ELSEVIER

Cognitive Science 27 (2003) 41–71

COGNITIVE SCIENCE

http://www.elsevier.com/locate/cogsci

Semantic grounding in models of analogy: an environmental approach

Michael Ramscar[a,b,*], Daniel Yarlett[a,b]

[a] *Edinburgh University, Edinburgh, Scotland, UK*
[b] *Department of Psychology, Stanford University, 450 Serra Mall, Jordan Hall, Stanford, CA 94305-2130, USA*

Received 20 March 2001; received in revised form 21 September 2001; accepted 28 May 2002

Abstract

Empirical studies indicate that analogy consists of two main processes: retrieval and mapping. While current theories and models of analogy have revealed much about the mainly structural constraints that govern the mapping process, the similarities that underpin the correspondences between individual representational elements and drive retrieval are understood in less detail. In existing models symbol similarities are externally defined but neither empirically grounded nor theoretically justified. This paper introduces a new model (EMMA: the environmental model of analogy) which relies on co-occurrence information provided by LSA (*Latent Semantic Analysis*; Landauer & Dumais, 1997) to ground the relations between the symbolic elements aligned in analogy. LSA calculates a contextual distribution for each word encountered in a corpus by counting the frequency with which it co-occurs with other words. This information is used to define a model that locates each word encountered in a high-dimensional space, with relations between elements in this space representing contextual similarities between words. A series of simulation experiments demonstrate that the environmental approach to semantics embodied in LSA can produce appropriate patterns of analogical retrieval, but that this semantic measure is not sufficient to model analogical mapping. The implications of these findings, both for theories of representation in analogy research and more general theories of semantics in cognition, are explored.

Keywords: Analogy; Semantics; Retrieval; Implications

* Corresponding author. Tel.: +1-650-725-2400; fax: +1-650-725-5699.
E-mail address: michael@psych.standford.edu (M. Ramscar).

0364-0213/02/$ – see front matter © 2002 Cognitive Science Society, Inc. All rights reserved.
PII: S0364-0213(02)00113-1

1. Introduction

"How do humans reason by analogy?" To take the case of the familiar analogy between the solar system and Rutherford's model of the atom: how does one of these seemingly disparate things remind us of the other in the first place? And on what basis are we able to make meaningful correspondences between them? Research in cognitive science has been remarkably successful at providing answers to these questions, at least in broad terms (Gentner, 1983; Goswami, 1992; Holyoak & Thagard, 1995; Hummel & Holyoak, 1997).

Retrieval (or reminding) is driven by a computationally inexpensive process that initially matches the semantics of surface elements in representations in a purely local fashion (e.g., the roundness of bicycle wheels and spectacles may lead to spectacles reminding us of a bicycle; see Gentner, Rattermann, & Forbus, 1993). Analogical mappings, on the other hand, are determined by a more computationally expensive process that is driven by global constraints. Systematic structural similarities between the items to be matched need to be detected in order to make the kind of "deeper," inference supporting correspondences that characterize analogy. (Despite the many surface similarities between *Romeo and Juliet* and *The Two Gentlemen of Verona*, underlying structural similarities of plot make the former seem more similar to *West Side Story* than to the latter; Gentner, 1983; Hummel & Holyoak, 1997).

The explanations offered by current theories and models of analogy appear compelling at one level of abstraction. However, all analogical theories and explanations make assumptions that beg important questions if one seeks a more thorough explanation of these processes. As one examines the question "How do humans reason by analogy?" in more detail, it appears that there are important gaps in existing theories and process models. This paper considers just one of these gaps: the problem of matching the "semantics" of elements during the retrieval and mapping phases of the analogical reasoning process.

This problem can be illustrated as follows. Suppose that in your representation of the atom you describe the motion of an electron in relation to the nucleus in terms of it "revolving around" it (perhaps this is how you ordinarily think about this motion). And on the other hand, suppose that in your representation of the solar system you conceive the motion of the planets in terms of their "orbiting" the sun. At one level of abstraction it may be sufficient to say that (as yet unspecified) similarities in the meanings of (or the concepts underlying) these words explains how one prompts the reminding of the other in the context of retrieval, or why they might be mapped on to one another should surrounding structures warrant it. However, in a more detailed account—and model—of analogy we might wish to do more than appeal to such ill specified intuitions about similarities of meaning. We might wish to account for the way in which these sub-elements of our representations of the atom and the solar system prime and are mapped onto one another in the same level of detail with which we account for the mappings between the representations themselves.

To fully explain high-level mapping in analogy, one must account for the way distinct but "semantically" similar items in representations are reconciled with each other in a way that allows representations to be retrieved and mappings to be made. This paper reviews the role that semantics plays in theories of analogy, subjecting current approaches to analogical semantics to a considered critique. The shortcomings in existing approaches are used to motivate an alternative approach to modeling semantics in retrieval, using a similarity metric derived from the linguistic environment (LSA, Landauer & Dumais, 1997). Then a series of simulations demonstrate that this kind of similarity metric can be used to model the availability of analogs in different contexts. However, they show that environmental information alone is insufficient to account for the process of structural alignment (e.g., Gentner, 1983) in analogy. Finally, the implications of these findings are discussed, and the need for more dynamic models (both theoretical and computational) of the retrieval and mapping processes is identified.

2. Semantic matching and the re-representation hypothesis

Perhaps the most straightforward way to explain the matching between distinct but similar ideas—a process required for both retrieval and mapping in analogy—is in conceptual terms. For example, if the concepts of "revolving around" and "orbiting" could be shown to decompose into some canonical conceptual representation (e.g., "circumnavigating"), then the link between them could be explained by reference to that common underlying concept. This proposal is put forward by Gentner et al. (1993, p. 553):

> [the] constraint of matching identical predicates assumes canonical conceptual representations, not lexical strings. Two concepts that are similar but not identical (such as "bestow" and "bequeath") are assumed to be decomposed into a canonical representation language so that their similarity is expressed as a partial identity (... [in this case] "give").

The main drawback to this proposal is the lack of any specification of what a canonical conceptual representation (or a canonical representation language) is. Research into the mental representation of concepts suggests that human conceptual representations are anything but canonical; the proposals for generalized theories of representation that exist in the concepts literature fall well short of providing the kind of "neat" account of concepts that a theory of canonical conceptual representation assumes (see Komatsu, 1992; Ramscar & Hahn, 1998, for reviews). This problem has not gone unrecognized. In conjunction with other factors, such as evidence of the important role that structural commonalities play in "ordinary" conceptual tasks (e.g., Ahn, 1998), and the sheer difficulty of distinguishing analogy from "ordinary" conceptual tasks (Ramscar & Pain, 1996), a widespread view has emerged that suggests that analogy itself may play an important role in the process of semantic matching (Forbus, Gentner, Markman, & Ferguson, 1998; Hummel & Holyoak, 1997).

The basic idea behind the canonical representation proposal is outlined by Forbus et al. (1998), who propose that semantic terms might be decomposed into sub-predicate re-representations, with matches between these being determined recursively by the same mapping process operative at the top-level of representation:

> re-representation allows relational identicality to arise out of ... analogical alignment, rather than as a strict constraint on the input descriptions. Forbus et al. (1998, p. 246)

A similar re-representation proposal is also advanced by Hummel and Holyoak (1997):

> With the notion of chunked predicates and objects, LISA hints at a kind of recursive representation for meaning that may ultimately ground itself in basic perceptual primitives. In its current implementation, LISA can represent and map hierarchical propositions of arbitrary depth ... Analogously, it is

> possible to imagine structures for roles and objects that are, themselves, deeply embedded recursive structures. The depth to which a role or object would need to be decomposed for the purposes of mapping would depend on the task at hand. For example, mapping "John lifted the hammer" onto "Bill raised the book" may require little or no decomposition of the predicates "lift" and "raise," which will have substantial overlap in their semantic features. On the other hand, mapping "John lifted the hammer" onto "Bill pushed the cart," where the predicates have less feature overlap, may be more likely to depend on decomposition of "lift" into "cause to rise" and "push" into "cause to move laterally," thereby making explicit the parallelism of their internal structures. Recursively, "rise" and "move laterally" might be decomposed into structures relating simpler predicates, with basic perceptual primitives representing motion and locations in space residing at the very bottom. Hummel and Holyoak (1997, p. 457)

Whilst re-representation is a popular idea in the analogy literature, its current status is largely hypothetical: re-representation proposals are usually couched in terms that relate to computational models, and as yet no evidence has been offered to support the psychological validity of the proposal. Furthermore, a process of recursive decomposition and matching via mapping, even if it descends only a few levels, opens up the possibility of a prohibitive combinatorial explosion in analogies between complex, real-world representations. Moreover, if the recursive decomposition proposal is not to result in an infinite regress, the terminating layer of semantic elements must be grounded in some way.

As reviewed below, studies have indicated that retrieval acts as a cheap pre-filter for the more computationally expensive—and conceptually rich—process of analogical structure-mapping (e.g., Gentner et al., 1993). However, re-representation in retrieval would imply multiple structural mappings to be carried out at the retrieval stage (as many as there are lexically distinct but potentially "semantically" similar items in any representations to be mapped). This is hardly compatible with the idea of retrieval as a computationally cheap process. Since re-representation seems to offer a particularly implausible account of semantic matching in retrieval, this investigation into modeling semantics in analogy begins with a detailed examination of the constraints that empirical results place on models of retrieval.

2.1. Retrieval and semantic matching

2.1.1. Four constraints on retrieval

Empirical studies conducted by Gentner et al. (1993) established that similarities between surface (or semantic) features of representations of analogs are largely responsible for their retrieval from long-term memory. From their studies, Gentner et al. derived four primary constraints that an appropriate model of retrieval should meet when given a specific probe item:

1. *Primacy of the mundane:* The majority of retrievals evoked should be literally similar to the probe item, meaning that they share both surface and structural characteristics (e.g., a bicycle should result in the activation of representations of other bicycles that have been encountered previously).
2. *Surface superiority:* Retrievals based on surface similarity alone (without structural similarity) should also be frequent (e.g., a fairy story about a frog might call to mind knowledge of other frogs, and castles and wells, although there may be no structural congruities between the retrieved items and the probe item).

3. *Rare insights:* Memories that are merely structurally similar to the probe item should be retrieved only occasionally (e.g., the orbits of the solar system reminding one of electrons orbiting an atom; see also Gick & Holyoak, 1980).
4. *Scalability:* The model must plausibly extend to realistically sized memory pools because people typically have vast numbers of memories that they are able to access in a matter of seconds.

In sum, Gentner et al.'s (1993) investigation demonstrated that retrieval is sensitive more to surface (or "semantic," Hummel & Holyoak, 1997) similarities between a target representation and a base analogy that needs to be to be retrieved, than to the shared relational structure which facilitates an analogical match. The picture of retrieval that emerges from their investigation is that the retrieval process, being relatively computationally cheap, acts as an efficient prefilter to the more expensive process of structural alignment (albeit at the expense of potentially passing over useful analogies that share structural commonalities with the target domain; see Duncker, 1945; Gick & Holyoak, 1980).

3. Environmental models of semantics

A promising approach to capturing some properties of lexico-semantic knowledge—enabling at least some of the semantic similarities between words to be determined—is to examine the distributional characteristics of the contexts in which words are used in the language stream. This is a data-intensive technique that analyzes a corpus of natural language output, and from it derives summarized information about the variety of different contexts that different words are used in. Such techniques measure the frequency with which different words co-occur with one another (that is, are used together within a particular context, such as a paragraph or moving-window). There is a growing body of evidence that co-occurrences provide useful information about the semantic properties of individual words (Burgess & Lund, 1997; Landauer & Dumais, 1997; Landauer, Foltz, & Laham, 1998).

In co-occurrence analyses, a contextual distribution is calculated for each word encountered in a corpus by counting the frequency with which it co-occurs with every other word in the corpus being analyzed. The contextual distribution of a word can then be represented as a vector showing the frequency with which it is encountered with the other words in a common linguistic environment. One can think of this information as defining a model containing a network of links between the words in a language, each with varying strengths, and representing the similarity of the contexts that words are used in, in the language in question. Two such co-occurrence models are the Latent Semantic Analysis (LSA) model (Landauer & Dumais, 1997; Landauer et al., 1998), and the Hyperspace Analog to Language (HAL) model (Burgess & Lund, 1997).

There is good evidence that co-occurrence analysis extracts information from corpora that can be used to model certain linguistic behavior. For example, Landauer and Dumais (1997) report that the LSA model is sufficiently sensitive at recognizing similarities to be able to pass a multiple-choice TOEFL test. Lund, Burgess, and Atchley (1995) present evidence that co-occurrence data can act as a good predictor of various priming effects. Also, Redington, Chater, and Finch (1998) and Burgess and Lund (1997) have shown that co-occurrence techniques appear to provide powerful cues to the syntactic category that words belong to.

Whilst the exact parameters of LSA and HAL are different, they both adopt the general approach outlined above to generate co-occurrence vectors. There are a number of attractive benefits to be gained from modeling the semantic information used in analogical and similarity-based retrieval in this way:

1. The proposed semantic metric is clearly specified. Proposing that the semantic information used in retrieval is learned from observing the varying contextual co-occurrences of words in a language, avoids postulating entities such as semantic primitives, whose theoretical and psychological nature is seriously under-specified.
2. The semantic information used could be easily derived from the environment, avoiding the problems inherent in positing entities whose learnability is somewhat controversial, and whose innateness might otherwise have to be treated as axiomatic (as canonical concepts seem to be; see Fodor, 1981; Laurence & Margiolis, 1999).
3. An environmental context model contains representationally cheap, summarized information, the usage of which makes only limited processing demands. Thus, it allows one to avoid the complexity issues besetting accounts in which retrieval is explained in terms of semantic decomposition and expensive structural alignment (cf. Forbus et al., 1998). This also makes it more likely that such an approach will meet the scalability constraint.
4. Models of context based on environmentally available information are objective: they do not require that a particular set of semantic features are defined before textual analysis begins. The co-occurrence technique takes the words themselves as features, and uses frequency relations between them to define how strongly they are associated. This is an advantage given the difficulty already highlighted of empirically grounding claims as to the identity of semantic features. Because co-occurrence techniques do not rely on a postulated set of semantic features (such as gender, plurality, animacy and so on), this eliminates the subjectivity resulting from decisions about: (i) what semantic primitives are to be used, and (ii) which primitives are exhibited in a particular representation.

3.1. Can co-occurrence data be used to model semantics in retrieval?

The success of co-occurrence techniques in accounting for priming effects (cf. Lund et al., 1995) has shown them to be useful models of lexical retrieval. The following experiments seek to establish whether the LSA model (Landauer & Dumais, 1997; Landauer et al., 1998) can be used to account for the retrieval of structured composite representations, and not just individual words, from a memory-pool.[1]

4. Retrieval experiments

4.1. The "Karla the Hawk" stories

The materials used in the following simulations are the "Karla the Hawk" materials used originally in Gentner et al.'s (1993) study, which consist of 20 sets of stories written in natural language. Each set consists of a base story and four systematic variations of that story, with

Table 1
The commonalities each variant category shares with its corresponding base

	Shared structure	No shared structure
Shared features	Literal similarity	Surface similarity
No shared features	Analogy	First-order relations

two factors (surface similarity and structural similarity) crossed over the four variant stories, as shown in Table 1. An example material set is given in Table 2.

The four story categories systematically vary the commonalities that are shared with the base story from which they are derived. Each variant either provides a surface match and/or a structural match to the corresponding base story. This 2×2 design allows for the controlled examination of the factors that various putative measures of retrieval are sensitive to. In a series of experiments, Gentner et al. (1993): (1) taught subjects the base stories and primed and tested their retrievability with regards to the variant story versions after 1 week's delay; (2) had subjects assess the various mutual similarities holding between the stories; and (3) asked subjects to assess the ability of the various stories to support inferences about one another. Gentner et al. found that the prime determinant of retrievability was shared surface features, whilst shared structural commonalities had a nonsignificant effect on the process.

Table 2
Example set from the "Karla the Hawk" stories (Gentner et al., 1993)

Base

Karla, an old hawk, lived at the top of a tall oak tree. One afternoon, she saw a hunter on the ground with a bow and some crude arrows that had no feathers. The hunter took aim and shot at the hawk but missed. Karla knew the hunter wanted her feathers so she glided down to the hunter and offered to give him a few. The hunter was so grateful that he pledged never to shoot at a hawk again. He went off and shot a deer instead.

Literal similarity

Once there was an eagle named Zardia who nested on a rocky cliff. One day she saw a sportsman coming with a cross-bow and some bolts that had no feathers. The sportsman attacked but the bolts missed. Zardia realized that the sportsman wanted her tailfeathers so she flew down and donated a few of her tailfeathers to the sportsman. The sportsman was pleased. He promised never to attack eagles again.

True analogy

Once there was a small country called Zardia that learned to make the world's smartest computer.

One day Zardia was attacked by it's warlike neighbor, Gagrach. But the missiles were badly aimed and the attack failed. The Zardian government realized that Gagrach wanted Zardian computers so it offered to sell some of it's computers to the country. The government of Gagrach was very pleased. It promised never to attack Zardia again.

First-order relations (mere appearance)

Once there was an eagle named Zardia who donated a few of her tailfeathers to a sportsman so he would promise never to attack eagles.

One day Zardia was resting high on a rocky cliff when she saw the sportsman coming with a cross-bow. Zardia flew down to meet the man, but he attacked and felled her with a single bolt. As she fluttered to the ground Zardia realized that the bolt had her own tailfeathers on it.

However, when subjects were asked to rate the various stories for similarity, and how strongly knowledge of a story would support inferences about another story, Gentner et al. found that subjects based their judgments upon structural similarities—rather than surface similarities—to a highly significant extent (see Fig. 1 for a summary). Conformity to this pattern of results is used to determine the adequacy of modeling performance in the following computational simulations.

4.2. Methods and representations for retrieval

The following retrieval experiments focus on comparing the straightforward feature-matching approach to retrieval, as exemplified in MAC/FAC, against the high-dimensional contextual approach, as implemented in LSA.

MAC/FAC models retrieval by generating a content vector for each representation that is stored in its memory-pool. A content vector summarizes the surface features of a representation by recording the frequency with which each lexically distinct predicate occurs in it. Thus, the following proposition:

```
(CAUSE (STRIKES-WITH JOHN CUE CUE-BALL)
(AND (POTS CUE-BALL) (POTS BLACK)))
```

would be assigned the following content vector:

```
((CAUSE . 1) (STRIKES-WITH . 1) (AND . 1) (POTS . 2))
```

A measure of the degree that two representations share the same surface features can then be derived by calculating the dot-product of their content vectors (if a particular predicate does not appear in a representation then it is implicit, through sparse-encoding, that it has a frequency of zero). It is important to note a consequence of the content vector approach to retrieval: only identical predicates shared between the two domains in question can contribute to the increased magnitude of their dot-product, and hence the retrievability of one representation given the other. There is no potential for multiplying the frequencies of distinct predicates in the dot-product calculation, because vector elements are matched strictly on the basic of lexical-identity (and so the presence of “orbits” and “revolves around” would not serve to increase the mutual retrievability of two domains containing those predicates). Content vectors thus take a straightforward feature-matching approach to retrieval.

In order to model the way that distinct features in stimuli prime one another for retrieval, the content vector proposal (CV) makes a commitment to a canonical representation (CR) theory of mental representation (see also Gentner et al., 1993). As noted earlier, according to CR theory, complex semantic elements can be recursively decomposed—or re-represented—until their canonical microfeatures are reached. At this point, a straightforward feature-matching approach can be used to assess whether their semantic overlap is sufficiently great to permit a match to be made. Hence, CR theory assumes that the mental encoding of semantically complex concepts can ultimately be analyzed in terms of a finite stock of canonical forms (at present the precise nature of these forms is not specified; see also Fodor, 1975, 1981).

The following simulation experiments compared the high-dimensional, contextual approach to modeling semantic features with the content vector method across a range of representational

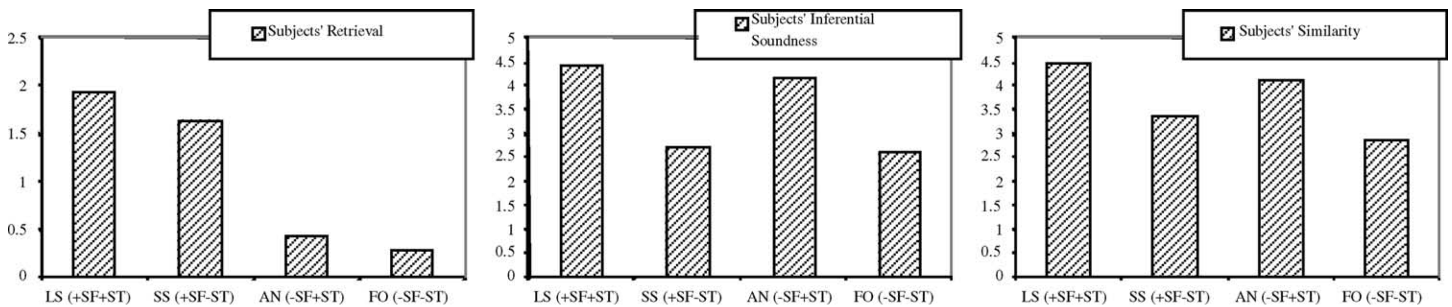

Fig. 1. The retrieval scores, and ratings of retrievability, similarity and inferential soundness, produced by subjects in Gentner et al.'s (1993) studies. Here and throughout, all participant data is on a scale 1 (least) to 5 (most).

formats to determine whether the former might offer an alternative paradigm for modeling of retrieval.

4.3. Simulation Experiment 1: natural language representations

Experiment 1 examines the natural language versions of the Karla the Hawk materials in order to determine whether appropriate patterns of retrieval could be produced by both models on raw representations requiring little or no preprocessing. Two simulations were conducted: the first used the natural language materials presented to participants in Gentner et al. (1993) verbatim, whilst the second used stripped versions of these texts to determine whether there is sufficient informational content in a reduced representation of the natural language versions of the Karla the Hawk materials to produce retrieval patterns conformable to the empirical data.

While verbatim texts provide a good initial test of the two theories of retrieval in a context where no representational assumptions have to be made, there is clear empirical evidence that people do not rely on such verbatim representations in retrieval. In addition to the accretion of structural information during comprehension, there is a concomitant loss of superficial verbatim information as internal representations are built up (Bransford & Johnson, 1972; Gernsbacher, 1985; Sachs, 1967). Since we wanted to simulate retrieval of what subjects in Gentner et al.'s studies actually stored, the second simulation in Experiment 1 tested the retrieval models on versions of Gentner et al.'s stimuli that had all of the closed-class[2] words removed from them.

Applying this principle resulted in a set of words for each story that constituted the words that are, in some sense, maximally informative about the context that the representation defines. For example, some words (particularly the closed-class words) may occur in almost any and every possible context (e.g., "the" can, and does, co-occur plausibly with an extremely diverse set of words). Thus, encountering such a word in a probe representation has little informational utility with respect to retrieval because it fails to narrow the set of candidate retrievals at all (this is in contrast to a word like "heartbroken," which occurs only in a very specific set of contexts).

The original Karla the Hawk base-story, after it had been pruned of all closed-class words, is given below as an example of the "bag of words" that remained once the natural language representations had been stripped of closed-class words:

```
Karla old hawk lived top tall oak tree afternoon saw
  hunter ground bow crude arrows feathers hunter aim
  shot hawk missed Karla knew hunter wanted feathers
  glided down hunter offered give hunter grateful
  pledged shoot hawk shot deer.
```

4.3.1. Method

The base of each of the 20 natural language material sets was compared to each of its four variants using the two different models of retrieval (in order to reproduce the experimental format embodied in Gentner et al.'s original retrieval experiments). Because MAC/FAC's content vector approach to retrieval was designed to operate on specially coded structured representations, it was adapted for use on the natural language materials by treating each word as a distinct feature, and thus generating content vectors based on the frequency with which distinct words occurred in each representation. The LSA model[3] was set to make its comparisons in document-to-document mode, using the first 300 factors of the "General Reading up to 1st_year college" training set (which most closely reflecting the linguistic exposure of the original undergraduate subjects in Gentner et al.'s study).

The 2×2 design of the experiment allowed repeated-measures ANOVA analyses (treating the items in material sets as a random-effect; see Clark, 1973, for a fuller discussion) to be conducted in order to determine which of the factors—surface similarity (±SF) or structural similarity (±ST)—the two metrics were sensitive to, and whether their performance could be expected to generalize beyond the specific materials examined to the population they were sampled from (these analyses are reported throughout the following simulation experiments).

4.3.2. Results

The results of the simulations are shown in Fig. 2.

4.3.2.1. Verbatim texts. ANOVA analysis revealed that for the content vector comparisons there was a main effect of surface similarity ($F(1, 19) = 13.870$, $p < .001$); a main effect of structural similarity ($F(1, 19) = 4.716$, $p < .05$); and a marginal interaction effect ($F(1, 19) = 3.557$, $p = .07$). For the LSA comparisons there was a main effect of surface similarity ($F(1, 19) = 55.161$, $p < .001$). There was also a marginal effect of structural similarity ($F(1, 19) = 3.180$, $p = .09$), and a marginal interaction between the factors ($F(1, 19) = 2.863$, $p = .11$).

4.3.2.2. Stripped texts (minus closed-class words). ANOVA analysis revealed that the content vector metric was sensitive to both surface similarity ($F(1, 19) = 11.965$, $p < .005$) and structural similarity ($F(1, 19) = 10.027$, $p < .005$), with a marginally significant interaction effect ($F(1, 19) = 3.717$, $p = .07$). For the LSA metric there was a main effect of surface similarity ($F(1, 19) = 68.985$, $p < .001$); a marginal effect of structural similarity ($F(1, 19) = 2.611$, $p = .12$), and a marginal interaction between the factors ($F(1, 19) = 2.428$, $p = .14$).

4.3.3. Discussion

In the first simulation, both measures are sensitive to the surface features of the verbatim materials in forming their retrieval patterns, in accordance with Gentner et al.'s (1993) retrieval findings, although this effect is less clear-cut for the content vector metric. The clustering in the mean LSA scores for each category of variant (LS and SS being similarly matched, and AN and FO scores being similarly matched) mirrored the subject data in Gentner et al.'s (1993) study closely (Fig. 2).

In the second simulation, whilst the LSA metric was primarily sensitive to the surface features of the stripped representations, the content vector metric showed a marked sensitivity to the structural characteristics of the representations—indeed, the content vector metric showed no preference at all for semantic over structural information.

These initial results show that LSA models the original empirical data more accurately than the content vector approach. It is also interesting to note that the LSA retrieval scores remain more or less the same in both of these simulations, whilst the content vector scores are

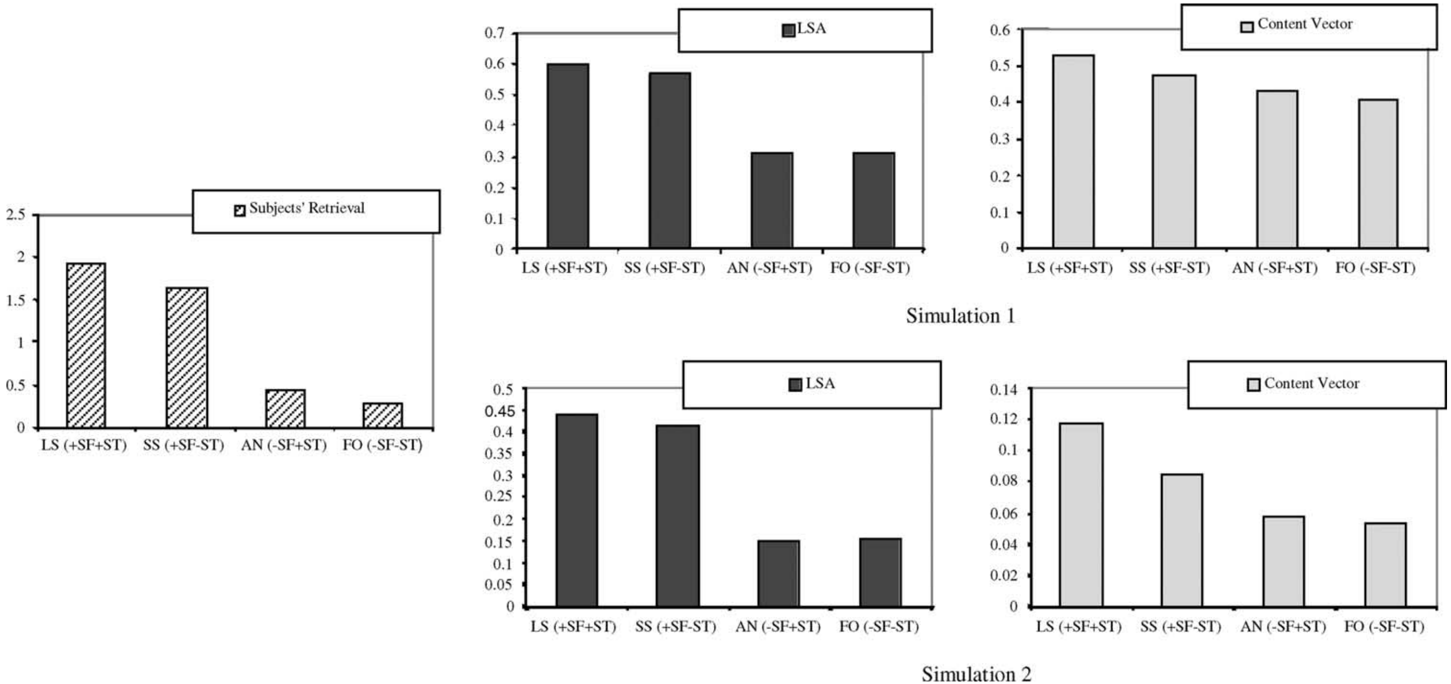

Fig. 2. Subject retrieval scores and content vector and LSA scores for each of the four variants as compared to the base stories. Materials are in their original, natural language form (Simulation 1) and the stripped ("bag-of-words") natural language form (Simulation 2). Note that here and elsewhere, all LSA scores reported are cosine values.

significantly reduced when the stripped texts are compared.[4] This seems to suggest that the LSA model provides a more robust measure of retrieval across varying representational formats.

4.4. Simulation Experiment 2: Original Dgroups

To test the MAC/FAC model, Forbus, Gentner, and Law (1994) developed "Dgroup" representations of 9 of the original 20 Karla the Hawk stories (we call these the *Original Dgroups*), e.g.

```
Base:

  (sme:defDescription base-5
    entities (Karla man1 feathers cross-bow F high deer loc1)
    expressions (((bird Karla)   :name isa-bird))
      ((hawk Karla)              :name isa-hawk))
      ((large Karla)             :name is-large))
      ((powerful Karla)          :name is-powerful))
      ((black Karla)             :name is-black))
      ((predatory Karla          :name is-predatory))
      ...

Literal Similarity Variant:

  (sme:defDescription ls-5
    entities (Zerdia man1 feathers cross-bow T F high)
    expressions (((bird Zerdia)  :name isa-bird))
      ((eagle Zerdia)            :name isa-eagle))
      ((large Zerdia)            :name is-large))
      ((predatory Zerdia)        :name is-predatory))
      ((brown Zerdia)            :name is-brown))
      ((powerful Zerdia)         :name is-powerful))
      ...
```

Dgroups are the standard representational format used by both MAC/FAC and SME to explicitly encode the information—such as higher-order causal and temporal relations—that is often only implicit in natural language. The typed predicate-calculus style language of the Dgroups is used to represent both surface and structural properties in an unambiguous form, permitting analogical mappings to be computationally generated across them. Evidence presented in Forbus et al. (1994) demonstrates that the content vector approach produces appropriate patterns of retrieval on the Original Dgroups. The following simulation experiment tested whether the LSA model was sensitive to the information carried by the Dgroups and whether it too could produce appropriate retrieval patterns.

4.4.1. Method

The LSA and content vector models were used to compare the base of each Original Dgroup set with its four variants for each of the 9 material sets that were originally coded by Forbus et al. (1994). MAC/FAC already contained routines to extract the surface information from a Dgroup to complete its content vector comparisons; these routines were also used to extract the information to be used for LSA comparison. The LSA model was used with the same settings as in previous experiments.

4.4.2. Results

The results of the simulations are shown in Fig. 3. In accordance with the previous data regarding MAC/FAC's performance, analysis revealed that for the content vector metric there was a main effect of surface similarity ($F(1, 8) = 84.396$, $p < .001$); a marginal structural similarity effect ($F(1, 8) = 4.716$, $p = .06$); and a marginal interaction ($F(1, 8) = 3.557$, $p = .10$). Analysis of the LSA metric revealed a main effect of surface similarity ($F(1, 8) = 74.974$, $p < .001$); no effect of structural similarity ($F(1, 8) < 1$); and no significant interaction between the factors ($F(1, 8) < 1$).

4.4.3. Discussion

As reported in Forbus et al. (1994), the content vector method exhibited sensitivity primarily to the surface characteristics of the materials. This is in agreement with the pattern of results produced by human subjects in Gentner et al. (1993). Significantly, the LSA metric also produces the correct retrieval patterns, being sensitive only to the surface characteristics of the materials. These results indicate that the LSA metric can produce appropriate patterns of retrieval from representations that are amenable to analogical mapping.

Further, the LSA metric offers a potential further advantage over the content vector metric. The Original Dgroups encode surface information that is not explicitly contained in the natural language materials from which they are derived. Comparing the Original Dgroup examples above with the original materials in Table 2, it is clear that much of the surface information in the Original Dgroups is not explicitly encoded in the natural language materials: for example, nowhere in the natural language materials is it explicitly mentioned that either Karla or Zerdia has the attribute of being either a bird, large, powerful or predatory, and yet these predicates are included in the corresponding Original Dgroup.

Although the addition of this information is justified under the assumption of a canonical theory of representation—it is postulated that the canonical representation of a concept decomposes into a specific set of features, i.e., eagle decomposes into a set of semantic primitives including largeness, powerfulness, etc.—the considerations advanced earlier against canonical theories of representation make it worth testing how well the two retrieval metrics perform on representations devoid of this sort of additional information. If fewer unconstrained representational features can be used in modeling, it may by possible to produce successful retrieval in a model with one less free parameter than MAC/FAC has at present. This would constitute a significant improvement in the modeling of the retrieval process in analogy.

To test this, surface information that was not directly warranted by the original natural language materials was removed from the representations used in the following simulation experiment. Further, because the Original Dgroups also failed to include lexical information that *was* present in the original natural language materials, a new series of Dgroups—the *Faithful Dgroups*—was developed to test the retrieval metrics on a representational format

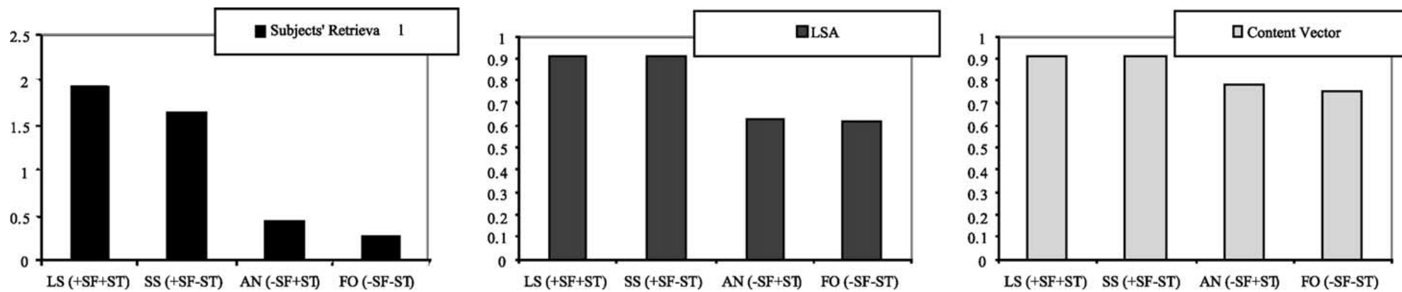

Fig. 3. Subject retrieval scores and content vector and LSA scores for each of the four variants as compared to the base stories. Materials are in Original Dgroup form.

both amenable to analogical alignment, and that reflect the lexical properties of the natural language materials as closely as possible.

4.5. Simulation Experiment 3: Stripped and Faithful Dgroups

Two simulations were conducted. In the first, surface information unwarranted by the original natural language materials was removed from the Original Dgroups, to see how reliant the retrieval metrics are on assumptions made in line with CR theory (e.g., because Karla is not explicitly described as being a bird, large, black, powerful or predatory, in the natural language materials, these statements were removed from the Original Dgroups). The modified Dgroup representations remaining after this process were termed the *Stripped Dgroups*. Since many of the attributes removed were lexically identical across story sets, and hence facilitated the success of the straightforward feature-matching method of the content vector metric at modeling retrieval, we hypothesized that the removal of surface information would be detrimental to the retrieval performance of the content vector metric, while leaving the performance of the LSA metric unchanged (because its approach to lexical similarity was not reliant on identicality).

As well as including information not explicitly mentioned in the natural language materials, the Original Dgroups also omitted some information that was explicitly referred to in the natural language materials (e.g., Karla is explicitly described as being "old" in the natural language materials, and yet (`old Karla`) does not feature in the corresponding Original Dgroup), and relational propositions that occurred in the original stories were translated into a canonical form. A second simulation in Experiment 3 tested the two retrieval metrics on Dgroup representations that were as faithful as possible to the original. Information was only included in these *Faithful Dgroups* if it was directly warranted by the natural language materials.

4.5.1. Producing the Faithful Dgroups

Humans extract more meaning from language than the basic information encoded in the surface structure of texts and dialogues might suggest. To take the following example:

```
John hit Mary; Mary cried. The Principal sent John home.
```

In interpreting this, a reader has to infer firstly that John's hitting Mary caused her to cry, and secondly that the relationship between John's hitting Mary, and her crying, caused the Principal to send John home. We might express this information in terms of the following nested propositional structure:

```
cause(cause(hit(john, mary), cry(mary)),
send(principal, john, home))
```

None of this causal information appears explicitly in the original utterance, so it is clear that it must in some way be inferred from a prior source. (The need for inference here is uncontroversial: all theories of comprehension agree that it requires a great deal of active involvement on the part of the comprehender when it comes to inferring information that is not explicitly encoded in language—see, e.g., Gernsbacher, 1985; McKoon & Ratcliff, 1992. Where there is disagreement, it is on what, and how much, inference actually happens.)

Whilst no commitment to a particular theory of comprehension was made in specifying the procedure for translating texts into Faithful Dgroups, an attempt was made to devise a method of representation that requires a minimal amount of inference, and is broadly compatible with the bulk of the available data in this area (again, see McKoon & Ratcliff, 1992).

4.5.2. Algorithm for constructing the Faithful Dgroups

The following procedure was used to form the Faithful Dgroups, seeking to maximally preserve open-class (content word) lexical information, from natural language samples:

1. Identify the objects that are referred to in the text, and define them as entities.
2. Identify all the word structures used to express attributes of the objects in the text, and express these as unary predicate expressions.
3. Identify the word structures used to express relations between the identified objects, and express these in the Faithful Dgroups as expressions with two or more arguments, where the expressions take only objects as arguments.
4. Next represent higher-order information (i.e., the temporal and causal information that is frequently implicit in natural language representations). This information should be represented by expressions which take other expressions as arguments. Note that because this information is often implicit in the natural language forms of the stories, a standard (or canonical) lexical identity for each expression must be adopted (this has the effect of minimizing the influence of inferred structures on retrieval, which is in accordance with Gentner's empirical findings). The set of inferred relations should be the minimum set required to articulate the narrative structure of the story.

This procedure seeks to minimize unwarranted inferences and the addition of features not warranted by their inclusion in the original materials. Below is an extract from the Faithful Dgroup corresponding to the natural language version of the Karla base story (see Table 2 for comparison):

```
(sme:defDescription base-5
 ;;list the entities
 entities (karla oak-tree one-afternoon hunter ground bow
 arrows feathers hawks deer)
 ;; define expressions
 ;; object attributes first
 expressions (((old karla)                 :name old-karla)
   ((hawk karla)                           :name hawk-karla)

   ((tall oak-tree)                        :name tall-tree)
   ...

((sees karla hunter-with-gear-at-ground)
                                           :name karla-see-hunter)

   ((aims-at hunter karla bow)             :name hunter-aims)
   ((shoots-at hunter karla arrow)         :name hunter-shoots)
   ((miss hunter karla)                    :name hunter-misses)
```

```
((wants hunter feathers)          :name hunter-wants-feathers)
((knows karla hunter-wants-       :name karla-knows)
  feathers)

((glides-down karla hunter)       :name karla-glides)
((a-few feathers)                 :name a-few-feathers)
((give karla hunter a-few-        :name give-feathers)
  feathers)
...
```

Where more than one word was required to do justice to the meaning of an expression the phrase was hyphenated and used as a functor in this form (e.g., "aims-at" or "a-few"). When an LSA comparison is performed the comparison is insensitive to hyphens, and so these functor names perform just as if they had been extracted from the natural language directly.

As with the Stripped Dgroup simulation, we hypothesized that the removal of canonically represented information in the Faithful Dgroups would be detrimental to the retrieval performance of the content vector metric, while the performance of the LSA metric, which is not reliant on canonical representations, would remain unchanged.

4.5.3. Method

The MAC/FAC routines to extract the surface information from a Dgroup to complete its content vector comparisons were again used to extract the information to be used for LSA comparison. The LSA and content vector models were used to compare the base of each Stripped Dgroup set and each Faithful Dgroup set with its four variants, for each of the 9 material sets available. The LSA model was used with the same settings as in previous experiments.

4.5.4. Results

The results of the two simulations are graphed in Fig. 4.

4.5.4.1. Stripped Dgroups. ANOVA analysis of the content vector data revealed a main effect of surface similarity ($F(1, 8) = 21.167$, $p < .005$); a main effect of structural similarity ($F(1, 8) = 10.027$, $p < .05$); and a marginally significant interaction effect ($F(1, 8) = 3.717$, $p = .09$). Analysis of the LSA data revealed a main effect of surface similarity ($F(1, 8) = 23.051$, $p < .005$). There was no effect of structural similarity ($F(1, 8) = 1.048$, $p = .34$), and no significant interaction between the factors ($F(1, 8) < 1$).

4.5.4.2. Faithful Dgroups. For the content vector method there was a marginal effect of surface similarity ($F(1, 8) = 3.647$, $p = .09$), a marginally significant effect of structural similarity ($F(1, 8) = 3.383$, $p = .10$), and no interaction effect ($F(1, 8) < 1$). For the LSA method there was an effect of surface similarity ($F(1, 8) = 66.091$, $p < .001$); no significant effect of structural similarity ($F(1, 8) = 2.190$, $p = .18$); and no significant interaction between the factors ($F(1, 8) = 1.094$, $p = .33$).

4.5.5. Discussion

As expected, in the first simulation the changes to the Stripped Dgroup representations inhibited the content vector metric's ability to perform in accordance with Gentner et al.'s empirical data. Not only was the metric sensitive to surface similarity, it was sensitive, as it should not have been, to structural similarity. In contrast, the performance of the LSA metric remained unchanged from the previous experiment, showing the required sensitivity only to surface similarity. This is significant because one free parameter in the modeling process—the addition of surface information to the Original Dgroups in an uncontrolled fashion—has been removed. The Stripped Dgroups only retain surface information contained in the Original Dgroups that is directly and explicitly referred to in the natural language materials.

In the Faithful Dgroup simulations, the content vector method showed a marginal sensitivity to *both* the surface and structural features of the stories, and hence exhibited no preference for either. As Fig. 4 shows, it failed to produce empirically adequate retrieval patterns. This is because there are few lexically-identical terms that occur both in the base of a Faithful Dgroup and its corresponding variants (reflecting the linguistic idiosyncracies of the original natural language materials). The LSA method, however, performs much better: its retrievals are only (and clearly) sensitive to the surface similarity factor required to model the empirical evidence.

It is particularly noteworthy that the LSA method assigned high retrieval scores to the LS and SS categories in this experiment, when their representations shared few identical words with the corresponding base representation. This indicates that the LSA model is not simply relying on straightforward feature-matching across distinct representations to facilitate retrievals, but is capturing instead a more complex kind of relationship between the ways that individual words are used in differing linguistic contexts.

4.6. Why a one-shot approach to analogy is unlikely to succeed

The performance of the LSA measure on the various styles of representation examined so far offers concrete evidence that it can act as a good predictor of retrieval. That it can do so even when operating on a style of representation that remains faithful to the natural language form of information, and relies on only a plausible range of inferences for its structure is encouraging. These modeling experiments have shown—by successfully simulating human performance in Gentner et al.'s (1993) studies—that a co-occurrence model can meet the first three of Gentner et al.'s constraints on analogical retrieval (primacy of the mundane, surface superiority, and rare insights) whilst relying on fewer free parameters in on its representations than existing models. It seems reasonable to add to this that co-occurrence measures, which make use only of simple associative mechanisms, are at least as likely as other existing proposals (if not more so) to meet Gentner et al.'s fourth constraint, that of scalability.

Given the success of the LSA metric at modeling retrieval, an obvious question to ask is "Can LSA account for the whole process of analogical matching—retrieval and mapping—on its own?" The answer to this, quite simply, is that it *cannot*. LSA is insensitive to the structural characteristics of the Faithful Dgroups—it treats the bundle of features in a representation as an unstructured conglomerate. Therefore, it cannot account for the sort of structural matching required to generate analogical interpretations, and to generate candidate inferences (Gentner, 1983). To see this most obviously, Fig. 1 shows the different patterns of retrievability and

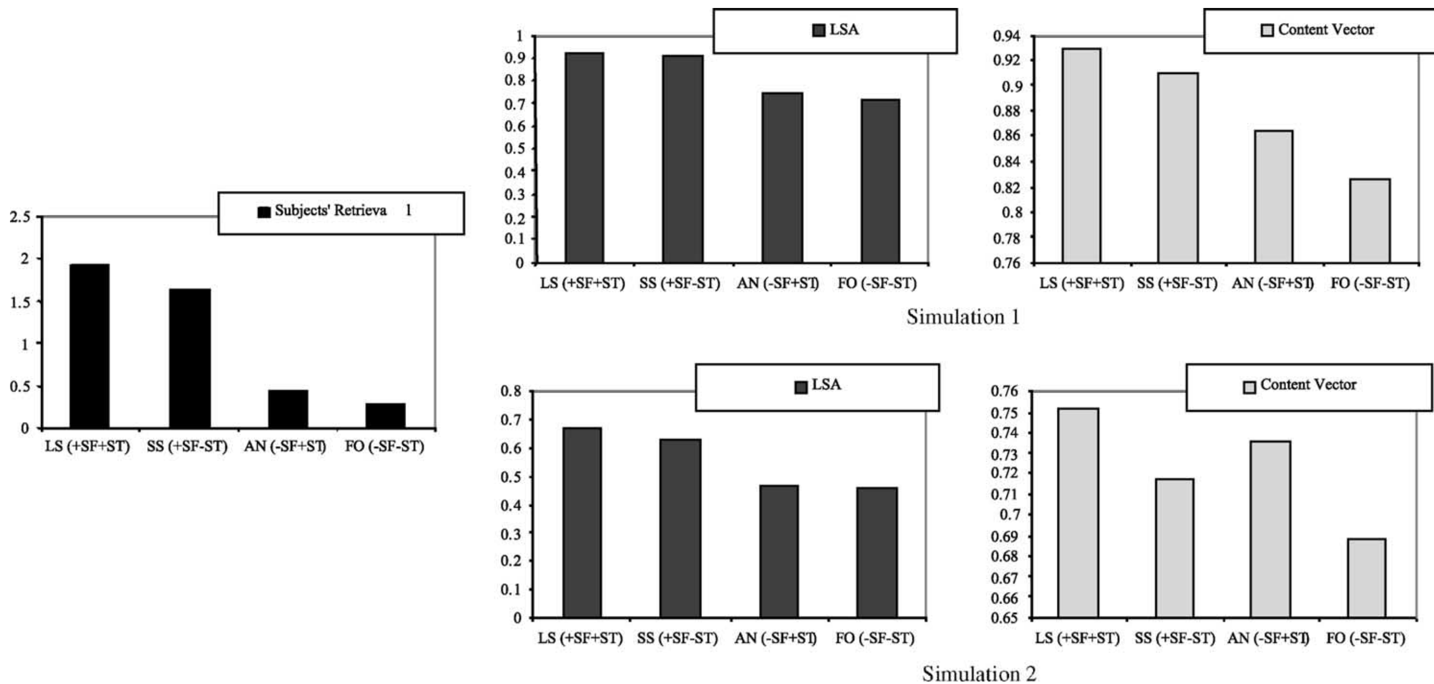

Fig. 4. Subjects' retrieval scores and content vector and LSA scores for each of the four variants as compared to the base stories. In Simulation 1, the materials were in the Stripped Dgroup form, and in Simulation 2, the materials were in the Faithful Dgroup form.

judgments of inferential soundness—based on the structural commonalities of representations—produced by subjects in the studies conducted by Gentner et al. (1993). Whilst LSA can model retrieval patterns successfully, it cannot account for the different similarity patterns that arise as a result of people's sensitivity to the structural aspects of representations (Clement & Gentner, 1991; Gentner, 1983). Indeed, given the dissociation in the human data in Gentner et al.'s study, it is hardly surprising that LSA is unable to model both retrieval and mapping; the overwhelming indication from this evidence is that two cognitive mechanisms are at work here (see also Clement, Mawby, & Giles, 1994). It appears that a different—structure sensitive—approach is required to account for matching global representations during analogical mapping.

5. Mapping

There is a consensus in modern theories of analogy that analogical mapping is largely a process of making systematic structural correspondences between two representations (Forbus et al., 1994; Gentner, 1983; Holyoak & Thagard, 1995; Hummel & Holyoak, 1997; Keane, Ledgeway, & Duff, 1994); e.g., making a globally coherent mapping between, say, the SOLAR SYSTEM and A SIMPLE MODEL OF AN ATOM that can warrant the transfer of structurally embedded elements that are present in only one representation to the other (Clement & Gentner, 1991; Gentner, 1983).

In mapping, the problem of semantic matching arises in the process of aligning individual elements in the construction of a systematic global mapping; e.g., mapping the SUN to NUCLEUS, or REVOLVES_AROUND to ORBITS. This can be illustrated by considering Gentner's (1983) structure-mapping theory, which proposes a set of constraints defining permissible mappings between a base and target domain in analogy (Gentner, 1983), and is implemented in the Structure-Mapping Engine (SME; Falkenhainer, Forbus, & Gentner, 1989). Structure-mapping theory constructs analogical mappings between discrete domains of propositional statements (the Dgroups), with the main focus being on mapping interconnected *relational* structure.

In detecting shared relational structure the structure-mapping theory permits mappings to be made between relations if, and only if, they have lexically-identical functors (the *lexical-identicality constraint*) and the same number of arguments. Thus, there are two constraints on the formation of an initial *match hypothesis*. With regards to the *lexical-identicality constraint*, it is important to observe that it carries a commitment to a canonical theory of representation because it requires that mappable relations are represented with identical names (or decompose into the same canonical forms; see earlier discussion and below). For example, SME would not permit an alignment between the following two relations, even though it might be appropriate given a wider context:

```
(ORBITS PLANET SUN)
(REVOLVES_AROUND ELECTRON NUCLEUS)
```

The fact that ORBITS is not lexically-identical to REVOLVES_AROUND also means that the corresponding analogical mappings between the arguments of the relations (PLANET with ELECTRON, and SUN with NUCLEUS) are not made. Holyoak and Thagard (1995, p. 258) have argued that this constitutes a significant weakness in structure-mapping theory: "with its emphasis on structure to the exclusion of all other constraints, SME does not simply discourage mappings between nonidentical but semantically similar items; it does not even permit them."

Both the ACME (Holyoak & Thagard, 1989) and LISA (Hummel & Holyoak, 1997) models of analogy avoid this objection by postulating semantic links that hold between the names of relations. These links are hand-coded into the propositional representations between which the analogical mappings are generated. If a sufficiently strong semantic link is coded between two relations then a mapping can occur between them. Thus, in the example above, ACME's or LISA's representations could incorporate a sufficiently strong semantic link between ORBITS and REVOLVES_AROUND to enable a mapping to be generated from one relation to the other.

5.1. Mapping and the CR theory

Holyoak and Thagard's criticism of the structure-mapping theory is not entirely fair, however, as it ignores SME's commitment to a CR theory. This extra assumption would allow the intuitively correct mapping to be made in the above case (ORBITS and REVOLVES_AROUND can be recast in identical terms). However, since the postulation of semantic links (in ACME and LISA) and the CR theory rely on human-based coding decisions—and neither subscribe to a worked out model of semantics—both are ultimately equivalent in terms of their explanatory power.

SME takes a modular approach to be taken to the cognitive modeling of analogical semantics. By mapping across canonical representations, questions of semantics are left outwith the scope of structure-mapping theory. SME (and by inheritance MAC/FAC) thereby remain noncommittal with respect to a theory of lexical semantics. The following simulation experiment exploited SME's modular approach to semantics in order to examine whether, once appropriately structured representation have been retrieved, semantic constraints between sub-elements in representations need be enforced at all in mapping: If SME could map successfully between noncanonical representations, underspecification inherent in the CR theory could be avoided.

5.2. Simulation Experiment 4

Simulation Experiment 4 was conducted to test the performance of SME both on the Original Dgroups, which are the original encodings of 9 of the 20 Karla the Hawk story sets, in order to provide a baseline measure of SME's performance, and then on the Faithful Dgroups as described earlier. The Faithful Dgroups feature the same words used in the natural language versions of the Karla stories: this means that SME will not be able to rely on the lexical-identicality of predicates in order to constrain the mappings it produces.

Since SME enforces a *lexical-identicality* constraint, it was expected that SME would fail to map appropriate structures in the Faithful Dgroups because lexical differences between otherwise semantically similar tokens in the Faithful Dgroup representations would preclude mappings under this constraint. Accordingly, simulations were also run in which SME was modified to eliminate the lexical-identicality requirement (and thus SME's commitment to CR theory), to see whether the SME would still be sensitive to global structure in the representations. (Thus, this simulation also examined the question of what role a matching criterion plays in SME: is it critical to the SME algorithm, or is it merely an efficiency measure?)[5]

5.2.1. Materials

The materials used were the 9 sets of the "Karla the Hawk" materials (described earlier) that had been coded using a canonical representation scheme (Forbus et al., 1994), and the 9 Faithful Dgroups that faithfully encode the lexical properties of the original Karla representations.

SME was then modified to eliminate the *lexical-identicality* constraint (M-SME), and used to perform the inter-set mapping task. The SES scores and number of match hypotheses formed were recorded for each mapping produced on the Faithful Dgroups between each base and its corresponding four variants.

5.2.2. Method

For each of the nine sets of Original Dgroups, SME was used to map the base Dgroup onto its four variants. The structural evaluation score (SES)[6] and number of match hypotheses formed for each mapping were then recorded.

5.2.3. Results

The results of the SME simulations are graphed in Fig. 4, and the results of the M-SME simulation are graphed in Fig. 5.

5.2.3.1. Original Dgroups (SME simulation, lexical-identicality constraint enforced). For the *SES scores*, a repeated measures ANOVA treating items as a random-effect revealed that the only significant effect was for structural similarity ($F(1, 8) = 5.43$, $p < .05$). Both the surface similarity ($F(1, 8) < 1$) and interaction ($F(1, 8) = 1.24$, $p = .30$) factors produced nonsignificant effects. With regards to *match hypothesis formation*, the only significant effect was for the surface similarity ($F(1, 8) = 51.44$, $p < .001$). Both the structural similarity ($F(1, 8) = 1.12$, $p = .32$) and interaction ($F(1, 8) = 1.29$, $p = .29$) factors produced nonsignificant effects.

5.2.3.2. Faithful Dgroups (SME simulation, lexical-identicality constraint enforced). Here, the ANOVA analysis showed that on the *SES scores* surface similarity had a nonsignificant effect ($F(1, 8) < 1$); structural similarity had a marginal effect ($F(1, 8) = 4.72$, $p = .06$); and there was no interaction effect ($F(1, 8) < 1$). Again, an analysis of *match hypothesis formation* showed that there was no effect of structural similarity ($F(1, 8) < 1$); a very marginal effect of surface similarity ($F(1, 8) = 3.21$, $p = .11$); and no interaction effect ($F(1, 8) < 1$). Testing on both the SES scores ($t = 11.37$, $df = 35$, $p < .001$) and the number of match hypotheses ($t = 8.38$, $df = 35$, $p < .001$) revealed that there was a significant decrease in the means of both from mapping on the Original Dgroups to mapping on the Faithful Dgroups.

5.2.3.3. Faithful Dgroups (M-SME simulation, lexical-identicality constraint disabled): SES scores. The only factor that produced a significant effect was structural similarity ($F(1, 8) = 19.00$, $p < .005$). Both surface similarity ($F(1, 8) < 1$) and interaction ($F(1, 8) < 1$) effects were nonsignificant. *Match hypothesis formation:* all three factors produced nonsignificant effects: structural similarity ($F(1, 8) = 2.09$, $p = .19$); surface similarity ($F(1, 8) < 1$); and interaction effects ($F(1, 8) = 1.40$, $p = .27$) (Fig. 6).

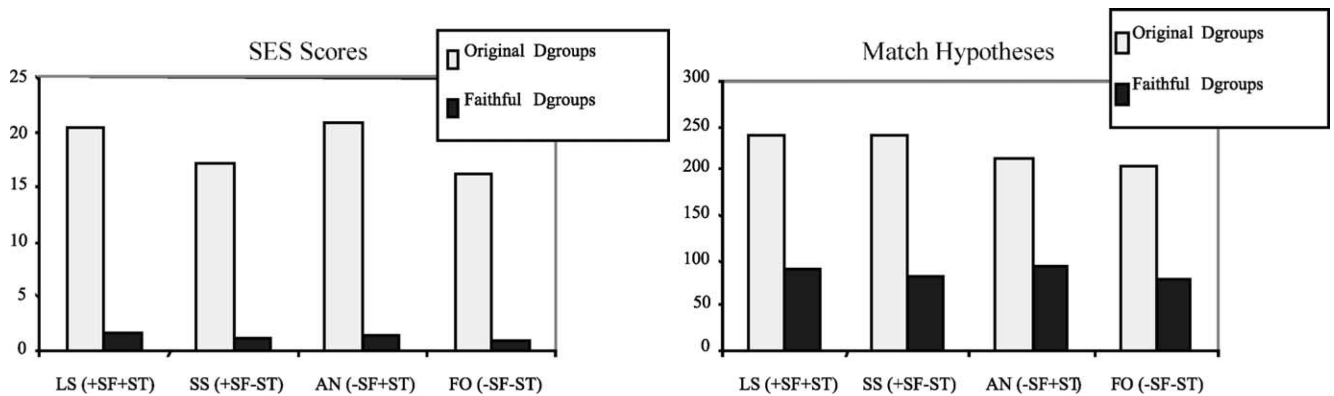

Fig. 5. The mean SES scores and mean numbers of match hypotheses formed with the original SME module on the Original and Faithful Dgroups.

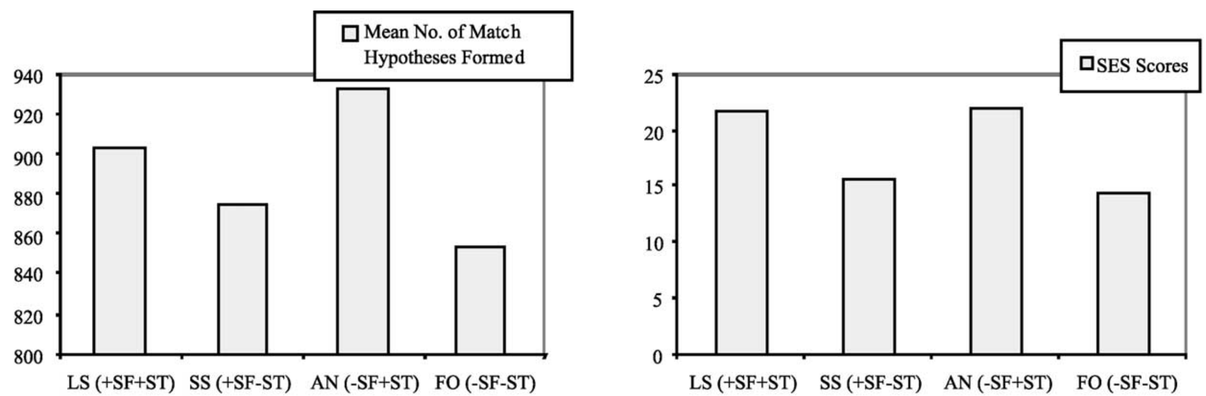

Fig. 6. The SES scores and number of match hypotheses formed with the LSA retrieval module and M-SME performing intra-set mappings on the Faithful Dgroups.

5.2.4. Discussion

As expected, in the first baseline simulation mapping the Original Dgroups, SME exhibited the required sensitivity to the structural commonalities of the Original Dgroups (witness the higher SES scores for the LS and AN mapping conditions). This is demonstrated by the fact that the only significant factor in the analysis of the SES scores was structural similarity. Interestingly, the number of match hypotheses formed for each category of match was sensitive to the surface similarity factor. This reflects both that lexically-identical functors are more likely to occur in the Original Dgroups when there are shared surface features, and that SME can only form match hypotheses between relations with lexically-identical functors.

On the Faithful Dgroups, which had not been hand-coded to meet lexical-identicality constraints, SME still exhibited marginal sensitivity to structural similarity (see Fig. 5). However, the greatly reduced magnitude of SES scores from its performance on the Original Dgroups show that it fails to map comparable quantities of structure on the Faithful Dgroups when compared to the Original Dgroups. This indicates that structure which should be analogically aligned may be being passed over by the SME algorithm when operating on the Faithful Dgroups. Furthermore, the greatly reduced number of match hypotheses formed for each category of mapping (reduced from an overall mean of 224.8 in the Original Dgroup simulation to 87.1 in the Faithful Dgroup simulation) suggests a possible explanation of this failure: the constraints on the formation of match hypotheses are too strict to allow the appropriate local alignments to be made on the Faithful Dgroups (because there are an insufficient number of lexically-identical relations between different domains. It is noteworthy, however, that even on the Faithful Dgroup representations, the overall assessments made by SME of which of the categories of stories provide the best category of matches are still in line with Gentner et al.'s 1993, findings; see Fig. 5). The pattern—if not perhaps the quality—of SME's mappings appears to be correct.

In the M-SME simulation where the lexical-identicality constraint in SME was disabled, the number of match hypotheses proposed by M-SME was insensitive to surface similarity (in contrast to the results for the Faithful Dgroups in the SME simulation). The pattern of results for the SES scores show that M-SME exhibits the appropriate sensitivity to structural similarity on the Faithful Dgroups, indicating that it can successfully generate structural mappings on representations that encode the lexical properties of the materials they are derived from.

This is interesting because it indicates that appropriate mappings can still be produced on representations which have not been hand-coded to reflect items that could be appropriately aligned (previously, because of the lexical-identicality constraint embodied in SME, elements in the Original Dgroups that were supposed to be aligned had to be coded with identical predicates meaning that the mapping process was constrained, to some extent, by human decision).

One way of interpreting this result would be to make the claim that M-SME can make appropriate analogical mappings in the absence of any semantic information. There are a number of problems with this, most notably that analogical mapping doesn't just *match* two representations together: analogical mappings form the basis of an interpretation of one representation in terms of another. Relaxing the requirement that relations be identical in a mapping renders unclear how it is that the similarity between the representations is to be interpreted. Although a match between "revolves around" and "orbits" might make sense at a surface level (it may be that in an analogy between the solar system and an atom, the surface similarity between these functors allows them to be treated as identical in the *psychological* processing of analogy), in a matching process unconstrained by semantic factors, not all mappings will be so obviously semantically similar. How can the interpretation of such matches be resolved?

One answer might be to say that such interpretations are outside the scope of the model—that the model can identify which information is mapped, and which inferences should be carried over, is a sufficient demonstration of its ability to model some of the key processing underlying interpretation.

Another answer leads to a very different interpretation of what these results imply. In Experiment 4, *both* SME and M-SME succeed in ordering the candidate bases appropriately in terms of their underlying structural commonalities with the target analog (i.e., in line with the results of Gentner et al., 1993) in the absence of any "semantic" information. That this is possible may have less to do with whether or not semantic factors actually play a part in mapping than with the fact that once representations that have appropriate structural commonalities have been retrieved, a structure mapping algorithm will, by definition, be able to determine which structures match best.

This consideration leads to a problem with all current models of analogy (including the one developed here): the representations used in the simulations reported here are static (they encapsulate knowledge in a "canned" format). However, the extent to which the knowledge used by people in analogical thinking is stored in neatly packaged structures is questionable. Although evidence suggests that analogical reasoning is incremental (Keane, 1996, 1997), with base representations being built up dynamically as reasoning proceeds, the processes underlying representation building are poorly understood. How are these representations built up? What are the further interactions between retrieval and mapping in building representations (perhaps through re-representations), and what is the relationship between what may well be qualitatively different semantic metrics (taking in structural and surface similarities between stored representations) in representation building? On this view, one of the more interesting aspects of the results of Experiment 4 is that they illuminate the limitations of studying analogy by focusing on fixed representations.

6. General discussion

This paper presented a series of simulation experiments modeling observed patterns of behavior in analogical retrieval and mapping. The environmental retrieval model (LSA) and M-SME components described here together comprise what we term an environmental model of analogy (EMMA). This model offers significant advantages over previous models such as MAC/FAC and LISA (though it should be noted that EMMA is very much a variation on a theme of the former model). EMMA achieves similar or better levels of performance in modeling especially retrieval data, whilst simultaneously exploiting fewer free parameters in achieving these modeling results. Significantly, and importantly, one of the elements that has been largely removed from the model is the requirement for externally represented, fully specified concepts.

This is particularly important because of the increasing amount of evidence that the mechanisms underlying analogy and similarity-based transfer play a crucial role in *all* conceptual thought, and critically in categorization (Ahn, 1998; Carey, 1985; Gentner & Wolff, 1997; Hummel & Holyoak, 1997; Lakoff & Johnson, 1980; Markman, in press; Medin, Goldstone, & Gentner, 1993; Medin, Goldstone, & Markman, 1995; Yamauchi & Markman, 2000). From a knowledge-representation and processing point of view, explaining both ordinary conceptual processes and analogy are very similar tasks. Just like analogy, the act of categorization involves comparing two representations. To decide whether or not some new stimulus is a penguin, for example, we need to compare our representation of penguins with that of the new stimulus.

The idea of explaining ordinary conceptual processes in terms of analogy is appealing. Explaining analogical thinking forces one to deal explicitly with the kind of structured representations that appear to underlie much of our conceptual knowledge. One massive drawback to this approach, however, is that if theories and models of analogy in turn rely on appeals to concepts to explain many of their workings, we are left with a circle to be squared: explanations of conceptual thought that use analogical mechanisms will be ultimately circular.

The approach put forward in EMMA offers at least a partial solution to this problem. In EMMA, analogizing (or any other meaningful comparison process) is seen as an operation that takes place between representations in working memory. However, canonical concepts are not required to explain the retrieval and mapping of these representations. In EMMA (as far as lexical representations go) at least one of these ancillary elements of analogical thought—retrieval—can be empirically grounded in simpler background knowledge extracted from the linguistic environment. Some support for this approach comes from a series of studies by Boroditsky and Ramscar (in preparation) which show both that people are sensitive to this kind of environmental (statistical) knowledge (words in an artificial language were perceived to be more similar in meaning if they co-occurred with similar contexts more frequently), and that this information can act to facilitate structural alignment as proposed in EMMA (when the names of novel objects were co-occurred together in context in the artificial language, the objects by themselves were later seen to be more similar than the same objects when rated without having had their names contextually co-occurred).

A logical extension of the view put forward here is that all conceptual thought might be seen as the processing of structured, context specific representations in working memory, with the processes operating on these representations making use of more general environmental knowledge (this, perhaps initially extreme sounding, proposal gains further support from such well documented phenomena as change blindness (see Simons, 2000 for an overview) and functional fixedness (Duncker, 1945; Glucksberg & Weisberg, 1966) which graphically indicate that the representations people actually use in thinking may be somewhat less than the sum of their total world knowledge).

Although EMMA is successful at simulating a simplified version of the kind of environmentally-grounded thinking we have in mind here, it is important to restate the fact that the representations used in the simulations reported here are static. The process by which working representations are built up in context, and the extent to which re-representation is constrained by memory capacity limits, has been little explored and remains largely unknown (but see Halford, Wilson, & Phillips, 1998). Like other analogy models, EMMA offers no insights on how psychologically plausible representations are actually built. What EMMA perhaps does offer, however, is a view of how far the modeling of analogical processes operating on static representations can go. These simulations have shown that environmental knowledge can provide a model for semantics in retrieving representations. Explaining how different kinds of "semantic similarities" interact dynamically in retrieval, mapping and representation-building will provide a stiffer challenge; although it may be that meeting it will allow us to account for much of high-level cognition.

Notes

1. Although we use the LSA model in the experiments reported in this paper, we see LSA as an exemplar of the more general co-occurrence approach.
2. Closed-class words are essentially function words: they are the words that are used in providing grammatical structure for 'open-class' content words. Since they tend to remain constant as languages change, the set of these function words is said to be closed under the grammatical rules of the language.
3. The LSA model is accessible on-line at http://lsa.colorado.edu/.
4. This is to be expected. Because the LSA model uses an entropy-weighting term, words which occur in a wide variety of contexts, such as closed-class terms, will be counted as less important in the comparison process.
5. See Yarlett and Ramscar (2000) for a discussion of the possibility of using LSA to provide a matching threshold in this context.
6. SES scores are automatically calculated by SME and provide a measure of the quantity of structure that has been mapped between two domains.

Acknowledgments

We are grateful to Ken Forbus for generously making the MAC/FAC code available to us, and to Andrew Wishart for his help in coding the Faithful Dgroups. We also thank Lera Boroditsky, Rob Goldstone, Graeme Halford, Walter Kintsch and Art Markman for many thoughtful comments on earlier drafts of this paper, and Ken Forbus, Dedre Gentner, Ken Kurtz and Alex Heneveld for insightful discussions of these issues. Portions of this paper have been presented at the 22nd annual meeting of the Cognitive Science Society at The University of Pennsylvania. The work reported was supported in part by EPSRC Grant GR/M59846 to Michael Ramscar.

References

Ahn, W. K. (1998). Why are different features central for natural kinds and artifacts? The role of causal status in determining feature centrality. *Cognition*, *69*, 135–178.

Boroditsky, L., & Ramscar, M. J. A. (in preparation). *Co-occurrence and comparison*.

Bransford, J. D., & Johnson, M. K. (1972). Contextual prerequisites for understanding: Some investigations of comprehension and recall. *Journal of Verbal Learning and Verbal Bahavior*, *11*, 717–726.

Burgess, C., & Lund, K. (1997). Modeling parsing constraints with high-dimensional context space. *Language and Cognitive Processes*, *12*, 177–210.

Carey, S. (1985). *Conceptual change in children*. Cambridge, MA: MIT Press.

Clark, H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior*, *12*, 335–359.

Clement, C. A., & Gentner, D. (1991). Systematicity as a selection constraint in analogical mapping. *Cognitive Science*, *15*, 89–132.

Clement, C. A., Mawby, R., & Giles, D. E. (1994). The effects of manifest relational similarity on analog retrieval. *Journal of Memory and Language*, *33*, 396–420.

Duncker, K. (1945). On problem-solving (L. E. Lees, trans.). *Psychological monographs, 58* (Chapter VII) (5, Whole No. 270).

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

Fodor, J. (1975). *The language of thought*. New York: Crowell.

Fodor, J. (1981). The present status of the innateness controversy. *RePresentations.* Cambridge: MIT Press.

Forbus, K., Gentner, D., & Law, K. (1994). MAC/FAC: A model of similarity-based retrieval. *Cognitive Science*, *19*, 141–205.

Forbus, K. D., Gentner, D., Markman, A. B., & Ferguson, R. W. (1998). Analogy just looks like high level perception. Why a domain-general approach to analogical mapping is right. *Journal of Experimental and Theoretical Artificial Intelligence*, *10*, 231–257.

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

Gentner, D., Rattermann, M., & Forbus, K. (1993). The roles of similarity in transfer: Separating retrievability from inferential soundness. *Cognitive Psychology*, *25*, 524–575.

Gentner, D., & Wolff, P. (1997). Alignment in the processing of metaphor. *Journal of Memory and Language*, *37*, 331–355.

Gernsbacher, M. A. (1985). Surface information loss in comprehension. *Cognitive Psychology*, *17*, 324–363.

Gick, M. L., & Holyoak, K. J. (1980). Analogical problem solving. *Cognitive Psychology*, *12*, 306–355.

Glucksberg, S., & Weisberg, R. W. (1966). Verbal behavior and problem solving: Some effects of labeling in a functional fixedness problem. *Journal of Experimental Psychology*, *71*, 659–664.

Goswami, U. (1992). *Analogical reasoning in children*. Mahwah, NJ: LEA.

Halford, G. S., Wilson, W. H., & Phillips, S. (1998). Processing capacity defined by relational complexity: Implications for comparative, developmental and cognitive psychology. *Behavioral and Brain Sciences*, *21*, 803–864.

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

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

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

Keane, M. T. (1996). On adaptation in analogy: Tests of pragmatic-importance and adaptability in analogical problem solving. *Quarterly Journal of Experimental Psychology*, *49A*, 1062–1085.

Keane, M. T. (1997). What makes an analogy difficult?: The effects of order and causal structure on analogical mapping. *Journal of Experimental Psychology: Learning, Memory* & *Cognition*, *23*, 946–967.

Keane, M. T., Ledgeway, T., & Duff, S. (1994). Constraints on analogical mapping: A comparison of three models. *Cognitive Science*, *18*, 387–438.

Komatsu, L. K. (1992). Recent views of conceptual structure. *Psychological Bulletin*, *112*, 500–526.

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

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The Latent Semantic Analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*, 211–240.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to Latent Semantic Analysis. *Discourse Processes*, *25*, 259–284.

Laurence, S., & Margiolis, E. (1999). In S. Laurence & E. Margiolis (Eds.), *Concepts and cognitive science*. Cambridge, MA: MIT Press.

Lund, K., Burgess, C., & Atchley, R. A. (1995). Semantic and associative priming in high-dimensional semantic space. In *Proceedings of the seventeenth annual conference of the cognitive science society* (pp. 660–665). Pittsburgh, PA: Erlbaum.

Markman, A. B. (in press). Structural alignment, similarity, and the internal structure of category representations. In U. Hahn & M. J. A. Ramscar (Eds.), *Similarity and categorization*. Oxford: Oxford University Press.

McKoon, G., & Ratcliff, R. (1992). Inference during reading. *Psychological Review*, *99*, 440–466.

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

Medin, D. L., Goldstone, R. L., & Markman, A. B. (1995). Comparison and choice: Relations between similarity processing and decision processing. *Psychonomic Bulletin and Review*, *2*(1), 1–19.

Ramscar, M. J. A., & Hahn, U. (1998, August). What family resemblances are not: The continuing relevance of Wittgenstein to the study of concepts and categories. In *Proceedings of the twentieth annual conference of the cognitive science society* (pp. 865–870). Madison-Wisconsin, USA: Lawrence Erlbaum.

Ramscar, M. J. A., & Pain, H. G. (1996). Can a real distinction be made between cognitive theories of analogy and categorisation? In *Proceedings of the eighteenth annual conference of the cognitive science society* (pp. 346–351). New Jersey: Lawrence Erlbaum.

Redington, M., Chater, N., & Finch, S. (1998). Distributional information: A powerful cue for acquiring syntactic categories. *Cognitive Science*, *22*(4), 425–469.

Sachs, J. (1967). Recognition memory for syntactic and semantic aspects of connected discourse. *Perception & Psychophysics*, *2*, 437–444.

Simons, D. J. (Ed.). (2000). Change blindness and visual memory: A special issue. *Visual cognition*. Brighton: Psychology Press.

Yamauchi, T., & Markman, A. B. (2000). Learning categories composed of varying instances: The effect of classification, inference, and structural alignment. *Memory and Cognition*, *28*(1), 64–78.

Yarlett, D. G., & Ramscar, M. J. A. (2000). Structure-mapping theory and lexico-semantic information. In *Proceedings of the 22nd annual conference of the cognitive science society* (pp. 571–576). University of Pennsylvania.