comparison process. We demonstrate the model's effectiveness by replicating the results from three psychological studies that bear on the time course of comparison.

Keywords: Analogy; Visual similarity; Same-different task; Computational modeling

1. Introduction

There is accumulating evidence that analogical processes play a role in many high-level cognitive operations, from language acquisition (Gentner & Namy, 2006), to scientific discovery (Dunbar, 1999; Holyoak & Thagard, 1995). However, equally interesting is analogy's role in low-level operations that are ubiquitous in our everyday lives. Gentner and colleagues (Gentner & Markman, 1997; Medin, Goldstone, & Gentner, 1993) have suggested that individuals determine the perceptual similarity of two simple images via the same processes of structural alignment that are used in conceptual analogies. For example, Markman and Gentner (1996) demonstrated that in judging the similarity of two images, participants attended more to *alignable differences*, those differences which were connected to the common structure of the images, than to differences unrelated to the common structure. This finding is predicted by Gentner's (1983) structure-mapping theory of analogy, in which individuals compare cases by aligning their common structure, thereby highlighting the alignable differences.

Further support for the claim that similarity is determined via structure mapping comes from studies conducted with the Structure Mapping Engine (SME) (Falkenhainer, Forbus, & Gentner, 1989). SME is a computational model based on structure-mapping theory. It has been used in a number of cognitive simulations and has successfully replicated human judgments in several studies, including the conceptual similarity of stories (Gentner, Ratterman, & Forbus, 1993) and the perceptual similarity of basic visual patterns (Gentner, Rattermann, Markman, & Kotovsky, 1995). However, previous work has not tested SME's ability to model the time course of the comparison process.

In this paper, we use SME to simulate the time course of comparison in three studies that utilize the visual *same-different* task. In the same-different task (Farell, 1985; Posner & Mitchell, 1967; Tversky, 1969), participants are shown two stimuli, a *base* image and a *target* image, and asked

* Corresponding author.
E-mail address: andrew-lovett@northwestern.edu (A. Lovett).

1389-0417/$ - see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.cogsys.2008.03.006

to judge whether the stimuli are the same. A consistent finding across a broad range of stimuli (see Farell, 1985 for a review) is that as the number of differences between the stimuli increases, the time required to respond "different" decreases. In contrast, the error rate for detecting that stimuli are different typically remains fairly stable (and low), regardless of the number of differences. Given enough time, participants are generally able to detect that the stimuli are different, even when there is only a single difference between them.

Goldstone and Medin (1994) studied the time course of the same-different task in greater detail by limiting the amount of time available for comparing the stimuli. They found a shift over time. Under very short deadlines, participants' similarity judgments depend only on the attribute matches between the two figures. However, when given more time, participants rely more on the relational matches. At this later stage, only those matches that are consistent with a global mapping between the structure of the stimuli contribute to the similarity of the stimuli.

A study by Sloutsky and Yarlas (submitted for publication) suggests that a further distinction must be made between the time course of *encoding* stimuli and the time course of *comparing* stimuli. Their results (described in detail below) indicate that individuals are faster to encode attributes than to encode relations when viewing a perceptual scene. Thus, there may be two factors that operate to make relational matches occur later during comparison than attribute matches: the time required for encoding (Lovett, Gentner, & Forbus, 2006) and the time required for comparison (Gentner & Sagi, 2006; Goldstone & Medin, 1994).

We present a model of the same-different task that takes account of both encoding and comparison processes. The comparison process is modeled using SME. We argue that the multiple stages that SME uses to construct mappings provide part of the explanation for the rapid difference judgments individuals make when either the stimuli are very dissimilar or there is limited time for comparison (Gentner & Sagi, 2006; Sagi, Gentner, & Lovett, in preparation). The encoding process is modeled as an incremental, any-time process in which attributes are encoded before relations. We use automatically generated structural representations of sketched stimuli, to reduce tailorability.

This paper describes our model of the same-different task, using incremental encoding and SME's multi-stage comparison process to explain both time course and error rate phenomena. We start with an overview of SME, focusing on how SME's operations are sampled to provide same/different judgments that depend on timing. Next we review three same-different studies that provide evidence about time course phenomena, focusing on the constraints they suggest for models. We then describe our model of the encoding process, including automatic generation of representations. We show that the combination of SME and our model of encoding can replicate both timing and error rate results from the three human studies.

2. The Structure-Mapping Engine

The Structure-Mapping Engine (SME) (Falkenhainer et al., 1989; Forbus & Oblinger, 1990) is a computational model of Gentner's (1983) structure-mapping theory of analogy. It takes as input two cases, a *base* and a *target*. Each case is a structural description containing entities and expressions. Expressions can describe attributes, such as the color or shape of an object, or relations, such as one object being larger than another, or above it. First-order relations connect entities, whereas higher-order relations connect other relations. SME finds one or more *mappings* between the base and target by aligning their common relational structure. It prefers mappings that maximize *systematicity*, i.e., include larger relational structures, especially those containing higher-order relations that constrain the lower-order relations.

A mapping has three parts: (1) A list of *correspondences* between items (entities and expressions) in the base and target. (2) A *structural evaluation score* measuring the degree of similarity between base and target. (3) A set of *candidate inferences* about the target, supported by what is known about the base and the correspondences.

SME computes mappings in three stages (see Fig. 1), which we illustrate with an analogy between basketball and soccer. Fig. 2 shows the expressions for the base case (basketball) and the target case (soccer), with entities in boldface. The expressions describing colors are examples of attributes. The expressions beginning with `score` and `moveThrough` are examples of first-order relations. The `cause` expressions are examples of higher-order relations.

In the local match stage, SME computes all possible match hypotheses between items in the base and target. A match hypothesis is created between every pair of expressions that share the same predicate. In the sports example, there would be a match hypothesis between `(white` **`net0`**`)` in the base and `(white` **`net1`**`)` in the target, as well as a match hypothesis between `(white` **`net0`**`)` in the base and `(white` **`soccer-ball`**`)` in the target. Match hypotheses are formed for both attributes and relations, so there would also be a match hypothesis between each of the two causal relationships in the base and the causal relationship in the target. SME also attempts to create match hypotheses between arguments of matched expressions. This generates match hypotheses between entities. For example, the match hypothesis between `(score` **`player0 points0`**`)` and `(score` **`player1 points1`**`)` generates match hypotheses between **`player0`** and **`player1`** and between **`points0`** and **`points1`**. This process of matching the arguments of matched predicates can also lead to matches between expressions with non-identical predicates: e.g., matching two terms with different functions, such as mapping pressure to temperature in a water/heat analogy. The result of the local match construction process is an inchoate, typically structurally inconsistent set of match hypotheses, out of which consistent global mappings emerge.

Stage	What Happens	Produces	Same-Different Decision Criterion
Local Match Construction	In parallel, match hypotheses conjectured between identical predicates and corresponding arguments	Forest of local match hypotheses. Globally inconsistent	Feature overlap
Kernel Construction	In parallel, find local maximal structurally consistent overlapping structure.	Kernels	
Mapping Construction	Serially find one or more global mappings, via greedy merge of kernels	Mappings	Candidate inferences

Fig. 1. The three stages of the Structure-Mapping Engine.

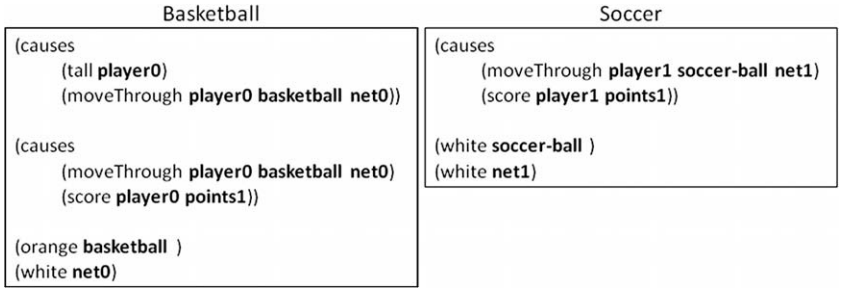

Fig. 2. Possible representations for basketball and soccer.

In the kernel construction stage, SME explores structural consistency: it identifies groups of match hypotheses that represent a local overlapping piece of consistent structure. These kernels are the pieces from which global mappings can be constructed. For example, the match hypothesis between

(causes (moveThrough **player0 basketball net0**)
(score **player0 points0**))

and

(causes (moveThrough **player1 soccer-ball net1**)
(score **player1 points1**))

and all of the match hypotheses for their subexpressions will be grouped together as a kernel. Note that kernels can overlap: there will also be a kernel for (white **net0**) and (white **net1**). Thus the correspondence between **net0** and **net1** will be a member of both kernels.

In the mapping construction stage, SME uses a greedy merge process to combine these kernels into globally consistent mappings. Both of the kernels above, for example, will be in the same global mapping. The kernel which maps (white **net0**) to (white **soccer-ball**) is inconsistent because it puts **net0** and **soccer-ball** into correspondence, which would violate the 1:1 constraint of structure-mapping.

A global mapping's structural evaluation score is the sum of the scores calculated for the match hypotheses. Match hypothesis scores are calculated by assigning a fixed score to each match hypothesis and then allowing scores to trickle down from match hypotheses between relations to match hypotheses between those relations' arguments, thus allowing match hypotheses that support deep aligned structure to receive particularly high scores. The mapping just described would receive a reasonably high structural evaluation score because it maps a higher-order relationship.

Candidate inferences are computed by examining unmapped base expressions that connect to the mapped structure. For example, the causal relationship between being tall and getting the ball through the net, found in the base, is not a part of our mapping. Therefore, SME will conjecture, based on the common relational structure,

(causes (tall **player1**)
(moveThrough **player1 soccer-ball net1**)

In other words, because being tall helps players score in basketball, it may also help them score in soccer.

For modeling the same-different task, several features of SME's processing are important to notice (see again Fig. 1). First, pairs of stimuli that are very different will give rise to small sets of match hypotheses, because they simply do not have very much in common. Pairs of stimuli that are very similar (or identical) will give rise to larger sets of match hypotheses. Thus the size of the set of match hypotheses computed by SME in its first stage provides one rapid decision criterion for same-different tasks. This criterion, which we call the *feature overlap*, will be reasonably accurate for saying that two things are different. It will be less accurate for saying that two things are the same, since the match hypothesis network is inchoate—many of the match hypotheses may be mutually incompatible.

To compare pairs of stimuli that have moderate-to-large feature overlap, a more fine-grained decision criterion is needed. For this purpose, we can look for candidate inferences after computing the full structural alignment. A candidate inference is generated only when there is a difference connected to the common structure (in many cases an alignable difference) between the two stimuli, and hence if a candidate inference is detected, the stimuli cannot be identical. The accuracy of this decision criterion depends on the accuracy of encoding. Decisions based on the presence of candidate inferences will be slower than decisions based on the number of match hypotheses because candidate inferences appear only after all three stages of SME's operation have finished.

3. The same-different task

We consider three studies of the same-different task, to identify constraints on models of the task. In the first study (Goldstone & Medin, 1994), participants were given limited time to compare two stimuli before responding "same" or "different." In this study, the time limit applied to both encoding the stimuli and comparing them; thus it is not possible to distinguish between encoding time and comparison time in interpreting the results. The second study (Sloutsky & Yarlas, submitted for publication) placed a limit only on the time available for encoding the base image, thus (partially) isolating encoding time from comparison time. The final study (Gentner & Sagi, 2006) gave the participants unlimited time for encoding and comparison, and examined the effects of varying the alignability of the pairs of stimuli.

3.1. A same-different task with limited time to encode and compare

3.1.1. Experiment

Goldstone and Medin (1994) investigated a same-different task in which the base and target images consisted of pairs of butterfly figures. Each butterfly varied along four dimensions: head shape, tail shape, body texture, and wing texture. The two butterflies in the base image[1] differed along all four dimensions (see the left image in Fig. 3). The two butterflies in the target image were systematically varied to produce different degrees of overlap with the butterflies in the base image. For example, if the two base butterflies were classified as AAAA and BBBB (where each letter stands for a value along one of the four butterfly dimensions), a target butterfly classified as AAAB would share three features with the first base butterfly and one feature with the second base butterfly.

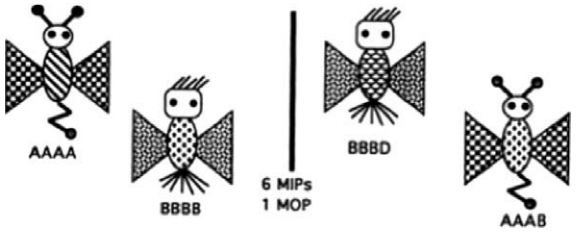

Fig. 3. A base and target image from Goldstone and Medin (1994).

In some cases, the attribute matches were consistent with the best global mapping between the two images; in other cases they were cross-mapped—that is, inconsistent with the best global mapping (Gentner & Toupin, 1986). Goldstone and Medin referred to these as matches-in-place (MIPs) and matches-out-of-place (MOPs), respectively. For example, consider the butterflies in Fig. 3. The best global mapping between the base and target butterflies would map the AAAA butterfly to the AAAB butterfly and the BBBB butterfly to the BBBD butterfly. In this case, butterflies AAAA and AAAB share three common attributes, which qualify as MIPs. Butterflies BBBB and AAAB also share one common attribute, but because these butterflies do not correspond to each other in the best global mapping, this would be considered a cross-mapping, or MOP.

Goldstone and Medin instructed participants to ignore the relative positions of the butterflies. Thus, the only information used for comparison should be the shape or texture of the four butterfly parts (head, tail, body, wings), along with the relations that tie those butterfly parts together in a single butterfly.

3.1.2. Results

Goldstone and Medin ran their study under three deadline conditions that varied within subject: short (1 s), medium (1.84 s), and long (2.68 s). The primary result of interest was the effect that the number of MIPs and/or MOPs had on overall accuracy on trials where the images were different. Their chief findings are summarized in Fig. 4. In the short deadline condition, when participants had only 1 s to encode and compare the stimuli, both MIPs

[1] While in this study the figures were seen simultaneously, for convenience we use the terms *base* and *target* to refer to the left and right images.

and MOPs had an equal effect on error rates. That is, participants' tendency to (incorrectly) respond "same" when the stimuli were different increased with the number of common attributes, regardless of whether those common attributes were cross-mapped or consistent with the best global mapping. However, in the medium and long deadline conditions, the effect of MOPs decreased significantly. At long deadlines, participants were still influenced by the number of structurally consistent matches, but they were largely able to ignore the cross-mapped attributes.

3.1.3. Constraints on models

Goldstone and Medin's results suggest that a model of human performance on the same-different task must exhibit at least two patterns of responses, depending on the time available for encoding and comparison. When participants are given a very small amount of time for comparison (1 s), all common features between the base and target contribute to error rates equally, regardless of whether they are consistent with the overall structural alignment. When participants are given more time (1.84 or 2.68 s), the common features consistent with the structural alignment have a much greater influence on error rates than those features inconsistent with the alignment—it is as though participants no longer attended to the stray feature matches that do not belong in the alignment.[2]

The pattern of responses found by Goldstone and Medin is roughly consistent with SME's local-to-global matching process: early on, the feature overlap criterion, which depends on the forest of local match hypotheses, will be equally sensitive to MIPs and MOPs; but the candidate inferences produced after a full structural alignment is completed will be sensitive only to MIPs. However, because the study did not distinguish between time for encoding and time for comparison, it is unclear whether the effect of a limited time for comparison should be modeled in the encoding process or in the comparison process.

3.2. A same-different task with limited encoding time

3.2.1. Experiment

Sloutsky and Yarlas (submitted for publication) conducted a study that at least partly separated encoding time from comparison time. Instead of a simultaneous display, they used a sequential display paradigm. Participants were shown a base image for a limited amount of time, followed by a mask. Afterwards, they were shown the target image and given as much time as needed to determine whether

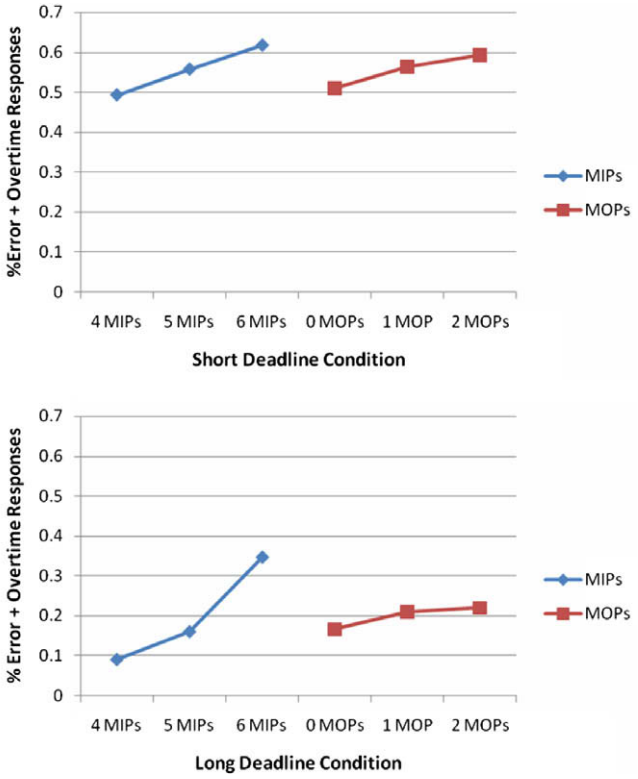

Fig. 4. Results from Goldstone and Medin (1994).

the two images were the same. Since only the time to encode the base image was manipulated, any effects on performance can be attributed to the encoding process, not to the comparison process.

The images used by Sloutsky and Yarlas were rows of three simple objects. The three objects all had different colors. However, two of them always had the same shape. The shapes appeared in one of three relational patterns: ABA, AAB, or ABB. For example, Fig. 5 shows a base image with an ABA pattern. For each base image, there were three target images that could differ from the base image along two dimensions: attributes and relations. An *element match* (E+) contained the same attributes as the base image, in other words the same colors and shapes, while

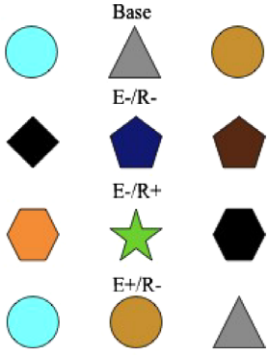

Fig. 5. Stimuli from Sloutsky and Yarlas (submitted for publication).

[2] Note that both of these patterns differ from the pattern of performing equally well regardless of the number of common features, a pattern found in most studies where participants have ample time to make their comparison (Farell, 1985). However, this high-accuracy pattern may simply reflect one pole of the speed-accuracy tradeoff: when participants have unlimited time to make a comparison, they can detect even a very subtle difference, but when their time is limited, the error rate increases as the number of differences decreases.

an *element mismatch* (E−) contained entirely different attributes. A *relational match* (R+) contained the same pattern of shapes (e.g., ABA and CDC), whereas a *relational mismatch* (R−) contained a different pattern of shapes. A key point for our purposes is that a target with element mismatches (E−/R+ or E−/R−) could be distinguished from the base image by the *attributes* of the three objects; but a target image that was purely a relational mismatch (E+/R−) could only be distinguished from the base image based on the *relations* between the three objects.

3.2.2. Results

In the *ample time* condition, the base image was shown for 2100 ms, but in the *limited time* condition, the base image was shown for only 150 ms. The results are shown in Fig. 6. When participants had ample time to encode the base images, they performed at about the same level across all three target types. However, when participants had limited time to encode the base image, performance varied markedly across conditions. Targets that contained element mismatches (E−/R+ and E−/R−) showed only a small drop in performance, but targets that were purely relational mismatches (E+/R−) showed a much greater decline. In other words, when participants had limited time to encode the base image, they had much more difficulty distinguishing it from a target image that differed only in relations. These findings suggest that people encode attributes before relations.

Regardless of time available for encoding the base, participants responded "different" more quickly after being shown the target image when it was an element mismatch—i.e., when the attributes were different (Fig. 7). These results support the priority of attributes over relations in the comparison process, consistent with Goldstone and Medin's (1994) work and with the idea of an early readout from the feature overlap stage of SME (followed later by a full structural alignment).

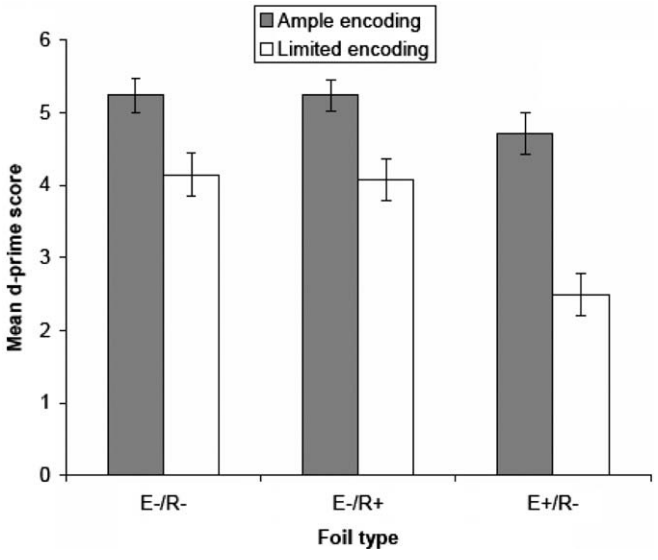

Fig. 6. Response accuracy in Sloutsky and Yarlas (submitted for publication).

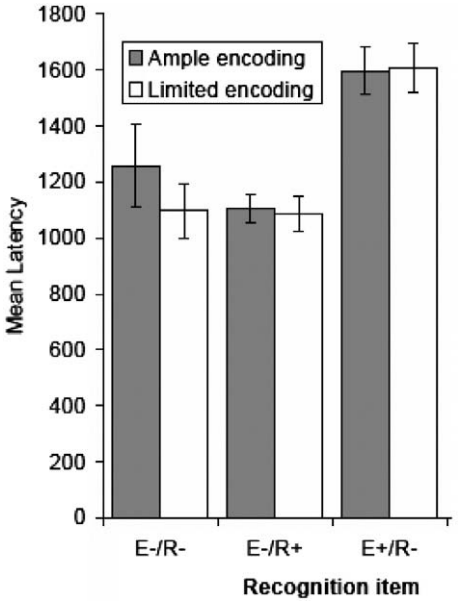

Fig. 7. Response times from Sloutsky and Yarlas (submitted for publication).

3.2.3. Constraints on models

The Sloutsky and Yarlas study suggests two additional constraints on models of the same-different task. First, it suggests that attributes are encoded before relations. Second, it suggests that the comparison process can operate over incomplete representations. Even when people were unable to completely encode the base image, the orderly results show that they were able to carry out comparisons between the partial representations and the new targets.

3.3. A same-different task with no time limit

3.3.1. Experiment

Gentner and Sagi (2006) conducted a same-different task with simultaneous presentation and no time limits for encoding or comparison. In one study, the materials were images of heraldic shields (see Fig. 8). Participants were presented with both high-similarity pairs (the rows in Fig. 8) and low-similarity pairs (the columns). The high-similarity pairs were highly alignable, containing most of the same spatial relations but differing in some small way, often a single attribute such as the color or shape of an element of the shield. The low-similarity pairs were entirely different.

3.3.2. Results

Unsurprisingly, participants were much faster to respond "different" for the low-similarity pairs (consistent with Farell's (1985) review). The reaction times were 1380 ms for the high-similarity pairs and 880 ms for the low-similarity pairs.

Fig. 8. Shield stimuli from Gentner and Sagi (2006).

3.3.3. Constraints on models

The rapid "different" response for low-similarity pairs is compatible with an early readout of feature overlap (number of match hypotheses) as a rapid decision criterion. If the raw feature overlap is extremely low, then the pair cannot possibly be identical (Gentner & Sagi, 2006), and a "different" response can be quickly given. If the number of match hypotheses is great enough so that the two items could be the same, then the process continues to a full structural alignment, resulting in a substantially slower reaction time.

Importantly, incremental encoding alone cannot explain these results. Attributes alone would be sufficient for distinguishing many of the high-similarity pairs, which often varied in a single or a few attributes. Thus, if individuals always engaged in a full comparison process but encoded and compared attributes before relations, they would be able to distinguish both the high-similarity and low-similarity pairs after encoding only the attributes. The high discrepancy in reaction times between high-similarity and low-similarity pairs supports the idea of a comparison process that can quickly render a rough measure of similarity.

3.4. Summary of constraints

To summarize, the constraints from the three studies suggest the following model. (1) The encoding process incrementally builds up representations over time, encoding attributes before relations. (2) The comparison process involves a local-to-global structural alignment process with the following signatures. (a) Rapid difference judgments can be made for very low-similarity pairs based on there being too few feature matches between the base and target (the feature overlap criterion), without consideration for whether these matches are consistent with a structural alignment between the base and target; (b) when there is limited time provided for comparison, participants' judgments will reflect only this feature overlap criterion; (c) judgments are slower for high-similarity pairs, for which people must compute a full structural alignment between the base and target and identify particular differences.

4. Modeling the same-different task

We start by describing our model of encoding, and then describe in more detail how we use the model of encoding and SME together to perform the same-different task.

4.1. Generating encodings using CogSketch

An important methodological technique in cognitive simulation is to use automatically generated representations. Hand-coded representations can be useful for some purposes, but they suffer from tailorability. By contrast, automatically generated representations provide more constraints, since the representation for each stimulus is computed by a precisely specified algorithm. This approach is particularly strong when the same representation encoding scheme is used across multiple simulations.

For our simulations, we use CogSketch (Forbus, Usher, Lovett, Lockwood, & Wetzel, 2008), an open-domain sketch understanding system, in the encoding process. CogSketch automatically constructs relational descriptions of sketches drawn by a user. Modelers draw one or more glyphs,[3] representing the objects in a sketch. CogSketch then automatically computes a number of spatial relations between the glyphs in a sketch. These include relative position, containment, and whether two glyphs are intersecting. CogSketch can also compare two glyph's shapes to see how they relate to each other, determining that two glyphs are the same shape or that one glyph's shape is a rotation, reflection, or rescaling of another. The relations and attributes encoded by CogSketch can be used to compare one sketch to another with SME. Examples of the representations that CogSketch produces and how these have been used in cognitive modeling can be found in (Lockwood, Forbus, Halstead, & Usher, 2006; Lovett, Forbus, & Usher, 2007).

The representation generated by CogSketch is the *ideal representation* of a sketch, modeling what a person might produce given ample time and attention. Our incremental encoding model simulates the time course of encoding such a representation. It produces an *available representation*, the subset of the ideal representation that is available for comparison, based on the amount of time provided for encoding. We simulate a short encoding time by including only the attributes, such as color and shape, in the available representation. We simulate a very short encoding time by

[3] In sketching, a *glyph* is the basic unit of a sketch, often corresponding to a particular shape or object drawn by the user.

including only a randomly selected subset of the attributes in the available representation. In cases where there is ample time available for encoding, the model will first produce an available representation containing only attributes and then produce a follow-up representation with both attributes and relations.[4]

4.2. Using comparison to make same-different decisions

As noted above, we use two distinct decision criteria, based on sampling SME's output at different stages in its processing (see Fig. 1). The first is based on the overall size of the match hypothesis network constructed in the first stage, the *feature overlap* decision criterion. The number of match hypotheses is normalized by the size of the base and target representations. If the base and target are identical, there should be at least one match hypothesis for every element in the base and for every element in the target, so the feature overlap should be 1.0 or higher. Of course, a feature overlap slightly below 1.0 might simply indicate a failure to encode one or two elements. A feature overlap well below 1.0 should provide strong evidence that the base and target representations are different, and thus should allow a participant to give a fast "different" response before completing the rest of the mapping. In the simulations below, we used a threshold of 0.6 for indicating fast "different" responses.

The second decision criterion is the *candidate inference* criterion, which requires full SME processing. Since this criterion is based on a structurally consistent global mapping, it is immune to match hypotheses that are inconsistent with that mapping. Normally, the presence of any candidate inferences should indicate that the stimuli being compared are different.

We note that this model is incomplete, in that it does not detect all non-alignable differences between stimuli. That is, a difference unconnected to the aligned structure would not produce a candidate inference, and thus would not be detected by the model. However, several findings (e.g., Gentner & Gunn, 2001; Markman & Gentner, 1996) suggest that alignable differences are more salient than non-alignable differences in comparisons. Although we have not so far found non-alignable differences to play an important role, this issue deserves further investigation.

4.3. Interaction between encoding and comparing

In our model of the same-different task, encoding and comparison interact as follows. As attributes are encoded, they are passed to SME. The comparison process is initiated before the full representation is built, but only the first stage (constructing the match hypothesis forest) is run at this point.[5] We assume that this is a fast, parallel process that can operate in parallel with continued encoding. Once all the attributes have been encoded and matched, the model checks whether the feature overlap has fallen below a set criterion (0.6). If the feature overlap is below the threshold, indicating that the two stimuli have far too few feature matches to be identical, a "different" judgment can be quickly rendered. Otherwise, the process continues.

Once the relations are encoded and SME has relations available, the comparison process continues through the other stages to a full structural alignment. At this point, candidate inferences are available, and more subtle differences can be detected.

5. Simulation experiments

We now describe three experiments in which we simulated the same-different tasks described above. The assumptions made in simulating the three experiments are described in Fig. 9.

5.1. Simulating a same-different task with ample time

5.1.1. Simulation

We begin by simulating the heraldic shield same-different task from Gentner and Sagi (2006) because, while it used the most complex stimuli, it is actually the most straightforward to simulate (Sagi et al., in preparation).

The original heraldic shields (Gentner & Sagi, 2006) contained a number of complex images that CogSketch would have had difficulties representing. Therefore, we constructed a new set of heraldic shields by replacing each complex object with a basic geometric shape, while maintaining the original spatial relations between the objects within the shields. Contrast Fig. 10 with Fig. 8 for an example. We ran a new set of participants on the same-different task using these simplified shields to ensure that the high-similarity/low-similarity differences would be maintained. Participants were able to respond "different" in 910 ms for the low-similarity shield pairs, but they required on average 1270 ms for the high-similarity shield pairs.

The heraldic shield stimuli were created in PowerPoint. They were then imported into CogSketch. CogSketch automatically created a glyph for each PowerPoint shape and computed the spatial and shape relations between glyphs, as well as shape and color attributes for each glyph.

[4] A further step would be to prioritize the encoding order according to the salience/psychological availability of particular predicates, but for now we have kept the model as simple as possible.

[5] We could also explore the possibility that some attribute kernels might be computed at this point (that is, that SME might enter its second stage and begin computing kernels before relations had been encoded). Then, after relations are encoded, new kernels could be created and the mapping could be updated, as when SME operates incrementally (Forbus, Ferguson, & Gentner, 1994). For the present set of studies, the results would not differ if this were the case.

Task	Condition	Representation	Decision Criterion
Butterflies (Goldstone & Medin 1994)	Short Deadline	Attributes only	Feature overlap
	Medium/Long Deadline	Complete	Feature overlap & Candidate inferences
Three Shapes (Sloutsk & Yarlas, submitted)	Limited encoding time	Partial (1-4) attributes for base	Feature overlap & Candidate inferences
	Ample encoding Time	Complete	Feature overlap & Candidate inferences
Heraldic Shields (Gentner & Sagi, 2006)	Unlimited encoding time	Complete	Feature overlap & Candidate inferences

Fig. 9. Assumptions made for each of the three simulated studies. The short and long deadlines in the butterfly study refer to time for both encoding and comparison.

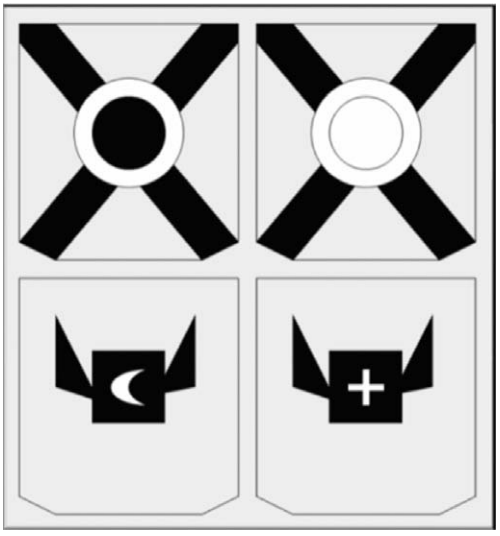

Fig. 10. Simplified shield stimuli.

5.1.2. Results

We presented the model with 24 high-similarity and 24 low-similarity pairs. As expected, the feature overlap (match hypotheses) scores were far lower for the low-similarity shield pairs. The low-similarity pairs received an average score of .45, while the high-similarity pairs received an average score of 1.12. 85% of the low-similarity pairs received a score below the .60 threshold for fast "different" responses, while 0% of the high-similarity pairs resulted in a fast response. These results match the finding that participants responded "different" faster for the low-similarity pairs.

When the feature overlap criterion does not provide an answer, the candidate inference criterion is used. In this simulation, our model was able to find at least one candidate inference in every comparison, indicating that, given sufficient time, the model could always correctly respond "different."

Fig. 11. Stimulus sketched for the Sloutsky and Yarlas (submitted for publication) simulation.

5.2. Simulating a same-different task with limited encoding time

5.2.1. Simulation

To simulate the Sloutsky and Yarlas (submitted for publication) study, we sketched six base images (see Fig. 11 for an example), along with the E−/R−, E−/R+, and E+/R− target images for each base. We sketched two base images for each of the three relational patterns (ABA, AAB, ABB). R− (relational mismatch) target images for these bases used each of the other two patterns. Thus, the six stimulus sets covered all possible combinations of relational patterns in the base and target images. Because there is no theoretical or functional difference between one shape or color and another, these six stimulus sets were sufficient for our simulation.

As in the previous study, CogSketch automatically computed spatial and shape relations between the glyphs, along with attributes. However, for this simulation, CogSketch also computed positional attributes for each glyph (`firstShape`, `secondShape`, or `thirdShape`), representing where in the row each shape was located.

For the ample encoding condition, in which participants had 2100 ms to encode the base image before it was masked, our encoding model fully encoded the base image. In the limited encoding condition, participants were given only 150 ms to encode the base image. To simulate this extremely limited encoding time, our model encoded a random subset (1–4 elements) of the attributes from the ideal base representation.

5.2.2. Results

Because the attributes encoded in the limited encoding time condition were chosen randomly, we ran the limited encoding time simulation 30 times for each base image/target image combination and averaged the results. Afterwards, we averaged the results across the six stimulus sets to get overall results. The error rate results are shown in Fig. 12 (compare this to the human results in Fig. 6). Because participants had unlimited time to make the comparison in this task, error rates were based on failure to distinguish between the base and target using either decision criterion.

The simulated error rates line up well with the human results found by Sloutsky and Yarlas. Given ample time to encode the base, participants were highly accurate in distinguishing the target from the base for all target types. When there was limited time to encode the base, accuracy went down a small amount for the E− targets, the targets that differed from the base in their attributes. However, the accuracy went down much more for the E+/R− targets, the targets that differed from the base only in their relations.

We also simulated the reaction times for the same-different judgment. As in the previous simulation, fast responses were assumed for cases when the feature overlap score fell below 0.6. The results are shown in Fig. 13 (compare this to Fig. 7). The model correctly predicted that the response time would be faster for the E− target images (the attribute mismatches) than for the E+/R− target images (the relational mismatches), regardless of the encoding time condition.

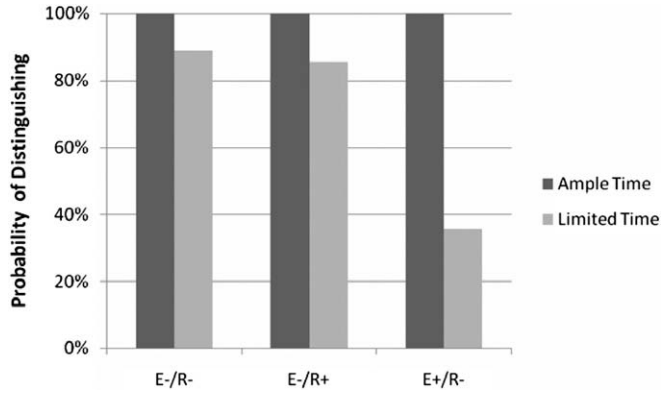

Fig. 12. Predicted error rates in the Sloutsky and Yarlas (submitted for publication) simulation.

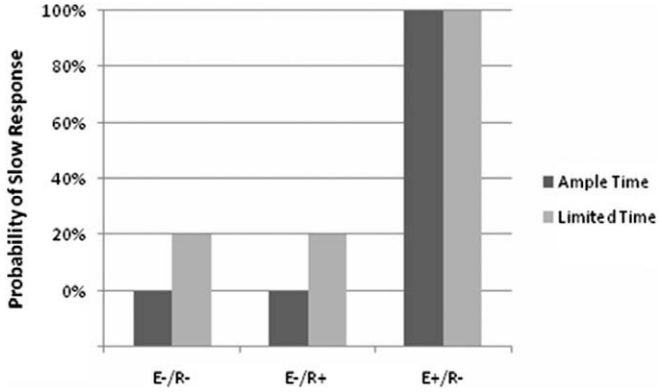

Fig. 13. Predicted reaction times in the Sloutsky and Yarlas (submitted for publication) simulation.

5.3. Simulating a same-different task with limited comparison time

5.3.1. Simulation

Goldstone and Medin's (1994) butterfly experiment is more difficult to simulate because the stimuli relied on a certain amount of semantic knowledge. Participants had to be able to segment the butterflies into four parts (head, tails, body, and wings) and assign attributes based on the shape or texture of each part. To simplify this task, we sketched each butterfly as four glyphs, corresponding to the four parts. Because there is no particular significance to one feature or another, we replaced all butterfly part attributes with colors. We used a different set of colors for each of the different butterfly parts to ensure that there would be no cross-mappings between, say, one butterfly's head and another butterfly's tail. After drawing four glyphs for a butterfly's four parts, we grouped them together using CogSketch's manual grouping function. This causes CogSketch to create a new group glyph and assert part-whole relations between the individual glyphs and the group glyph in its representation.

We sketched a single base image (Fig. 14) and 13 target images, representing the variations of MIPs and MOPs examined by Goldstone and Medin in their study. In the original study, the base images varied in the shape or texture assigned to each butterfly part. However, because there is no theoretical difference between one shape or texture and another in the original study and no functional difference between one color and another in our simulation,

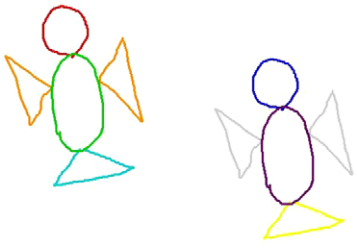

Fig. 14. Stimulus sketched for the Goldstone and Medin (1994) simulation.

a single base image is sufficient for simulation. Since participants were instructed to ignore the relative positions of the butterflies, the spatial relations that CogSketch normally computes between glyphs were not included. Only the attributes of the individual glyphs (colors) and the part-whole relationships that tied the parts of a butterfly together were included.

To conduct this simulation, we assumed that in the short deadline condition, participants only had time to compute the feature overlap between the base and target. In the medium and long deadline conditions, participants would have had time to encode relations and calculate the candidate inferences between the base and target. Because same-different judgments had to be made under a tight deadline, we assume that in both cases, as the evidence for "different" decreased (as the feature overlap rose towards 1.0, and as the number of candidate inferences fell towards 0), the error rate would increase.

5.3.2. Results

As in the original study, we focused on the effect that adding one or two MIPs (attribute matches consistent with the global mapping) or MOPs (attribute cross-matches) had on the decision criteria. Our results are shown in Fig. 15 (compare to the human results in Fig. 4). As the graph shows, when the model was only able to compute a feature overlap score, MIPs and MOPs contributed equally to that score. Each MIP or MOP increased the number of match hypotheses between attributes in the base and target. However, after SME computed a global mapping between the base and target, only the MIPs contributed to the perceived similarity of the stimuli, as indicated in a drop in the number of candidate inferences detected.

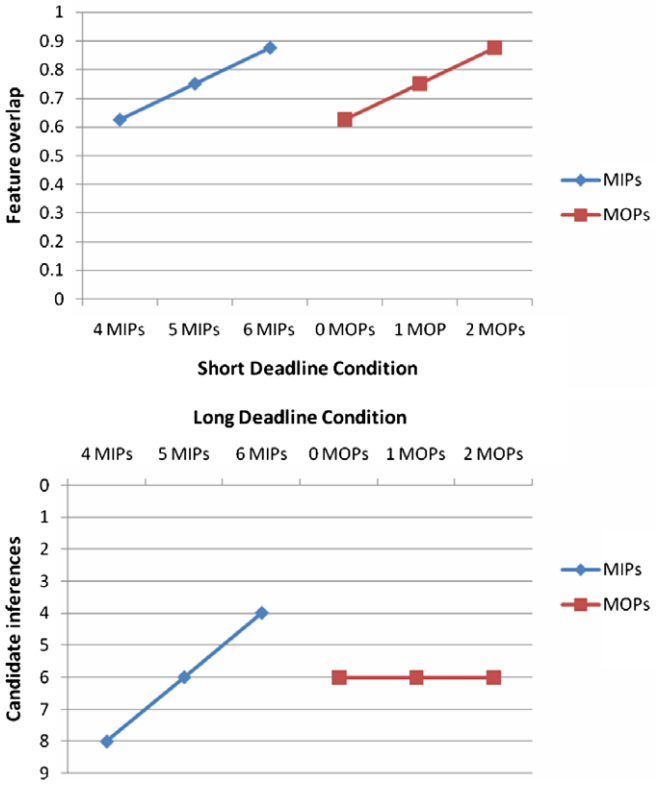

Fig. 15. Results from the Goldstone and Medin (1994) simulation.

6. Related work

Evans (1968) was the first to show that analogical comparison could be modeled in the visual domain. His early system was able to solve geometric-analogy problems, i.e., problems of the form "A is to B as C is to...?" Evans' system was an important milestone historically, but it was only designed to model the output behavior of people (i.e., to choose the same answer that people did). It performed analogical mapping via brute-force search, and thus it could not be used to produce predictions of human timing data.

Goldstone's SIAM model (1994) is a connectionist model of similarity which builds upon the idea of analogical mapping. It uses an interactive activation process between features and objects. Goldstone and Medin (1994) used SIAM to successfully replicate the results from their butterfly same-different task, thus demonstrating that it could explain the time course of similarity in that task. However, SIAM was limited in two respects: (1) it could only perform similarity comparisons, not abstract analogical mapping; and (2) it operated on hand-coded representations. Rather than simulating the encoding process, Goldstone and Medin made the assumption that the base and target images would be encoded in their entirety before the comparison process began.

Recently, it has been argued that SIAM models similarity more accurately than SME because it can account for the effect of attribute cross-matches (MOPs) on similarity. Larkey and Markman (2005) conducted a study in which participants were shown pairs of images and asked to rate their similarity on a numerical scale. Each image consisted of two geometric shapes that varied along two dimensions: shape and color/texture. The experimenters found that participants' similarity scores increased for every common attribute between the two images being compared. Even those common attributes that were not part of the best overall global mapping between the shapes, i.e., the MOPs, contributed to similarity, although MIPs contributed more. In contrast, SME's structural evaluation scores are based only on those correspondences found within the global mapping.[6] The experimenters argued that this result demonstrated a weakness of SME as a model of similarity.

[6] This critique does not characterize SME entirely correctly. While it is true that attribute cross-matches (MOPs) like those described in this study do not affect SME's mappings, it is possible for relation cross-matches to affect mappings. In particular, when match hypotheses are first constructed and scored, a match hypothesis may receive initial support from a parent match hypothesis (a match hypothesis between relations whose arguments are matched in the present match hypothesis), even though that parent match hypothesis may not end up in the same global mapping.

We believe there are two problems with this critique. First, we believe that when participants compare simple stimuli with very little structure, they are more likely to attend to commonalities outside of the mapping between the stimuli. This is particularly true when participants are instructed to rate similarity among many very simple pairs, a task that encourages them to search for any factor that can differentiate the pairs. Essentially, we believe that the actual global mappings are so trivially small that participants implicitly see further criteria, such as the presence of a second possible mapping, on which to rate the pairs. We suspect that as the complexity of the stimuli increases, the influence of MOPs will decrease.

In addition, we believe that in Larkey and Markman's (2005) comparison of the computational models, they are failing to distinguish between the mapping process and the similarity function. Because SIAM is a connectionist model, its output is a set of node activations. In order to produce a similarity measure from this, a similarity function is run over all the activations. Nodes that are more active, such as those consistent with the best overall mapping (MIPs), are weighted more highly than nodes which are less active, such as those inconsistent with the overall mapping (MOPs). Thus, the prediction that MIPs will contribute strongly to similarity while MOPs will contribute weakly is as much a product of this function as it is a product of the mapping process itself.

The output of SME is a structurally consistent mapping between elements in the base and elements in the target. SME's similarity score is based upon the elements in and the structural depth of this mapping. However, in order to compute the mapping, SME finds all possible correspondences between the base and target. Thus, it would be quite easy to apply a similarity function to a completed SME mapping that allowed correspondences not in the mapping to contribute. Such a function, like the one used in SIAM, could weight correspondences outside of the mapping lower than correspondences within the mapping. However, until there is more evidence that MOPs play a significant role in similarity, we believe it is better to leave SME's similarity score unchanged.

7. Discussion

By combining an incremental encoding process with SME's multi-stage analogical mapping process, our model explains results across three same-different tasks that vary in the time available for encoding and comparing the stimuli. The encoding process (using CogSketch) computes representations automatically from stimuli that are similar or identical to those shown to human participants. The comparison process makes use of the multiple stages in the Structure Mapping Engine to compute decision criteria whose profiles (including accuracy) vary according to the time available for comparison. Because SME's operation here uses the same processes that have been used to simulate a large set of analogical tasks, the simulations demonstrate that low-level perceptual similarity judgments can be made using the same comparison process used in high-level analogical reasoning.

Because our model of similarity distinguishes between the encoding and comparison processes, it is able to make a more complete set of predictions than models that begin with a fully encoded representation. Any time participants have ample time and resources, their similarity comparisons should be based on the best global mapping between the stimuli being compared (we believe the question of whether MOPs should contribute to similarity remains open). However, any task that interferes with either participants' perception of the stimuli or their comparison of the stimuli will result in their computing a more superficial comparison based primarily on the attributes common to the stimuli. Thus far, studies have demonstrated this effect when there is limited time for encoding stimuli (Sloutsky & Yarlas, submitted for publication), or when there is limited for both encoding and comparing (Goldstone & Medin, 1994). However, no study has yet isolated the comparison task from the encoding task and demonstrated the predicted effect when there is limited time only for comparison. We believe this is an important direction for further studies of the time course of similarity.

Acknowledgement

This work was supported by NSF SLC Grant SBE-0541957, the Spatial Intelligence and Learning Center (SILC).

References

Dunbar, K. (1999). The scientist in vivo: How scientists think and reason in the laboratory. In L. Magnani, N. Neressian, & P. Thagard (Eds.), *Model-based reasoning in scientific discovery*. Minneapolis, MN: University of Minnesota Press.

Evans, T. (1968). A program for the solution of geometric-analogy intelligence test questions. In M. Minsky (Ed.), *Semantic information processing*. Cambridge, MA: MIT Press.

Falkenhainer, B., Forbus, K. D., & Gentner, D. (1989). The structure-mapping engine: Algorithm and examples. *Artificial Intelligence, 41*, 1–63.

Farell, B. (1985). Same-difference judgments: A review of current controversies in perceptual comparisons. *Psychological Bulletin, 98*, 419–456.

Forbus, K., Ferguson, R., & Gentner, D. (1994). Incremental structure-mapping. In *Proceedings of the 16th annual meeting of the cognitive science society*.

Forbus, K., & Oblinger, D. (1990). Making SME greedy and pragmatic. In *Proceedings of the 12th annual meeting of the cognitive science society*.

Forbus, K., Usher, J., Lovett, A., Lockwood, K., & Wetzel, J. (2008). CogSketch: Open-domain sketch understanding for cognitive science research and for education. In *Proceedings of the fifth eurographics workshop on sketch-based interfaces and modeling*. Annecy, France.

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

Gentner, D., & Gunn, V. (2001). Structural alignment facilitates the noticing of differences. *Memory and Cognition, 29*, 565–577.

Gentner, D., & Markman, A. B. (1997). Structure mapping in analogy and similarity. *American Psychologist, 52*, 42–56.

Gentner, D., & Namy, L. L. (2006). Analogical processes in language learning. *Current Directions in Psychological Science, 15*, 297–301.

Gentner, D., Rattermann, M. J., Markman, A. B., & Kotovsky, L. (1995). Two forces in the development of relational similarity. In T. J. Simon & G. S. Halford (Eds.), *Developing cognitive competence: New approaches to process modeling* (pp. 263–313). Hillsdale, NJ: LEA.

Gentner, D., Ratterman, M. J., & Forbus, K. (1993). The roles of similarity in transfer. *Cognitive Psychology, 25*, 524–575.

Gentner, D., & Sagi, E. (2006). Does "different" imply a difference? A comparison of two tasks. In *Proceedings of the 28th annual meeting of the cognitive science society*.

Gentner, D., & Toupin, C. (1986). Systematicity and surface similarity in the development of analogy. *Cognitive Science, 10*, 277–300.

Goldstone, R. L. (1994). Similarity, interactive activation, and mapping. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*, 3–28.

Goldstone, R. L., & Medin, D. L. (1994). Time course of comparison. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*, 29–50.

Holyoak, K., & Thagard, P. (1995). *Mental leaps: Analogy in creative thought*. MIT Press.

Larkey, L. B., & Markman, A. B. (2005). Processes of similarity judgment. *Cognitive Science, 29*, 1061–1076.

Lockwood, K., Forbus, K., Halstead, D., & Usher, J. (2006). Automatic categorization of spatial prepositions. In *Proceedings of the 28th annual conference of the cognitive science society*. Vancouver, Canada.

Lovett, A., Forbus, K., & Usher, J. (2007). Analogy with qualitative spatial representations can simulate solving Raven's Progressive Matrices. In *Proceedings of the 29th annual conference of the cognitive science society*.

Lovett, A., Gentner, D., & Forbus, K. (2006). Simulating time-course phenomena in perceptual similarity via incremental encoding. In *Proceedings of the 28th annual meeting of the cognitive science society*.

Markman, A. B., & Gentner, D. (1996). Commonalities and differences in similarity comparisons. *Memory and Cognition, 24*(2), 235–249.

Medin, D. L., Goldstone, R. L., & Gentner, D. (1993). Respects for similarity. *Psychological Review, 100*(2), 254–278.

Posner, M. I., & Mitchell, R. F. (1967). Chronometric analysis of classification. *Psychological Review, 74*, 392–409.

Sagi, E., Gentner, D, & Lovett, A. (in preparation). "Different" need not entail a difference: A telling disassociation in the psychology of difference.

Sloutsky, V. M., & Yarlas, A. S. (submitted for publication). Processing of information structure: Mental representations of elements and relations.

Tversky, B. (1969). Pictorial and verbal encoding in a short-term memory task. *Perception and Psychophysics, 6*(4), 225–233.