Copyright
by
Marc Thomas Tomlinson
2010

The Dissertation Committee for Marc Thomas Tomlinson
certifies that this is the approved version of the following dissertation:

Building BRIDGES: Combining analogy and category learning to learn relation-based categories

Committee:

Bradly C. Love, Supervisor

Catharine H. Echols

Jeffrey Loewenstein

Arthur B. Markman

Bruce W. Porter

Building BRIDGES: Combining analogy and category learning to learn relation-based categories

by

Marc Thomas Tomlinson, B.S.,M.S.

DISSERTATION

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT AUSTIN

May 2010

Dedicated to my wife Erica, without her support and patience this would not be possible.

Acknowledgments

I wish to thank, first and foremost, my advisor for suggesting the idea and for his patience with my feeble writing skills. Also, Erica Tomlinson, and Aaron Hoffman for reading multiple copies of this material. As well as the rest of the Love lab that has stuck it out (Tyler Davis) and helped to make these experiments and this paper possible.

Building BRIDGES: Combining analogy and category learning to learn relation-based categories

Publication No. _____

Marc Thomas Tomlinson, Ph.D.
The University of Texas at Austin, 2010

Supervisor: Bradly C. Love

The field of category learning is replete with theories that detail how similarity and comparison based processes are used to learn categories, but these theories are limited to cases in which item and category representations consist of feature vectors.This precludes these methods from learning relational categories, where membership is determined by the structured relations binding the features of a stimulus together. Fortuitously, researchers within the analogy literature have developed theories of comparison that account for this structure. This thesis bridges the two approaches, describing a theory of category learning that utilizes the representational frameworks provided by the analogy literature to learn categories that may only be described through the appreciation of the structured relations within their members.

This theory is formalized in a model, Building Relations through Instance Driven Gradient Error Shifting (BRIDGES), that shows how relational

categories can be learned through attention-driven analogies between concrete exemplars. This approach is demonstrated through several simulations that compare similarity-based learning and alternatives, such as rule-based abstractions and re-representation. We then present a series of experiments that explore the reciprocal impact of relational comparison on category structure and category structure on relational comparison. This work provides a theoretical framework and formal model suggesting that feature-based and relation-based categories are a continuum that are learned through selective attention and similarity-based comparison.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Category learning researchers investigate how learners come to associate groups of stimuli with a common label, and generalize these labels to novel examples. Understanding how learners represent these stimuli is critical to understanding how categories are learned and used. Most approaches assume a stimulus is represented by features which act as continuous dimensions in space (Rips, Shoben, & Smith, 1973), discrete sets (Tversky, 1977), or combinations of the two approaches (Goldstone, 1994b). For example, a bird could be represented as the collection of features: has wings, flies, has feathers, etc. These representations are adequate for explaining many examples of categorization behavior, but they all suffer from a lack of accounting for the structural information present within exemplars (e.g. has wings longer than its body, see Markman, 1999).

In contrast to these feature-based views of categorization, relational approaches suggest that accounting for this structure is a necessary component of the categorization process (Genter & Kurtz, 2005; Marcus, Vijayan, Bandi Rao, & Vishton, 1999; Markman & Stilwell, 2001). These theories introduce a separation between feature-based categories, which can be captured

by featural descriptions, and relation-based categories, where categorization requires an appreciation of the structured relationships that exist within exemplars of the category. For example, instances of *giving* have few, if any, features in common with one another, neither what gave (e.g. weather, people, rocks), what is given (e.g. colds, books, headaches), nor what received. Instead examples of *giving* contain a particular set of relationships between entities (e.g. transfer of possession).

These relation-based categories have been demonstrated to occur across phylogeny and ontology. Fully developed relational reasoning is reserved to adult humans (see Gentner & Ratterman, 1991). However, infants and many animals exhibit an understanding of simple relational categories such as relational match-to-sample, *same-different* learning, and simple grammars (Gerken, 2006; Hauser & Weiss, 2002; Marcus et al., 1999; Thompson, Oden, & Boysen, 1997; Young & Wasserman, 1997; Vonk, 2003), while older children are able to succeed with even more complex tasks requiring appreciation of the relations between relations (Halford, 1984, 1993; Kotovsky & Gentner, 1996; Richland, Morrison, & Holyoak, 2006). The difference between infants and children is attributed to a phenomenon named the relational shift (Gentner & Ratterman, 1991). A complete theory of relational learning needs to account account for this diversity.

Discussion of how relation-based categories are learned is generally limited to the analogy literature. This general approach sees these categories as being formed by making analogies between exemplars and either pruning irrel-

evant structures (Kuehne, Gentner, & Forbus, 2000) or extracting high-order relational commonalities from the examples (Doumas, Hummel, & Sandhofer, 2008; Penn, Holyoak, & Povinelli, 2008a). Integral to this approach is also the idea that novel relational structures can be acquired through re-representation. Re-representation is the idea that as new knowledge is acquired, it is integrated into existing knowledge, thus changing its structure by altering links or creating new relations. As one acquires new relation-based categories, these labels can be added to the representation of existing stimuli, this is seen as a motive force in the development of relational competence (Kotovsky & Gentner, 1996) and the relational shift (Doumas et al., 2008).

An alternative theory of relational category learning is a rule-based approach. This approach has been used to explain how infants can learn to differentially respond to member and non-member examples of simple grammars. For instance, when infants are repeatedly exposed to *words* where the phonemes follow an abstract AAB pattern (e.g. *didila* or *leleta* vs. *diladi*), they will respond differently when they are then exposed to a *word* from a different grammar, ABA (e.g. *tidedi*). The rule-based explanation suggests that learners have use of a system that can extract algebraic rules (such as, first = second $\neq$ third, in the case of the AAB pattern) and classify stimuli according to whether or not the example satisfies the abstract rule (Hauser & Weiss, 2002; Marcus et al., 1999). This approach results in learners abstracting away the irrelevant structure, with the same end result as the analogical approach. Learners end up possessing a compact abstract template that they

can use to determine category membership.

While both the analogical approach of extracting commonalities existing across multiple exemplars and the rule-based approach make inroads into understanding how these categories are formed, they have a limited ability to explain results fundamental to the category learning literature, such as typicality effects, correlations between the typicality of an exemplar (how similar it is to other members) and how quickly it is identified as belonging to the category (such as robins being identified as birds more readily than penguins, Rips et al., 1973). [1] Most of these approaches are also unable to account for the full diversity of possible category structures, such as the exclusive-or category, where members exhibit one feature or the other, but not both. In addition, categories such as *game* whose membership cannot be well described by a rule or single prototype prove problematic for these approaches.

An alternative account of relation-based categorization can be developed through the category learning literature and the family-resemblance approach of Rosch and Mervis (1975). Family resemblance stipulates that novel items are assigned to the category whose members most resemble the item. For example, most birds share many of the same features as robins (e.g. same size, have wings, fly) and exceptions to this resemblance, such as penguins, take longer for people to categorize. While this theory has historically been focused on implementations using only feature-based resemblance (Kruschke,

[1] It should be noted that DORA (Doumas et al., 2008) can account for this affect through the use of memory, but it is not inherent in the category representation.

1992; Love, Medin, & Gureckis, 2004; Medin & Schaffer, 1978; Nosofsky, 1986), family resemblance does not preclude the use of relational structure in determining resemblance. In fact, as shown below, it is almost trivial to add this accounting.

Importantly this addition resists theories that relation-based and feature-based categories are dichotomies (Markman & Stilwell, 2001), but instead suggests that they are inseparable dimensions, in that features play a role in determining membership of abstract relation-based categories, and in-turn, relations can play a role in concrete feature-based categories (Jones & Love, 2007). For example, most *chairs* exhibit a feature-based family-resemblance structure of having four legs, a back, and a flat surface to sit upon, but even a very non-chair-like object such as a bumpy rock would be classified as a chair when it is playing the instrument role in the *sit* relation, because of its shared relational structure with other chairs.

By resting our account of relation-based learning within the categorization literature we directly extend established theories of learning effects, such as category and exemplar base-rates (Medin & Edelson, 1988), typicality (Nosofsky, 1988), learned attention (Kruschke, 1992), and category structure (Shepard, Hovland, & Jenkins, 1961), to relation-based categories. Within this thesis we will only look in detail at how a small set of these learning effects interact with, and extend to, relation-based categories, but other category-learning work, such as feature primacy (Sloman, Love, & Ahn, 1998), causal reasoning (Ahn, 1999), or category coherency (Rehder & Ross, 2001) and the

effects of prior knowledge or contexts, should play as large a role in categories relying on relational information as they do in those that rely on only featural information.

This thesis develops a theory bridging the category learning and analogy literatures in order to formalize a model, Building Relations through Instance Driven Gradient Error Shifting (BRIDGES), that leverages the internal structure of exemplars to successfully classify exemplars as members of a novel relational category. Below, we first review the three required parts of a theory of classification learning (category representation, stimulus representation, and learning) and compare feature-based and relation-based views in these areas. We then present a formal description of a hypothesis linking structured representations with learning theory and provide simulations showing how this can account for numerous examples of category learning (exhibited by animals, children, and adults) that suggest appreciation of the relational structure within the stimuli. This is followed by novel experiments showing the necessity of accounting for relational structure in classification experiments. Lastly, we provide an experiment extending the idea of a reciprocal interaction between category structure and stimulus structure, which tests a basic assumption of the approach.

1.1 Category representation

The oldest view of category representation describes category membership as being by a set of rules, such as mammals have fur and give birth to

live young. This approach to category learning has existed since Aristotle and fits people's intuitive notion of how they think they represent categories (E. E. Smith & Sloman, 1994). However, even as evidenced by the mammal example, not all mammals fit this simple rule (e.g. platypuses); this representation has short-comings in that many categories do not allow descriptions by simple sets of rules (Rosch & Mervis, 1975; Wittgenstein, 1953). Some of these shortcomings can be addressed by storing exceptions which augment the rules (e.g. platypuses, Nosofsky et al., 1994). Nevertheless, this approach does not significantly change the fact that rules use a strict accounting of membership which is unable to account for the complete role of family resemblance in category membership.

Many experiments show discrepancies in how people categorize objects and strict rule-based representations. Rips et al. (1973) showed that people identify some birds and mammals as being birds or mammals faster than other birds and mammals, and that the response time correlates with the rated distance between the exemplars and the prototype. This experiment establishes the idea of a typicality effect, or that some members of a category are more typical than others. Several experiments also show that even when people are explicitly given rules to use to categorize stimuli they are still affected by irrelevant information within the stimuli (Allen & Brooks, 1991; Sakamoto & Love, 2004).

In contrast to the rule-based approach, family-resemblance theories of category learning stipulate that categories are made up of similar items. There

are three main approaches to representing the similarity structure with a category. One straightforward approach is prototype theory (Posner & Keele, 1968; Reed, 1978; J. D. Smith & Minda, 1998), where each category is represented by a single abstract prototype, which in turn represents the average stimulus of the category (that may, or may not exist). A novel item is classified into the category that contains the most similar prototype. This representation easily accounts for typicality effects by being able to determine how well any particular exemplar matches the prototype.

One strength of prototype approaches is their ability to represent a complex category as a single entity; however, this also restricts their applicability. Often categories can contain separate and distinct sub-categories – in one group all the items could be *small* and *dark*, whereas in the other they could be *large* and *bright*. Prototype theory would incorrectly create one prototype, *medium* and *medium*, which would not capture the more nuanced structure present in the category. Even infants demonstrate this behavior, forming sub-groups around exemplars containing correlated attributes (Younger, 1985). In this way, prototype approaches are similar to rules in that they reduce a complex category to a simple abstraction which often misses some of the finer points of the category.

In contrast to prototype theory, which represents each category with a single prototype, is exemplar theory (Medin & Schaffer, 1978), which represents each category as a collection of all of the known examples of the category. A novel item is compared to every exemplar and scored according to how well

it matches every exemplar in each category. The item is classified into the category which contains the highest number of most similar exemplars. This approach still captures typicality effects (as more typical members are closer to more other exemplars than atypical members), although it does so in a way that is mathematically distinct from prototype approaches (see Nosofsky & Zaki, 2002; J. Smith, 2002). In addition, this approach can correctly account for subgroups because it maintains the idiosyncrasies of the exemplars in its representation.

Using the above example of the *small* and *dark*, and *large* and *bright* sub-groups, an exemplar model would compare a novel stimulus to all of the exemplars in both sub-groups. Because exemplar theory classifies a stimulus according to its similarity to other members (non-similar exemplars have little impact on classification), a stimulus falling into one of the two sub-groups will be rated as most typical. A mathematically average stimulus will not be similar to any of the exemplars, so would not be rated as being typical. In contrast, a prototype model would consider the average stimulus as the prototypical example.

The above category describes a non-linearly separable category, as a single line cannot be drawn that separates members from non-members. Prototype theory is limited to learning these types of categories. In contrast, people are quite able to learn non-linearly-separable categories, with some structures being as fast, or faster, than their linearly-separable equivalents (Medin & Schwanenflugel, 1981) as predicted by exemplar models.

However, this strength of exemplar models can also be a weakness. Exemplar models are theoretically able to represent almost any category distinction where their similarity function is a veridical representation of the similarity between the exemplars. Ashby and Alfonso-Reese (1995) showed that exemplar models (under assumptions present in most implementations) are equivalent to non-parametric kernel-based estimators, assuming a multivariate normal kernel function. In contrast, the researchers showed that prototype-models are equivalent to parametric classifiers, assuming that the category members are distributed according to a multivariate normal. This alternative representation of these two models clarifies that they hold different beliefs about the complexity of categories. Exemplar models assume that categories are inherently complex; this can lead to generalization errors if the true category structure is simple.

The last approach to category learning addresses this concern by following a mixed approach where categories are represented by multiple prototypes, or clusters (such as one for bats and one for the rest of the mammals, Anderson, 1991; Love et al., 2004). This approach allows the representation to start simple and only add complexity when it is necessary. Clustering models make the assumption that categories are generally well-behaved, only storing exemplars when faced with an unexpected error (or an especially unique exemplar) occurs while classifying a stimulus.

These models make an explicit prediction about how exemplars are remembered. They predict that unusual exemplars should be remembered better

than those that would get folded into a cluster. Those unusual exemplars can be those that differentiate themselves from other category members by being dissimilar to other exemplars in the same category (Sakamoto & Love, 2006). They can also be category exceptions, in that they violate simple rules for categorization (Nosofsky et al., 1994; Palmeri & Nosofsky, 1995) or collections of imperfect rules formed during learning (Sakamoto & Love, 2004).

For the purposes of demonstrating that relation-based categories can be captured without recourse to rules or abstract prototype formation, this thesis is framed in terms of an exemplar model. As will be discussed later, we postulate that abstraction forms through learned attention and not through the formation of a prototype. Exemplar models allow us to clearly demonstrate how abstraction forms, since they can not form a prototype or learn rules in other ways. In addition, unlike prototype or rule-based approaches, exemplar models maintain all of the interesting irregularities that might exist within a category. However, none of the assumptions made within the model preclude the use of cluster-based models, but an additional mechanism for abstraction would need to be integrated into the model in order to form the cluster's prototype.

1.2 Stimulus representation and similarity

Family resemblance theories of categorization suggest that category members resemble one another (or a prototype) more than they do non category-members. This stipulates that any complete theory must also establish a

method for determining the similarity between category examples. Central to a theory of similarity is an understanding of representation which establishes limits on how similarity will be determined. It is important that this measure of similarity be consistent with how people reason with the stimuli, otherwise any resulting model making use of the similarity function will not be able to effectively capture performance.

1.2.1 Featural approaches

Within the category learning literature, stimuli are generally represented as a collection of unrelated features (e.g. size=1,2,3; voiced=true/false). These features are then arranged in a vector, representing a point in a multidimensional space (Kruschke, 1992; Nosofsky, 1986; Rips et al., 1973) where each dimension represents a feature. A large blue square could be represented as [1,1,1], while a small red square would be [0,0,1]. This representation suggests that the similarity of any two items could be the distance between the points. In most cases the Minkowski distance is used, generally using an exponent of 1 (city-block) or 2 (euclidean), with an exponential decay. The exponential decay suggests that the effect of small differences between stimuli diminish as the stimuli get further apart. The exponential decay of similarity has been shown across species and suggested as a general law of generalization (see Shepard, 1987).

Rips et al. (1973) showed one of the first uses of a multidimensional approach for complex stimuli. In this study people rated the relatedness of

various animals and birds while the researchers recorded response times. Rips et al. (1973) found that a person's representation of the two groups, animals and birds, could be fit well by two dimensions representing size and ferocity. This approximation for the similarity between objects is now pervasive throughout the category learning literature due to its simplicity and sufficient accuracy for many tasks; however, it introduces several assumptions that are not fully supported.

Tversky (1977) showed that similarity ratings do not always follow metric axioms, and instead forwarded the *contrast model*, an idea based on set theory. The similarity between items is a function of the number of shared and unshared features between the stimuli. This does not assume that features have dimensions. While this approach addresses some irregularities between human similarity judgements and metric spaces, it is still not suitable for representing structure (Goldstone, Medin, & Gentner, 1991), as it assumes that features are independent. Instead, it is suited for situations where the features do not behave as continuous dimensions (Lee & Navarro, 2002).

While there is no formal definition stopping any of the above feature types from representing relations (configural cues, e.g. a bigger-than a breadbox feature), doing so poses several problems. The first is one of scaling: as the number of features in a scene grow, the number of encodings needed to represent all of the possible relations as features reach unaccountably large numbers. The second, related argument, is that such an encoding lacks structure specifying the relations between the features in a way that is parsimonious with

experimental evidence suggesting the role of structure in human reasoning.

Feature-based views of similarity hold that there are no relations between the features; they are independent. In contrast to this, several experiments have shown that changing certain features result in different affects on similarity depending on the relational support provided by other features (Goldstone et al., 1991; Goldstone, 1996). These experiments show that by changing features values from one scene to the next (which should reduce similarity), one can actually increase similarity by introducing shared relations between the two scenes. Goldstone et al. (1991) demonstrated that these effects cannot be accounted for by encoding the relations as features.

1.2.2 Structure mapping theory

In contrast to the purely feature-based approaches developed for category learning, relation-based approaches were developed to address how people compare stimuli containing structured relations; to explain how people perform analogies. Structure Mapping Theory (SMT, Gentner, 1983) emerged as one of the most influential descriptions of the analogical process, detailing both a representation and a means of comparing stimuli. Gentner (1983) suggested that stimuli should be considered a system of objects, attributes, and relations represented using a propositional framework of nodes and predicates. Single argument predicates represent attributes, i.e. bright(x), similar to Tvserky's (1977) view of features, and that mutli-argument predicates would encode relations between the attributes, i.e. brighter(x,y). Gentner also stipulated

that these representations should mimic the way people use and understand the system and not be considered a logically compact representation of the stimulus.

Structure allows for the encoding of relationships between the features within a stimulus. For example, Gick and Holyoak (1983) show how people extract complex structure from a set of examples and apply that structure to a featurally novel problem. People read stories where the solution to a problem followed a convergence schema, such as using a number of different lasers to destroy a tumor because any one laser powerful enough to destroy the tumor would also destroy any intermediary tissue. They found that when participants read two convergence stories before being asked to solve a problem that required convergence the participants were much more likely to solve the problem. The participants were able to extract the common relational structure present in the stories and apply it to a novel problem.

Work across children and experts suggests that the development of a structured knowledge base facilitates more complex reasoning (Chi, Feltovich, & Glaser, 1981; Gentner & Ratterman, 1991). By embedding structure into concepts it clarifies which parts of a stimulus are more central and should be reasoned about together. This approach has been used in areas other than psychology as well. The idea of compositionally, or developing reusable parts, is popular in computer science because it reduces the size of the knowledge base. Knowledge queries are more efficient when performed on a representation that makes effective use of structure and compositionally (Clark & Porter,

1997).

While logical compactness is important, Gentner (1983) actually specifies that it is not appropriate with regard to human reasoning; structures should be arranged according to use, and not be logically compact representations. Consider that Rips et al. (1973) reduces a complex category with complex members (e.g. birds) into two dimensions (e.g. size and ferocity). While these dimensions may accurately represent the similarity ratings and response times to similarity ratings of word pairs, these dimensions are not sufficient for explaining how people actually use and reason about the category of birds.

In order to explain how analogies between scenes take place, such as the above convergence schema, Gentner (1983) introduced structure mapping. Structure mapping is proposed as the resultant methodology explaining the comparison process between stimuli using a structured representation. Structure mapping is the idea that analogy and comparison is a mapping from one system onto another. These mappings are governed by a set of rules: attributes do not come with the object; relations should be preserved; and the systematicity principal, that preservation of higher-order relations in the mapping are preferred. Systematicity address the idea that analogy is about relational similarity, or the conveyance of structure. As such, good analogies should be at the deepest level possible, resulting in the most connected group of relations.

Structure mapping is directly modeled by the Structure Mapping En-

gine (SME) of Falkenhainer, Forbus, and Gentner (1989). SME implements structure mapping theory as a way to evaluate the quality of an analogy. SME does not attempt to address other issues related to analogy, such as retrieving exemplars from memory. Memory retrieval is often considered to happen over surface features (see MAC/FAC Forbus, Gentner, & Law, 1994), which explains why retrieval of structurally relevant exemplars is so poor (e.g. Gick & Holyoak, 1980).

SME places additional global constraints on the mapping process beyond Gentner (1983), namely one-to-one correspondence and the idea of support or parallel-connectivity. One-to-one correspondence requires that a mapping only link one element from each structure. Parallel-connectivity specifies that if two predicates are mapped, their arguments must also be mapped; this enforces that mappings maintain the structure of the analogs, an example is provided below.

The original version of SME follows a strict set of rules for which predicates are able to match one another. The rules specify that relations can only match identical relations, in that *greater* can only match *greater*, while entities are allowed to only match other entities that fill the same argument (or role) within the matching relation. Later iterations of SME relax the assumption that entity matches must preserve their role. Cross-mappings are those where the role of entities are not preserved, such as mapping the actor in one relation to the patient in another relation (Gentner & Toupin, 1986), see Figure 1.1.

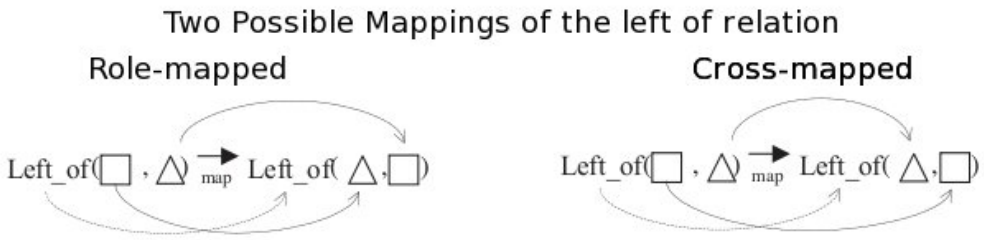

Figure 1.1: Role-based mappings preserve the argument structure of the entities within the mapped relations, cross-mappings do not

SME first generates a list of all possible mappings based on the local matching constraints and then prunes the list based on observance of the global constraints, such as parallel connectivity and systematicity. The list of allowable mappings are then scored according to how well they observe these constraints and a list of candidate mappings is generated.

The well known analogy between a solar system and an atom is a good medium through which to illustrate examples of these principals (Gentner, 1983; Falkenhainer et al., 1989). When making an analogy between the planets orbiting the sun in the solar system and electrons orbiting a nucleus in an atom, the best mapping would align the revolves relation that is present in both exemplars to one another (the sun with the nucleus, the planets to the electrons). This example is considered a good analogy because it observes one-to-one correspondence (each element is mapped to only one other element), parallel connectivity (the arguments of the revolves relations are mapped to one another), and there is no cross-mapping – the arguments play the same role in the mapped relations (the electrons and the planets are both *revolvers*).

Structure mapping has also been extended to a theory of similarity (Markman & Gentner, 1993, 1996). In order to determine the similarity between two items, a mapping is made between the nodes in one graph and the nodes in the other graph. The alignment is graded according to how well it maintains several key principals: one-to-one correspondence, a node can only be mapped to one other node; parallel connectivity, if two relations are mapped then their arguments should be mapped also; cross-mapping, if the role each node plays in its parent relation is preserved by the mapping; and systematicity, higher-order mappings have precedence. Once the best mapping is found between two graphs similarity is calculated as a function of how well the alignment preserves the structure between the two graphs (Markman & Gentner, 1993).

One notable effect of alignment on similarity is clarification of the importance of alignable and non-alignable differences. Alignable differences are those where a relational match dictates an alignment between non-matching entities, such as aligning a dog chasing a ball with a dog chasing a car. The ball and car serve as alignable differences. In contrast, non-alignable differences are those that occur outside of the common relational structure. Most feature-based views do not discriminate between these two types of differences (i.e. Tversky, 1977) – differences are differences. However, people actually report more differences in total, as well as more alignable differences, between similar objects than between dissimilar objects (Gentner & Markman, 1994), which demonstrates the relative accessibility of the two types of differences.

Markman and Gentner (1996) show how this accessibility affects similarity. Participants rated the similarity of pairs of stimuli that varied either by an alignable difference or a non-alignable difference. Importantly the scenes differed in one of four ways: a small alignable difference, a small non-alignable difference, a large alignable difference, or a large non-alignable difference. The researchers found that differences in alignable entities had a greater affect on similarity than differences in non-alignable entities.

Interestingly, the kinds of relational matches that are accepted plays an important role in determining similarity as well. SME limits itself to only mapping those relations that have identical labels, disallowing matches across different relations (e.g. *love* cannot map to *hate*). This is done to reduce the scope of the mapping problem, otherwise the number of possible mappings grows exponentially with the number of relations. However, structure mapping theory allows for circumventing the identically constraint by utilizing notions of re-representation and compositionally, which suggest that relations are made up of sub-relations, such as *give* and *take* both involving a $transfer - of - possession$, and that relations can be decomposed into these sub-relations to allow for mapping non-identical relations (Gentner & Kurtz, 2006).

In contrast to the above approach, the ACME model of Holyoak and Thagard (1989) considers all possible mappings, but uses the relational labels to grade the mappings after the fact . instead of positing re-representation, this approach suggests that mappings between non-identical relations imply dissimilarity between the relations. Several researchers suggest that this sim-

ilarity could be a product of how these relations are used, suggesting vector space models (Ramscar & Yarlett, in press; Turney & Littman, 2005) as appropriate measures. In this thesis we ascribe to the view that people are capable of entertaining mappings based on non-identical relations without recourse to conceptual re-representation, but that these mappings are constrained by the similarity of the relations.

In addition to the ACME model there are several other popular alternatives to SME, most of them also utilize the notion of structure mapping, but add complexities by accounting for additional effects. Models such as LISA (Hummel & Holyoak, 1997) and STAR (Halford et al., 1994) augment SMT with neural constraints within a connectionist representation. The constraints account for representational issues as well as capacity limitations such as working memory. AMBR extends the idea of structure mapping to memory and storage, focusing on integrating analogy with the rest of cognition (Kokinov, 1994).

While the above models add complexity to structure mapping beyond SME, the Connectionist analogy builder (CAB) of Larkey and Love (2003) simplifies the process of structure mapping. One pertinent effect demonstrated within CAB is that the principal of systematicity can be accounted for implicitly. A mapping that involves higher-order relations will necessarily have more of its structure mapped through the parallel connectivity constraint, than one that does not, allowing systematicity to fall out of models that select mappings based on a similarity measure that includes parallel constraints and one-to-one

correspondence.

Alternative accounts of similarity have also been proposed that attempt to account for the time-course of similarity without supporting the notion of fully-fledged structure mapping during the comparison process. These alternatives follow the idea that structure mapping is a complex process and doesn't occur without considerable effort. They suggest that simpler, faster processes can operate when time, or other resource limitations, are a factor.

The Similarity, Interactive Activation, and Mapping approach of SIAM (Goldstone, 1994b) is one such model. SIAM represents similarity as a function of the Matches In Place (MIPS, features that match on the same attributes, e.g. head color) and Matches Out of Place (MOPS, features that match on different attributes, e.g. head vs. tail color). This approach focuses on the importance of dimensional correspondence between different attributes allowing it to mimic more complex structured approaches for simple stimuli that do not have a deeply structured representation. SIAM has proven itself by fitting the time-course of similarity judgements (Goldstone & Medin, 1994) as well as being useful in suggesting a role for analogy in category learning (Lassaline & Murphy, 1998).

In contrast to SIAM, transformational theory or representational distortion (Hahn, Chater, & Richardson, 2003) is agnostic to the notion of a structured representation and posits that it is the number of transformations required to change one scene into the other that determines similarity. For example, to turn the string *ab* into *ba* a swapping transformation is required.

This approach does not specify the set of allowable transformations, instead it is suggested as a framework for discussing an alternative account of similarity.

Structure mapping, SIAM, and the transformational account provide ways of accounting for analogical-similarity, but in this paper we will frame the discussion in terms of structure mapping theory because of its completeness and historical significance to the field. However, it is important to note that similarity-based categorization does not require a particular measure of similarity allowing the use of any of these measures and might provide an additional avenue to explore the relative merits of the various approaches.

1.3 Learning and attention

1.3.1 Analogy

Research in analogy generally looks at how people extract, reason about, and recollect the relational structure within examples. For instance, Gick and Holyoak (1983) show how people are poor at abstracting schemas from single examples, but are more successful if they are provided with multiple similar examples which they can compare. Markman and Gentner (1993, 1997) expounded on this as being a product of a synergy between structural alignment and similarity; that similarity judgements make use of structure and in turn promote the understanding and memory of the deeper structure. Kurtz and Loewenstein (2007) build on this theory, demonstrating that after a schema is extracted through a comparison process that schema can be used to facilitate recollection of earlier read, relationally related material. These

experiments accomplish their task by using a few structured examples.

Complementing the above work investigating reasoning and memory, is a trend for work in analogy to focus on learning. This is accomplished through multiple trials. For instance, Kotovsky and Gentner (1996) introduce a series of experiments looking at two important theories of learning through analogy. The first is re-representation, which is the idea that as new knowledge is acquired the structure of older knowledge is changed to reflect it. The researchers suggest that children extract a novel higher-order relation (e.g. greater-than) from comparing concrete examples of lower-order relations (e.g. taller-than and brighter-than) and in turn re-represent those examples as containing a higher-order greater-than relation.

A second idea, which builds off of re-representation, is progressive alignment. Progressive alignment suggests that the learning of structure is progressive and supported by similarity. Learners first extract low-order structure by comparing highly similar examples, and this in turn supports the future extraction of higher-order relations because examples containing those relations are now more similar due to their re-representation in terms of the now known low-order relations. As exemplified in Kotovsky and Gentner (1996), children progress from a knowledge of taller-than and brighter-than to the more abstract greater-than relation, allowing them to solve more complex problems. The success of the similarity portion of progressive-alignment is shown by experiments such as Gentner, Loewenstein, and Hung (2007), which shows that children can be bootstrapped into extending part-labels (e.g. arm) to more

dissimilar looking stimuli by first teaching them the part name with more similar looking examples.

Re-representation is seen as one part of the explanation for the relational shift, the change in children from a feature-based understanding of a domain, to a relational understanding (e.g. grandpa as man with grey hair to your parents' fathers, see (Gentner & Ratterman, 1991)). A more developed relational view of a domain, utilizing relations with more arguments, or more abstract representations, will contribute more relational support to a matching problem reducing the likelihood of being distracted by mere-appearance matches. However, in addition to this support, issues such as the ability to suppress attention to feature-based matches and working memory capacity also have been found to play a role (Richland et al., 2006).

1.3.2 Category learning

In contrast to the general approach in the analogy literature, one of the primary focuses of category learning research is charting the evolution of learning throughout tens to hundreds of trials. A common format for a category learning study is as a classification task. Subjects are presented with a series of stimuli that come from a small number of categories. The subject is shown a stimulus and then asked which category it came from. Feedback is then provided to the subject as to whether or not they correctly classified the stimuli. This process is repeated until the subject reaches some learning criterion.

Shepard et al. (1961) provided data from a set of experiments using this classification paradigm that is still a subject of investigation to this day. The researchers investigated how long it took people to learn a binary classification problem depending on which of six category structures was used. These data are still relevant because they form an initial test of many models of categorization, and because Shepard et al. (1961) specified that the pattern of learning across the six categories could not be accounted for without a model that used selective attention. Kruschke (1992) presented a model supporting this hypothesis, and showed that learned selective attention is critical to successfully fitting the pattern of difficulty across Shepard et al.'s category structures.

Selective attention expresses the idea that different dimensions of a stimulus will receive more weight than others during the comparison process (Nosofsky, 1986). For example, when categorizing things as red or white, differences along the color dimension are more important than difference along other, irrelevant, dimensions (e.g. size). Selective attention theory has been integrated with learning approaches, suggesting that learners learn to attend to the dimensions that are relevant to the categorization task (Kruschke, 1992). In the case of a multidimensional stimulus representation selective attention has the effect of stretching space along the attended dimension, and shrinking it along non-attended dimensions. In this way attention can work like abstraction, in effect, the unattended dimensions are abstracted away and are no longer relevant to the categorization process.

Importantly learned selective attention has moved beyond merely being a weight attached to a dimension during a similarity decision. For instance, it has been shown it is faster to learn novel associations between attended dimensions and outcomes than between inhibited dimensions and outcomes (Kruschke & Blair, 2000). Also, learners exhibit greater perceptual sensitivity to differences along dimensions that discriminate between categories than on those dimensions that do not, even after learning (Goldstone, 1994a). In addition, changes in attention weightings predict looking times to stimulus dimensions during a category learning task. Learners spend more time looking at attended dimensions that predict the category label than those that do not (Kruschke, Kappenman, & Hetrick, 2005; Rehder & Hoffman, 2003). These effects ground attention as a cognitive process that has implications outside of a simple parameter within a classification model.

Kruschke (1992) introduced the ALCOVE model of category learning to formalize the application of learned selective attention to category learning. ALCOVE used a multilayer artificial neural network. In the network each node in the input layer corresponded to a stimulus dimension, while there was one node in the hidden layer for each exemplar in the stimulus space, and one output node for each category label. Activation of the hidden layer is a function of the attention-weighted distance between the input and the exemplar represented by that node, in turn the category nodes are activated according to the weighted sum of the activation of the hidden nodes. This is an implementation of exemplar theory. Learning in the network takes place

using error-driven learning – the attention to each dimension and the association between each exemplar and the category nodes is adjusted to minimize classification error on each trial.

ALCOVE provided a framework that could be used to investigate the theories of error-driven learning and selective attention, and to account for several interesting findings in the learning literature and how they pertain to classification. As already mentioned, ALCOVE provided support for Shepard et al.'s (1961) hypothesis appropriately capturing the difficulty of certain category structures requiring selective attention. Selective attention is a means to abstraction and can lead to rule-like behavior. In addition, several phenomena have also been shown to require selective attention: *blocking*, *highlighting*, and the *inverse-base-rate-effect*.

Blocking describes a situation where a novel predictive cue is added after an association has already been formed between an existing cue and an outcome; learners do not associate the new cue with the old outcome, even though it is predictive of the outcome (Kamin, 1969). Blocking effects can be captured by assuming that learning is error-based, or a response to surprising events (Rescorla & Wagner, 1972), and not correlational. Because the first cue perfectly predicts the outcome, there is no need to learn the secondary association.

However, more recent work has shown that blocking is actually slightly more complicated; in addition to not learning an association between the cue and the outcome, inattention to the cue is also learned (Kruschke & Blair,

2000). The addition of a limited pool of selective attention is able to fully account for this phenomena. The introduction of the novel cue and required redistribution of attention to that cue forces additional error in the prediction system, since that dimension is not predictive of the outcome. The attention learning mechanism fixes this classification error by reducing attention to the novel dimension and increasing it to the older, predictive dimension.

Highlighting is similar to blocking and also explainable through selective attention. Highlighting occurs when two cues have been paired with an outcome (A and B $\rightarrow$ X) followed by the interleaving of trials pairing a novel outcome with a conjunction of a novel cue and one of the already trained cues (B and C $\rightarrow$ Y). After learners have seen both pairings the same number of times, they are tested with cue B and a conjunction of A and C. Both of these are equally correlated with both X and Y; learners strongly prefer outcome X for cue B, and outcome Y for cue A and C. Highlighting cannot be explained simply through error-driven learning, and instead requires selective attention (Kruschke et al., 2005).

In addition to a more refined notion of blocking and an account of highlighting, selective attention allows for the capturing of the inverse-base-rate effect. The inverse-base-rate effect (Medin & Edelson, 1988) is when people pick the less frequent response when presented with ambiguous information that should have a stronger association with the more prevalent outcome. For example, if you often see that a runny-nose and fever mean you have the flu, but once in a while you see that a fever and pink-spots mean the chicken-pox,

when presented with a case of runny-nose and pink spots, people diagnose it as the chicken pox, even though runny-noses should have a more established prediction of the flu, because you see them more often. Interestingly, this effect is a problem for many approaches to learning, but can be captured through the use of selective attention (Kruschke, 2003, and see Kruschke, 2001).

The general framework behind the ALCOVE model has inspired many different iterations of models that focus on more accurately modeling different aspects of category learning. These modifications easily translate to different parts of the model. For instance, SUSTAIN changes the hidden layer to enable use of a cluster-based category representation and the output layer to an auto-encoder for unsupervised learning (Love et al., 2004). RASHNL looks at creating more accurate fits to the effects of attention-shifting during category learning, by changing the amount of learning that happens per trial. People appear to make rapid attention shifts as a form of hypothesis testing (Kruschke & Johansen, 1999). EXIT, with some other modifications, introduces exemplar specific attention (Kruschke, 2001). ALCOVE also allows for easily changing the definition of similarity within the model; Lee and Navarro (2002) created a contrast model of ALCOVE (as mentioned in the stimulus representation section).

We build off these core ideas and create our theory by enriching the notion of exemplar representation and similarity within this general connectionist framework. We stipulate that the learning and attention process is no different whether one is learning to discriminate categories based on feature

values or the presence of certain relationships between the feature values. This means that many of the important findings associated with exemplar-based, and attention-based learning should apply equally as well to tasks involving relations. Before describing the model in detail, we briefly present how attention will be integrated with relation-based similarity.

1.3.3 Integrating attention with analogy

Similarity-based category learning models contain well developed theories of learning (i.e. error-based learning, and attention), but the models lack a representation that is capable of representing the similarity between structured exemplars in a manner that is consistent with experimental evidence This suggests the importance of this structure (e.g. Goldstone et al., 1991). Analogy-based theories of representation (and similarity) can provide category learning models with the necessary representation to learn the more complex associations between relations and category labels.

The key to learning relation-based categories is to bridge the theories from analogy and category learning. By integrating the notions of similarity-based categorization, attention-driven learning, and analogical similarity, we develop a theory where learners can shift attention from feature-based to more abstract, relation-based responses when necessary. Importantly, there is no need to fundamentally change any of the underlying architectures, so the new approach reduces to any of its components, when applicable.

Instead of relying on a feature-based comparison method, as is common

in category learning, BRIDGES will use an analogical alignment between the stimuli because it is a better measure of similarity for exemplars containing structure (Goldstone et al., 1991). However, the mapping process will be augmented with attention. In this way, abstractions may be formed when suggested by the category structure by shifting attention from a mapping driven by the concrete features of a stimulus to one driven by the relations between those features.

For example, imagine aligning a dog chasing a ball with a scene of a person chasing a dog. In this case, mapping the dog from the first scene to the person in the second scene is the better relational alignment as it preserves the roles of the mapped agents, and it would be chosen by a model of analogy. However, the alternative mapping, the dog from one scene to the dog from the other scene, is a better feature-based alignment as it preserves the features of the scenes. By adding a learned attention shifting mechanism, the choice is dictated by what is being learned. If one is learning about the relation-based category *chasing*, then the first, role-based alignment is the more likely choice, and instead, if you are learning about the category *dog*, the second, cross-mapped alignment would be preferred as it maintains feature-based regularities (see Figure 1.1 for a comparable example).

1.4 Building BRIDGES

Building Relations through Instance Driven Gradient Error Shifting (BRIDGES) is an exemplar based category learning model where new items

are categorized according to their similarity to other items. Novel stimuli are placed into the category with the most similar members. By extending the notion of similarity normally used in categorization models (vector distance in a multi-dimensional space (Shepard, 1964)) to make use of a similarity measure based on relational theories (i.e. structural alignment, Gentner & Markman, 1997), it can be shown that relational categories, such as same-different or simple grammars, can be learned by making comparisons between concrete category members and novel stimuli (Tomlinson & Love, 2006).

BRIDGES is based on the ALCOVE model (Kruschke, 1992), an artificial neural network based exemplar category learning model. Exemplar theory holds that each category exemplar is stored in memory and that novel items are classified according to their similarity to those exemplars (Medin & Schaffer, 1978). When a stimulus needs to be classified a similarity score is computed between the stimulus and all exemplars in memory (hidden layer consists of exemplars). This similarity score can be adjusted by an attentional mechanism that weights some aspects of stimuli more than others (Nosofsky, 1986). Based on these similarity scores, the stimulus is predicted to belong to the same category as the most similar exemplars (output unit activation). ALCOVE made a critical update to older models by adding the ability to learn the attention weights through an error-based gradient descent method.

BRIDGES acts like ALCOVE: storing each exemplar in memory, making comparisons between novel stimuli and previous exemplars, classifying based on attention-weighted similarity, and learning through gradient descent.

However, in the BRIDGES model stimuli are assumed to have a richer representation and a richer notion of similarity. The exemplars have a structured representation and similarity is derived by finding the best alignment between the two stimuli. The quality of the alignment, how well it preserves parallel connectivity, role information (i.e. agent aligning with agent, vs. agent with patient), and how well it preserves feature matches, determines the similarity between two stimuli. Unlike most instantiations of this process (Falkenhainer et al., 1989; Larkey & Love, 2003), BRIDGES allows weighting of the alignment by attention. If relational information is learned to be more important, then alignments preserving role mappings will be preferred; however, if a particular feature is more diagnostic of category membership, then an alignment that preserves feature matches is preferred (see Figure 1.1). In a similar manner, if transformational similarity is assumed, it is the number of transformations where each transformation is assigned a weight based on learned attention that determines similarity.

BRIDGES is effective for a wide array of learning tasks. In the case where relational information is not pertinent, such as most experimental tests within the category learning literature, it reduces to the standard ALCOVE model. But when relational information is applicable, the model shows its analogical roots. Furthermore, since BRIDGES is a simple combination of extensively developed theories in category learning and analogy, many of the extensions proposed in either literature are applicable to the model. For example, BRIDGES is compatible with a variety of attention shifting theories

(RASHNL, Kruschke & Johansen, 1999), choice rules (see Wills, Reimers, Stewart, Suret, & Mclaren, 2000 for an alternative to the Luce choice rule), or even prototyping or clustering theories such as SUSTAIN (Love et al., 2004).

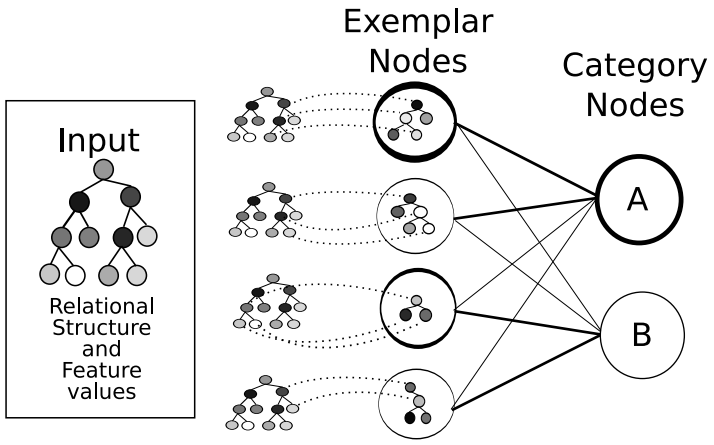

Figure 1.2: Depiction of BRIDGES. The input is mapped separately to each exemplar. Each exemplar is activated proportional to the similarity between the input and exemplar and then activation flows through the network to the category nodes weighted by their connection. The luminance of the filled circles represent the attention strength to each attribute, while the density of the nodes represents activation level and the density of the lines represent weights.

1.5 Formal Description

BRIDGES is a three-layer feed forward artificial neural network, more specifically a radial basis-function network (Lippmann, 1989). Figure 1.2 illustrates an example BRIDGES network. The input to BRIDGES contains both featural and relational information in the form of propositions and arguments. Each possible feature or relation has an associated attention weight

(connection to hidden-layer). The activation (i.e., similarity) of each hidden unit (exemplar) j by stimulus s is

$$h_j = max_{m \in M}(exp[-c(\sum_i \alpha_i \cdot Mismatch(e_{ji}, s_i))]) \tag{1.1}$$

where m∈M is all possible one-to-one mappings, preserving parallel structure between nodes (i.e., features, entities, and relations) forming stimulus s and exemplar j, i ranges over nodes in stimulus s, while α_i is the attention weight associated with node type s_i. e_{ji} is the node in exemplar j that is mapped to s_i, and c is the specificity parameter that determines the rate at which activation falls off with increasing mismatch. Mismatch is defined to be 1 for features or entities in stimulus s that map to non-identical nodes in exemplar j, otherwise 0. For relations, Mismatch is defined to be 1 for relations in stimulus s not exhibiting role-mappings, otherwise 0. For the purposes of determining similarity all possible mappings are considered, and the mapping that results in the highest similarity between the units is chosen.

Activation passes from exemplars to category unit o_k:

$$o_k = \sum_j w_{kj} \cdot h_j \tag{1.2}$$

where w_{jk} is the association weight between exemplar j and category unit o_k. The probability of selecting the category corresponding to category unit r is

$$Pr(r) = exp(\phi o_r) / \sum_k exp(\phi o_k) \tag{1.3}$$

where ϕ is a decision parameter and k ranges over all category units.

Learning is accomplished via gradient descent error minimization using a supervised or unsupervised error function. In the supervised rule the target value for the category unit corresponding to the correct category is set to 1 and other category nodes are set to 0. A "humble teacher" scheme is used so that category unit output values in the correct direction greater that 1 or less than 0 are not penalized. The error function E minimized is

$$E = 1/2 \sum_{k} (t_k - o_k)^2 \tag{1.4}$$

where t_k is the target value for o_k. The association weight w_{kj} from exemplar j to output unit k is adjusted by

$$\Delta w_{kj} = \lambda_k (t_k - o_k) h_j \tag{1.5}$$

where λ_w is the learning rate for association weights. Attention weights are updated by

$$\Delta \alpha_i = -\lambda_\alpha \sum_{j} [\sum_{k} (t_k - o_k) w_{kj}] h_j \cdot c \cdot Mismatch(e_{ji}, s_i) \tag{1.6}$$

where λ_α is the learning rate for attention weights.

Simulations 1 and 2 model an unsupervised learning case, where the network is predicting familiarity. For those simulations we adopt an approach that follows the modeling in Love et al. (2004). A single category output unit with a target value of 1 is used for all stimuli. In effect, this category unit is a familiarity detector, and stimuli familiar to the network will give high (around 1) activation while unfamiliar stimuli will yield lower activation. Learning

proceeds as above and the association and attention weights in the model are adjusted to uncover the underlying regularities across the inputs.

Simulation 5 models a two alternative forced-choice matching paradigm, which requires selecting which of the choices is most similar to a standard. As such, the model calculates the similarity between each choice and the standard, this is equivalent to treating the two choices as nodes within the hidden layer with the standard as the input, and treating the activation at the hidden layer as the output (e.g. $o_j = h_j$). The model probabilistically selects which choice *goes with* the standard, according to

$$Pr(a) = \frac{h_a}{h_a + h_b} \tag{1.7}$$

The probability of a given response is maximized to reinforce the decision, or minimized to reinforce the alternative decision depending on which choice the model selected. Steps are taken along the gradient to reinforce the choice selected by the model.

$$\frac{\partial Pr(a)}{\partial \alpha_i} = \frac{h'_a(h_a + h_b) - h_a(h'_a + h'_b)}{(h_a + h_b)^2} \tag{1.8}$$

where

$$\frac{\partial h_j}{\partial \alpha_i} = h_j * c * mismatch(e_{ji}, s_i) \tag{1.9}$$

to minimize or maximize $Pr(a)$, depending on the choice made by the model. Based on the above gradients, the attention updating equation for this case can be simplified

$$\Delta\alpha_i = \pm\lambda \left[mismatch(e_{ai}, s_i) - mismatch(e_{bi}, s_i)\right] * \frac{h_a * h_b}{(h_a + h_b)^2} \tag{1.10}$$

and it becomes easy to see that the model shifts its attention to highlight differences between the mappings of one choice and the standard, and the alternative and the standard. Shifts are only made when there are differences in the diagnosticity of the mappings for a particular attribute. The size of this update is a function of the relative similarity of the two choices. When the choices are maximally similar, the model makes the largest shifts in attention, and conversely if the two choices are not similar the model will not shift attention.

Chapter 2

Simulations

The goal of these simulations is to evaluate the theory that relation-based categories may be learned without recourse to rules or re-representation through similarity-based comparison of concrete exemplars. These simulations detail how relation-based categories may be learned by animals, children, and adults, and will highlight the differences between our theory and other approaches to relation-based category learning. To demonstrate these points the simulations cover a range of different category types, from learning simple grammars to relational match-to-sample tasks, learning effects, such as typicality, and feature-based interference. We then conclude by looking at the differences between re-representation and similarity-based categorization.

Simulation 1 models a short experiment, Marcus et al. (1999), which showed that infants are capable of learning to distinguish sequences of speech sounds based on relationships between the phonemes within the sequences. This study has evolved into a de-facto standard for any model of relation-based categories (Altmann & Dienes, 1999; Kuehne et al., 2000; Shultz & Bale, 2001; Seidenberg, Marcus, Elman, Negishi, & Eimas, 1999). It provides a good demonstration of BRIDGES as it has a simple experimental structure

and clearly shows how learning takes place.

Simulation 2 looks at classification advantages of a similarity-based learning system over a rule-based one. Gerken (2006) expanded on Marcus et al. (1999). She showed that infants are only willing to form abstractions to the degree specified by the stimuli. In contrast to rule-based approaches which do not clearly specify when an abstraction should be formed, appropriate stimulus-based abstractions are fundamental to the design of similarity-based categorization.

Simulation 3 investigates typicality effects, which provide another fundamental argument against rule-based views of category learning. Some members of feature-based categories are more typical of a category than others, garnering more accurate and faster responses than other members (Rips et al., 1973). For example, people are generally faster at identifying dogs as mammals than at identifying bats as mammals despite the fact that a simple rule – does it have a mammary gland – identifies members of the category. These effects are taken as strong evidence against rule-based approaches to categorization. Simulation 3 models an experiment showing the same phenomena for relation-based categories; some exemplars of a relation are more typical of the relation than others.

Simulation 4 demonstrates another divide between categories defined by similarity or rules. Allen and Brooks (1991) showed that even when provided with a uni-dimensional rule for the category, people are still affected by irrelevant features of the stimuli. An attention-shifting explanation of relation-

based category learning predicts that this effect will occur for relations as well. If the features and relations within a stimuli both predict the category label, then both will be used in classification decisions.

Finally, simulation 5 directly compares BRIDGES to theories of re-representation. Re-representation predicts that learners acquire novel structure when learning relation-based categories (Doumas et al., 2008; Kotovsky & Gentner, 1996). In contrast, BRIDGES stipulates that the the learners are simply shifting attention and comparing exemplars in a way that produces the same effect. Simulation 5 looks at Kotovsky and Gentner (1996) which presents evidence that children can learn to re-represent stimuli through progressive alignment. We suggest that what is shown is not re-representation, but a change in attention that looks like a restructuring of knowledge.

2.1 Simulation 1 - Abstraction Without Rules

A stringent test for modeling abstraction is in being able to fit Marcus et al. (1999), which shows abstraction over an artificial grammar. Marcus et al. (1999) showed that infants could learn to distinguish instances of two different artificial grammars even when the distinction was not supported by any featural regularities within the grammars. The researchers proposed that infants must learn algebraic rules (substitution, where variables could stand in for a class of segments) coupled with variable binding to succeed in their task.

In Marcus et al.'s (1999) study, seven-month-old infants were exposed,

for 2 minutes, to 16 unique sentences that followed either an AAB pattern or an ABB pattern. The sentences were made up of simple monosyllable sounds (words) such as "GA GA TI". The researchers then measured the infant's looking time to speakers producing novel instances of the training grammar and the opposite testing grammar. Those infants that were trained with one grammar showed a familiarity response to novel instances following the trained grammar (e.g. there was no significant change in looking time), and a novelty response to instances not following that grammar (e.g. looking time significantly increased). Below we show how BRIDGES is able to capture that same pattern of responding without recourse to rules.

Each syllable is represented as an entity. Each syllable's position in the speech stream is encoded by a positional feature. These syllables have a number of phonetic features that are not represented in these simulations. Not including such features simplifies the simulations (and follows Marcus et al's (1999) presumption that no significant regularities exist across these features). More importantly, not including these additional features provides a strong test of BRIDGES as there are no featural regularities that BRIDGES can rely on to discriminate between the grammars.

2.1.1 Method

2.1.1.1 Representation

Each sentence (e.g., "JE WI WI") was represented as an exemplar. Table 2.1 shows a tabular representation of the stimuli. Each phoneme is

Table 2.1: BRIDGES's representation of "GA TI TI."

Entities	Features	Relations
JE_1	$Position(JE_1) = 1$	$TypeOf(JE_1, JE)$
WI_1	$Position(WI_1) = 2$	$TypeOf(WI_1, WI)$
WI_2	$Position(WI_2) = 3$	$TypeOf(WI_2, WI)$

represented as an entity with associated feature values corresponding to its position in the speech stream and a single value standing in for its phonetic features. Following Marcus et al.'s (1999) presumption that no significant regularities exist across these features, phonetic features are not represented in these stimuli, except as match/no match. With no featural regularities, this simulation becomes a strict test of BRIDGES capabilities to use relations to discriminate the grammars.

Critically, relational information was included in BRIDGES's representations by making a distinction between types and tokens. In effect, the model assumes that infants in Marcus et al.'s (1999) study have developed categories of speech sounds (Eimas, Siqueland, Jusczyk, & Vigorito, 1971). These type relations allow for abstract patterns to be uncovered through analogy to stored exemplars as one category of sound can mapped to another.

A graphical rendition of the actual input to the BRIDGES model for "JE WI WI" is shown in Figure 2.1. Starting at the top of the image, the top node acts to tell the matching utility that this is a single stimulus. The next three circles describe a type/token relation between the concrete phonemic tokens ($je_1,\ wi_1,\ wi_2$) present in the auditory stream and the abstract types

instantiated by those tokens (aje, and awi). The third row contains the tokens and types. The fourth row contains feature values for the entities on the third row, position, phonetic features, and a class node defining the entity as a type or token (entity). The names along the arcs between the third and fourth rows represent the value of that feature for that entity.

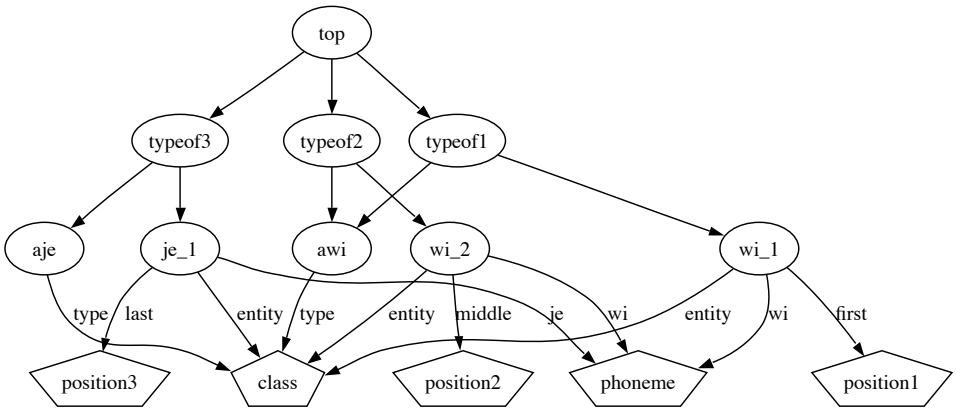

Figure 2.1: A graphical representation of one of the stimuli from Marcus et al. (1999). This image was generated using GraphViz

2.1.1.2 Fitting details

The training and test regimen followed the original study as closely as possible. During habituation, the 16 unique sentences were presented three times each to BRIDGES. On each presentation, association and attention weights were updated. Though not critical, we assumed that the salience of positional features is sufficiently great to constrain the mapping process (i.e., words in sentences align temporally). As the data points provided by Marcus et al. (1999) are only consistent in their pattern (e.g. looking times

to unfamiliar stimuli differ significantly across the three experiments) we only considered the overall pattern of results when fitting the studies, consequently any parameter set with non-zero learning will show the same pattern.

2.1.2 Results and Discussion

After training, BRIDGES correctly responded to novel items from the old grammar with higher activation (.94) than to novel items from the new grammar (.1). The model was able to decide the grammaticality of the items because parallel connectivity was perfect between novel exemplars of the new grammar and the training exemplars it had already learned to recognize. In contrast, ungrammatical items did not maintain parallel connectivity when mapped to the training exemplars. For instance, "GA TI TI" is isomorphic to "LI NA NA" in that all token and types in the type relation (see Table 2.1) can be mapped to one another and preserve parallel connectivity.

BRIDGES was only able to achieve this result because it learned to shift its attention away from the inconsistent, hence non-diagnostic, features and to the consistent relational alignment. This shift makes BRIDGES sensitive to the underlying grammar; novel sentences following the original grammar are now familiar. Sentences not following the learned grammar do not analogically match stored exemplars (parallel connectivity is violated), making these items less familiar and resulting in greater looking time as infants dis-habituate.

This experiment has been modeled by numerous researchers (Altmann & Dienes, 1999; Kuehne et al., 2000; Shultz & Bale, 2001; Seidenberg et

al., 1999). These models generally fall into two groups. The first are different variations of artificial neural networks that rely on a distributed representation. These approaches suffer issues with regards to the massive amount of pre-training that is generally needed, and the assumptions that are built into the models architecture which restrict the generalization of the network to other problems. Additionally, these networks generally require numerous exposures to the training stimuli. In contrast, BRIDGES learns to discriminate between the two grammars in the same number of trials as the infants. The second group (Kuehne et al., 2000) uses structural alignment to compare examples and extracts a prototype of the grammar through the alignments. This approach is very similar to our own, except that they generate a prototype instead of using a process of attention-based similarity. The disadvantages to this approach were described in the Introduction.

Marcus et al. (1999) has criticized other accounts (e.g., Seidenberg & Elman, 1999) of these results for including a same/different detector within a learning mechanism. The BRIDGES simulations do not explicitly label speech sounds as identical, rather the model assumes that infants can categorize speech sounds (Eimas et al., 1971), as embodied by the type/token distinction. BRIDGES's solution does not hinge on a same detector. In fact, the patterns that can be discriminated by analogical mapping (even in simple domains in which only the type relation is present) are more encompassing than same/different. The analogical mapping process in these simulations aligned the current stimulus to stored exemplars — BRIDGES did not label

words within sentences as same or different nor did it shift attention to a *same* feature. Abstract responding arose through analogy to stored exemplars and attention shifting from concrete features to relations. Furthermore, the model is able to achieve this result without pre-training, with a similar number of training trials, and BRIDGES is generalizable to other tasks.

2.2 Simulation 2 - Appropriating Abstraction

The above simulation showed that infants can form relational abstractions over phonemes without recourse to rules by shifting attention away from the concrete features of the stimuli to relations within the stimuli. However, the study can also be well fit by rule-based approaches, raising the question of the necessity of a similarity based approach. To demonstrate the ineffectiveness of rule-based accounts of grammar learning, Gerken (2006) designed a series of experiments which are problematic for rule-based approaches as rules lack a clearly specified mechanism designating at which point generalizations should/not be formed. In contrast, BRIDGES specifies that the amount of abstraction that will occur is dependent on the statistical regularities in the environment.

Gerken (2006) conducted two experiments to determine what generalizations infants would extract from different strings of phonemes following either an AAB (two identical and one non-matching phoneme) or AA*di* (any two identical phonemes, always followed by the *di* phoneme) pattern. The critical manipulation in her study was whether or not infants would still learn

an abstract AAB pattern if they were presented with an AA*di* pattern instead of a general AAB pattern (any two phonemes followed by any other phoneme). She found that if infants were presented with an AA*di* pattern, they were not able to generalize to an AAB pattern. They had learned AA*di* instead, and could distinguish AA*di* from AAB, ABA and A*di*B patterns (see Table 2.2).

Table 2.2: Sample stimuli and results from Gerken (2006) and BRIDGES' simulation of the study

	Training	Test 1	Test 2	Test 3
Abstraction	*AAdi*	*AAdi*	*ABA*	*AAB*
Sample	*je je di*	*la la di*	*wi ja wi*	*le le we*
Infant Response	Familiar	Familiar	Novel	Novel
BRIDGES	Familiar	Familiar	Novel	Novel

This experiment is one of the key arguments against a rule-based interpretation of the results from Marcus et al. (1999). If one is simply extracting rules, it is unclear when to stop generating more abstract rules. In contrast, BRIDGES relies on building a statistical model of the variability within the the stimuli to dictate how much attention should change. The below simulation shows that this mechanism is able to find the appropriate level of abstraction.

2.2.1 Method

2.2.1.1 Representation

This simulation used the same representation as Simulation 1.

2.2.1.2 Fitting

Gerken (2006) followed the same procedure as Marcus (1999), except that only 4 different training exemplars were used instead of 16. Likewise, the BRIDGES model was trained in the same manner as that used in the above simulation, consistent with the reduced variety of training trials given by Gerken (2006). The model received 50 training trials with each of the four stimuli. Like the above simulation the model was reinforced to produce a 1 for a familiar item and a 0 otherwise. Again, as above, we fit the model to the pattern of results, and not the exact differences.

2.2.2 Simulation and Discussion

In Gerken (2006), infants trained with stimuli following an AAB pattern showed a high familiarity to novel stimuli following an AAB pattern and a low familiarity to stimuli exhibiting an ABA pattern, exactly as in Marcus, et al. (1999). In addition infants trained with a set of stimuli following an AA*di* pattern exhibited a low familiarity to stimuli following an A*di*A, AAB, and ABA pattern, and a high familiarity to a novel AA*di* pattern. In contrast to the results, if the children (or BRIDGES) over-generalized, you would expect the model to respond as strongly to the novel AAB pattern as to the AA*di* pattern.

BRIDGES does not overgeneralize, and captures the pattern of responding perfectly – familiarity for the model is a number between 0 and 1 with 1 being highly familiar and 0 being not familiar. In the first case (trained

on AAB), the model responds with .02 (Novel) for a novel ABA stimulus and .95 (Familiar) for a novel AAB stimulus. In the second case, when the model is trained with AA*di*, the model responds with a .13 to AAB (Novel), .01 (Novel) to ABA, and .02 (Novel) to A*di*A, but with a .99 (Familiar) to a novel AA*di* stimulus, exactly as the children (see Table 2.2).

In the simulation, as in the original study, the key difference between the two cases is the consistency of the last phoneme. When that phoneme is consistent, BRIDGES learns to look for that phoneme in that position; however, when the phoneme is inconsistent, BRIDGES only focuses on the consistent relationship between the phonemes (that the first and second are the same and the third is different).

BRIDGES provides a perfect fit to the pattern of the data found in Gerken (2006). It suggests that abstraction happens by way of a comparison process to previous training items and a shifting of attention to the items that exhibit statistical regularities. BRIDGES shows how the statistics function not just over single phonemes or abstract place holders, but over groups of phonemes bound together by relations.

2.3 Simulation 3 - Typicality effects in relation-based categories

In the two preceding simulations, we showed how infants could learn to discriminate between stimuli that followed an abstract grammar, or did not. We demonstrated that they could learn to attend to the consistent relational

mappings between grammatically correct stimuli and ignore the inconsistent features. Simulation 3 looks at the implications of using similarity functions to delineate relation-based categories and shows how typicality effects provide more evidence against a rule-based view of abstraction.

The idea of typicality, or graded membership, is classically feature-based (Barsalou, 1985; Lynch, Coley, & Medin, 2000; Rosch & Mervis, 1975), meaning some category members are considered better examples of a category and are often identified faster. For example, to most individuals, robins are better examples of birds than are penguins. Because BRIDGES inherits this graded notion of category membership by relying on a similarity-based exemplar model, it extends it to relation-based categories as well. BRIDGES predicts that relation-based categories exhibit the same tendency for graded membership; some exemplars of the *on* relation are better examples than others.

To evaluate BRIDGES's predictions, we will consider results from a series of studies exploring how pigeons and humans learn notions of same and different. To illustrate how BRIDGES learns the concepts *same* and *different*, we will apply BRIDGES to Young and Wasserman (1997) study of *same/different* discrimination learning in pigeons. To foreshadow, Young and Wasserman's results indicate that pigeons can master a notion of same and different that cannot be explained by featural similarity. At the same time, the pigeons are sensitive to the particular examples they experienced during training and display a graded notion of same and different. Although fascinat-

ing, it would be easy to dismiss these results as relevant to pigeon cognition, but not human cognition. However, work by Young and Wasserman (2001) found the same pattern of performance with human subjects. Humans as a group are slightly more deterministic than pigeons, but this group difference is within the range of individual differences. The bottom and top 20% of humans clearly bracket the mean performance of pigeons.

In Young and Wasserman (1997), pigeons learned to respond differentially to displays containing 16 identical and 16 different icons. On each trial, the 16 icons were randomly placed within a 5 X 5 grid. The pigeons were reinforced for pushing a green button when presented with a same stimulus and a red button when presented with a different stimulus. Training consisted of blocks of 16 same stimuli and 16 different stimuli in a random order. An identical set of icons was used to form stimuli for both the same and different items, making it impossible to correctly associate an icon or icon feature with a response. Training continued until the pigeons reached 85% accuracy.

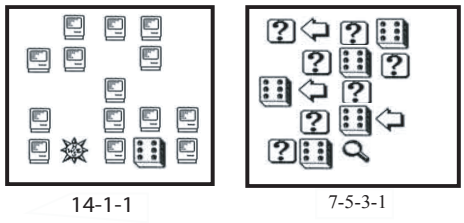

Figure 2.2: Two examples of intermediate stimuli are shown. The numerical code below each stimulus indicates its experimental condition and is explained in the main text.

The test phase consisted of intermediate stimuli that were somewhat

similar to both the same and different stimuli experienced in the training phase. Examples of intermediate stimuli are shown in Figure 2.2. These stimuli can be viewed as forming a continuum between the pure same stimuli (all 16 icons identical) and the pure different stimuli (all 16 icons different) used during the training phase. Eleven distinct conditions of intermediate stimuli were used. The 11 conditions can be characterized by their groupings of identical icons. For example, in Figure 2.2, the right most stimulus contains seven question marks, five dominoes, three arrows, and one magnifying glass and thus is an example of condition [7,5,3,1]. Adopting this nomenclature, the eleven intermediate conditions were [14,1,1], [8,8], [13,1,1,1], [12,1,1,1,1,1,1], [10,3,2,1], [7,5,3,1], [4,4,4,4], [8,1,1,1,1,1,1,1,1], [2,2,2,2,2,2,2,2], [4,1,1,1,1,1,1,1,1,1,1,1], [2,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1]. The pigeon's performance in these intermediate conditions, as well BRIDGES's predictions, are shown in Figure 2.3 (with the data points ordered left to right in the order the conditions are introduced in the previous sentence).

Table 2.3: The description of three icons from a *same* array, and three from a *different* array by BRIDGES

Entities	Features	Relations
$Rose_1$	$IconFeature(Rose_1) = Rose\ F.$	$TypeOf(Rose_1, A.Rose)$
$Rose_2$	$IconFeature(Rose_2) = Rose\ F.$	$TypeOf(Rose_2, A.Rose)$
$Rose_3$	$IconFeature(Rose_3) = Rose\ F.$	$TypeOf(Rose_3, A.Rose)$
$Robot_1$	$IconFeature(Robot_1) = Robot\ F.$	$TypeOf(Robot_1, A.Robot)$
$Sign_1$	$IconFeature(Sign_2) = Sign\ F.$	$TypeOf(Sign_1, A.Sign)$
$House_1$	$IconFeature(House_3) = House\ F.$	$TypeOf(House_1, A.House)$

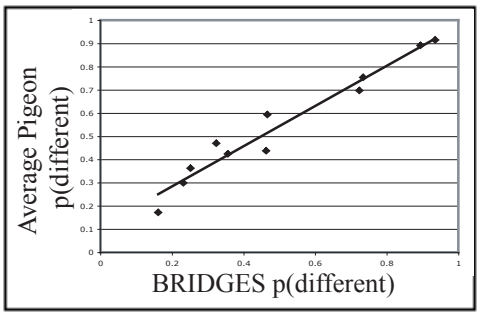

Figure 2.3: The results from Young and Wasserman's (1997) studies and BRIDGES's predictions are shown. The 11 intermediate conditions, forming a continuum between pure same and pure different stimuli, are described in the main text.

2.3.1 Method

2.3.1.1 Representation

This simulation used the same components as Simulation 1. Each array contained 16 entities with an associated single feature representing the objec; however, for these simulations no position feature was encoded as it was randomized in the original study and again would simply add noise to the learning process. Details specific to this study are presented in Table 2.3.

2.3.1.2 Fitting

The training regimen mimicked the procedure used in the original study as closely as possible. BRIDGES, like the pigeons, was trained to an 85% accuracy threshold before initiating the test phase. The reported numbers are from the parameters that yielded the best fit after training.

2.3.1.3 Simulation and Discussion

Strong support for BRIDGES's similarity-based discovery of the *same* and *different* relations is found in its fit of the test phase. BRIDGES correctly orders the intermediate conditions (see Figure 2.3). BRIDGES correctly predicts the probability that pigeons respond *different* in the intermediate conditions. Similarity-based activations are not all or none, and these intermediate cases activate stored exemplars to varying degrees, leading to the successful fit.

The fit arises because of the alignment process BRIDGES uses to determine similarity. In the model, a *same* stimulus aligns perfectly with *same* exemplars stored in memory. For example, consider aligning a stimulus containing 16 squares to another stimulus containing 16 triangles. Each triangle entity is put into correspondence with a square entity. This results in a perfect feature mismatch, but parallel connectivity is preserved. This alignment leads to attention shifting toward the relations and away from the mismatching entities. In contrast, only 1 out of the 16 relations will exhibit parallel connectivity when aligning a *different* stimulus with a *same* stimulus.

Looking at an example intermediate stimulus containing 12 triangles and four squares, the model considers it somewhat analogous to a stimulus containing 16 circles—mapping the triangle type to the circle type preserves parallel connectivity in 12 out of 16 relations. Along similar lines, an item with two matching icons and 14 icons that are all different from one another better matches a pure different exemplar than a pure same exemplar. Thus,

it is straightforward for BRIDGES to discriminate between *same* and *different* stimuli in the absence of featural support. In this regard, BRIDGES's solution for these simulations is the supervised learning analog to BRIDGE's discrimination of grammatical and ungrammatical sentences in the Marcus et al. (1999) simulations.

BRIDGES learned the same/different relation and achieved an excellent fit ($R^2 = .95$) of the test results involving intermediate stimuli. The simulations demonstrate how abstract concepts can be acquired through storage and analogy to concrete examples. BRIDGES's excellent fit of the intermediate conditions is a natural consequence of similarity-based processing. Like natural categories, BRIDGES suggests that relational categories have a graded structure.

2.4 Simulation 4 - Confounding Features in Relational Categories

Prototype and rule-based approaches to category learning represent categories as abstractions of their category members, devoid of their concrete features. In contrast, BRIDGES classifies stimuli according to their weighted similarity to exemplars of a category. The exemplars retain all of their concrete features. BRIDGES suggests that the features of the exemplars used during training will always play a role in the decision making process (although due to attentional shifting the size of this role could be indistinguishable from zero). Furthermore, BRIDGES's learning mechanism makes testable predic-

tions about the distribution and time-course of attention shifting, during and after the learning process, as well as predictions of how factors such as the category structure, stimulus distribution, and presentation order affect those shifts.

Several studies demonstrate that features play a role in the recognition of relations, showing a form of feature-based interference (Gentner & Toupin, 1986; Gomez, 2002; Huttenlocher, Duffy, & Levine, 2002; Richland et al., 2006; Sheya & Smith, 2006). Gibson and Wasserman (2004) provide an interesting example that directly manipulates the importance of features and relations on classification performance. The experiments are an extension of Young and Wasserman (1997), which is modeled above. In Young and Wasserman (1997) the authors showed that pigeons, with training, are capable of responding differentially based on icons in an array exhibiting either the *same* relation (all icons are identical), or *different* relation (all icons are different). Furthermore, the pigeons responded preferentially, on a graded basis, to mixed-arrays exhibiting a combination of unique icons and copies of icons depending on how what percentage of icons in the array were the same, or different.

In Gibson and Wasserman (2004), the researchers trained the birds on arrays where the features were confounded with the relations. Icons from set A were only found in arrays where all of the icons were identical (*same* arrays), whereas icons from set B were only found in arrays where all of the icons were different (*different* arrays). The pigeons were trained to differentiate *same* arrays, and *different* arrays. After the pigeons reached criterion on perfor-

mance, the pigeons were then tested with arrays using novel icons from group C. The pigeons were able to correctly categorize the novel arrays according the whether the constituent icons were all identical (*same*) or unique (*different*).

However, the researchers also tested the pigeons on arrays using icons from set A or set B. When the pigeons were tested on arrays where the items conflicted with the relation (Icons from set A in a *different* array), the pigeons would respond based on the icon and not the relationship between the icons. The pigeons' behavior properly reflected the relational information, except when it was put at odds with the feature information, in which case they responded based on the features.

Initially, a reader might be skeptical of whether the pigeons are truly learning relations. In fact, an alternative, entropy detection model, was put forward as an explanation of the pigeons' performance in (Young & Wasserman, 1997; Young, Ellefson, & Wasserman, 2003). The entropy explanation says that the pigeons are detecting the amount of visual entropy present in the system; the more variability that exists in the display icons, then the higher the entropy (Young & Wasserman, 1997). This explanation turned into a formal model of the 1997 finding, the Finding Differences Model, that fits the data only slightly better than the BRIDGES model, $R^2 = .98$ compared to $R^2 = .95$. However, Gibson and Wasserman (2004) provides evidence against an entropy based explanation. Entropy is a measure of variability, and two different displays containing the same number of matching icons should have the same entropy regardless of the icons used. Unlike entropy, BRIDGES stip-

ulates that performance on the test trials is a function of both feature- and relation-based similarity.

Simulation 4 illustrates the trade-off between featural similarity and relational similarity and how the two different types of similarity affect the way items are categorized. As in the simulation of Gerken (2004) attention shifting is carried out to maximize the models accuracy in determining the category label; abstraction is only as far as is dictated by the environment.

2.4.1 Method

2.4.1.1 Representation

This simulation used the same components as the previous simulation (see Table 2.3)

2.4.1.2 Fitting

The training regimen followed the procedure used in the original study as closely as possible. BRIDGES, like the pigeons, was trained to an 85% accuracy threshold across both same and different distinctions before initiating the test phase. Two sets of results are presented, the first set was generated by finding the parameters the yielded the best fit to the data, while the second set utilized the same parameters as from the above simulation of Young and Wasserman (1997).

2.4.2 Simulation and Discussion

As is shown in the first and second columns of Table 2.4 the model is able to provide a reasonable fit to the data obtained in the original study. BRIDGES is able to correctly categorize the training instances, as well as the test arrays which use novel icons. Also like the pigeons, BRIDGES responds incorrectly when faced with the test arrays which cross the icon set with the category label. When faced with the icons used for training the *different* arrays the pigeons, and the model, have a propensity to respond *different*, even when the array contains all of the same icon.

Table 2.4: Showing the pigeons and BRIDGES chance of correctly identifying the relation depending on the stimulus set. Results are provided for BRIDGES for both the parameters used in modeling Young and Wasserman (1997), and the set that gave the best overall fit

	Pigeons'	BRIDGES	
Stimulus Set	Results	Best Fit	Y&W(1997)
Training	.87	.90	.90
Novel	.66	.70	.51
Crossed	.28	.40	.49

The model is responding to both the relations and the features, and since the features are perfectly confounded with the relations in their predictive ability of the category label, attention shifts equally to both attributes. In (Young & Wasserman, 1997), the only correlation between the arrays and the labels is the relations, which causes a large shift in attention away from the non-predictive features and positions of the icons to the relations. Yet, in Gibson and Wasserman (2004), both the features of the icons and the relationships

between the icons are correlated with the category label. This linking causes attention shifts to both the features of the icons and the relations. Thus, in Gibson and Wasserman (2004) when the pigeons were tested with the features and relations put at odds with one another, the features won out as there was more attention on those components to begin with.

2.5 Simulation 5 - Re-representation and progressive alignment

The previous simulations show how BRIDGES can learn relation-based categories and predict a learner's performance on a transfer task. The modeled experiments show that a learner's category representation can be influenced by both the feature- and relation-based regularities present in the environment. Further, we showed that BRIDGES is able to capture this effect through its attention shifting mechanism. BRIDGES is able to shift attention between relations and features depending on which is predictive of the category label. This ability allows it to successfully learn relation-based categories as well as account for feature-based interference in those learned categories. This collection of results presented evidence favoring a similarity and attention-based account of relational categories over a rule-based account. In Simulation 5 we look beyond issues of classification to see how similarity-based category learning can clarify some of the processes that analogy researchers have identified as occurring during relation-based learning.

Re-representation and progressive alignment are two ideas of learning

considered in the analogy literature. Re-representation is the idea that as new knowledge is acquired, the structure of older knowledge is changed. Re-representing concepts can allow one to draw more accurate or easier analogies between domains, and is often cited as a major component of advancement in science(Gentner et al., 1997). However, it is also often considered to occur in simple laboratory based experiments that take a matter of minutes (Doumas et al., 2008; Kotovsky & Gentner, 1996). Attention-driven analogies suggest an alternative view of re-representation. Just like abstract relation-based responding can happen without rules, similarity-based relational abstraction can produce results almost indistinguishable from re-representation, without any change in the underlying representation. This suggests that these two types of learning, conceptual change through analogy and long bouts of careful consideration, and behavioral changes through quick attention shifting as a result of stimulus-response learning, could be different processes.

Progressive alignment is a theory of how complex abstractions can be formed from simple ones. It suggests that learners bootstrap to more complex representations through a series of progressively more complex representations formed by comparing gradually more disparate stimuli that are only similar at progressively higher levels of abstraction. BRIDGES provides a way to quantify the effect of similarity on the comparison process.

Kotovsky and Gentner (1996) describe a set of experiments which show both re-representation and progressive alignment through match to sample tasks. In each trial of the experiments a child is asked which of two test scenes

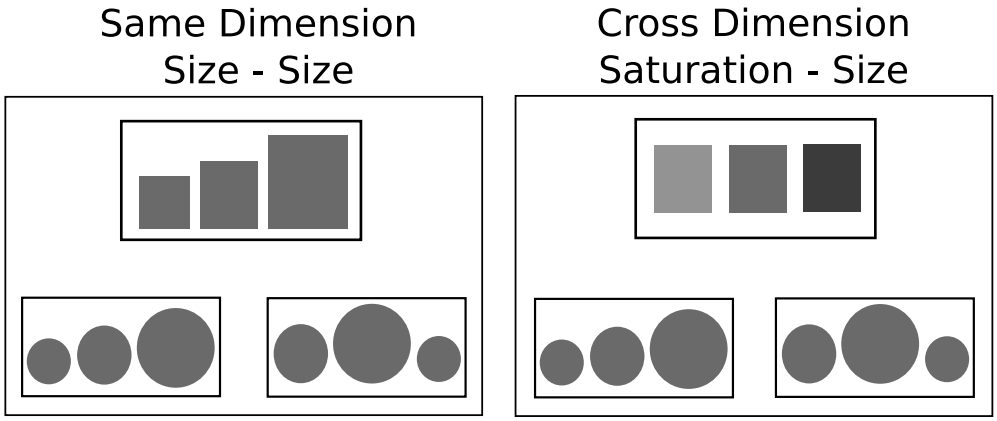

Figure 2.4: Example stimuli from Kotovsky and Gentner (1996). The left hand triad shows a same-dimension match (size) and the right hand a cross-dimension match (saturation→size)

is more like a standard. Each standard shows a group of three objects that exhibit one of two different second-order relations between them: a monotonic change, or symmetry (i.e. two identical objects, flanking a central object) over a size dimension or a hue dimension. The comparison scenes are made up of a relational-choice and a non-relational match that does not exhibit the same relation as the standard. Two different relational matches were examined, in one the relational choice exhibited the same high-order relation over the same dimension (i.e. size-size, same-dimension trials), and in the second the relational choice exhibited the same high-order relation over the opposite dimension (i.e. size-saturation, cross-dimension trials, see Figure 2.4).

While older children (and adults) will always select the relational match, whether same-dimension or cross-dimension, young children (4-year-olds) only select the relational choice above chance levels in the same-dimension task.

When the children are presented with triads where the standard exhibits a different low order relation (size) compared to the relational choice (hue), they select at random This pattern of results shows a clear difference between the two age groups. Younger children do not represent the similarity between the cross-dimension relational choices in the same manner as older children and adults. Kotovsky and Gentner's (1996) theory embodies the idea that re-representation is a critical part of the relational shift. Because the older children have already experienced the re-representation, they have shifted to a more relational view of the stimuli.

Table 2.5: BRIDGES fit of the children's responses across age and task orders.

Experiment	Children	BRIDGES
Experiment 1		
4-year-olds same dimension	.68	.66
4-year-olds different dimension	.49	.49
6-year-olds same dimension	.90	.90
6-year-olds different dimension	.75	.75
Experiment 2		
Size-only Control	.50	.45
Progressive Alignment	.80	.79

Interestingly, Kotovsky and Gentner (1996) found that some young children would select the relational-choice at above chance levels on the cross-dimension triads if the trials were presented in a special order. If the experiment started with the same-dimension trials where the standard was monotonic size, for example, then followed with trials over the saturation dimension, the 4-year-olds then went on to perform relationally on cross-dimension trials

(see 2.6 for actual trials, children were presented with two of each trial type). The authors posit that this is because 4-year-olds start with a domain-specific representation of the increasing size relation (increasing is tied to changes in size) and are able to re-represent the stimuli in terms of the higher-order relation, through progressive alignment during learning.

In contrast to the authors theory that children undergo a fundamental change in their representation of the stimuli, the theory behind BRIDGES suggests that the children do not necessarily experience a re-representation of the stimuli. Instead the model assumes that the older children start out with more attention focused on the relations, and less on the features. while the younger children are learning to shift their attention away from the concrete features of the objects to the relational structure between the objects. That is, children learn to focus on maintaining parallel connectivity when aligning the two sets of stimuli regardless of the featural distractions. This process is supported by the presentation of multiple challenging stimuli that the children can succeed with. The combination of attention shifting and exemplar comparison allows for the emergence of more complicated behavior than is suggested by the representation alone.

2.5.1 Method

2.5.1.1 Representation

BRIDGES simulation of Kotovsky and Gentner (1996) used a more complex representation than that in the previous simulations in this paper. In

Table 2.6: Order of triads for progressive alignment.

Dimension	Dimension of Standard	High-Order Relation
Experiment 2		
same	size	monotonic-increase
same	size	symmetry
same	color	symmetry
same	color	monotonic-increase
cross	size	symmetry
cross	color	symmetry
cross	size	monotonic-increase
cross	color	monotonic-increase

addition to the type-token relations used consistently throughout, each entity (square or circle) had an associated feature describing its size, hue, shape, and position. Furthermore, relations between the objects were encoded, such as bigger, and brighter. These relations were domain-specific, in that bigger did not match brighter.

2.5.1.2 Fitting

The match to sample task used in the original experiments required a modification to the architecture and to the learning rule. On each trial, the standard (i.e. the exemplar at the top of the display, see Figure 2.4) was assumed to be the input and the two choices were taken as the exemplars. The model preferentially chose the item that was most similar to the standard and adjusted its weights to reinforce the choice, according to Equation 1.10 from

the Introduction.

Informally, the model shifts its attention to the features or relations that match between the first item (A) and the input and mismatch between the second item (B) and the input, those parts of the structure that discriminate the two mappings. The size of this attention shift is a function of the relative similarity of the stimuli and the standard, large steps will be taken if both stimuli are highly similar to the standard (as more differentiation needs to be made), and small steps will be taken if they are disparate in their similarity to the standard. The direction of the shift changes depending on which stimulus the model selects as matching the standard; it reinforces that selection.

In Experiment 1 of Kotovsky and Gentner (1996) the stimuli are presented in a random order to children of different age groups. Like the children, the model received same-dimension size, same-dimension color, and cross-dimension triads in a random order. In order to represent the differences between the two groups, an "older" model was created from the results of the simulation of Experiment 2 that had more initial attention on the relations to represent an increased knowledge of the role of relations in the world.

Experiment 2 of Kotovsky and Gentner (1996) tested the progressive alignment theory. The stimuli were presented in a fixed order. The control condition presented 8 same-dimension triads (all color or size followed by the cross-dimensions triads), while the experimental condition presented 8 mixed same-dimension triads followed by the cross-dimension triads (see Table 2.6). In order to fit the results from Experiment 2, the relative attention to the size

and saturation relations, and the size and saturation features was adjusted. This is consistent with the idea that young children have more attention on features, than on relations.

2.5.2 Results and Discussion

Table 2.5 shows the fit of BRIDGES to the experimental results of Kotovsky and Gentner (1996). BRIDGES captured the full gamut of the responses. The model representing the youngest children selected the relational-choice on the same-dimension triads, but selected at chance on cross-dimension trials. In contrast, the model representing the oldest children always selected the relational choice because of its increased attention to the relational structure. As discussed in the methods section, the difference between the age groups is in the initial starting weight assigned to the relation condition. We reasoned that the older children place more value on relational information, or can more easily represent the relations, and therefore, are more likely to respond in accord with the relations. This approach is different from that proffered in (Richland et al., 2006), in that they proposed that children need a more mature inhibitory system to disregard the feature matches. BRIDGES simulation by adjusting the attention to relations suggests that learned attention to relations could be an important part of the relational shift.

An alternative approach would have been to change the representation used by the model for the two different age groups by adding a domain-general monotonic relation. This would have the same effect. A domain-general mono-

tonic relation would match the other domain general monotonic relation, and be less similar to the non-relational choice. This would increase the disparity in the similarity between the two standards and the relations choice. We remain agnostic as to which is more correct, but feel that this helps to demonstrate that changes in attention and changes in representation are sometimes not easily distinguishable.

Table 2.5 also shows that Bridges captured the second experiment on progressive alignment and re-rerepresentation. Recall that the critical manipulation in the second experiment was a change in whether the children received all size same-dimension triads, or a mixture of size and saturation triads. In the progressive alignment condition, those 4-year-olds that succeed on the same-dimension triads go on to succeed on the cross-dimension triads. In the control condition, the children the 4-year-olds do not succeed in reliably selecting the relational match. BRIDGES demonstrates the same pattern of results as the children. The model in the progressive alignment condition continues to do well if it did well on the initial trials, whereas the model is at chance in the size-only control condition.

The success of the model in selecting the cross-dimensional relational choice in the progressive alignment condition is due to its exposure to both features before being required to select in the cross-dimensional task. The training allows the model to shift attention to the size and saturation relations. In contrast in the control condition, the model is only exposed to one feature type; it does successfully increase its attention to the size relation during

presentation of the same-dimension triads. However, when faced with the cross-dimension triads the saturation relations do not have enough attention to provide sufficient support for the relational choice. This results in the relational and non-relational choice being confusable because both are highly dissimilar to the standard.

In both sets of experiments, the key to BRIDGES fit is shifting attention to what makes the relational-choice and the standard more similar, e.g. the relations. When the exemplars are distinguishable and the most similar exemplar reinforces the relational choice, BRIDGES will shift attention to the discriminating relations. Importantly, the relational choice must be similar enough to the standard that it will be chosen by the model, but not so similar that there is no error signal. If the relational choice is too dissimilar from the standard BRIDGES will probably not find a good alignment and could easily select the non-relational choice reinforcing the wrong relations or the features. By showing the model simple examples of a variety of instantiations of the relation, it learns to disregard the features and concentrate on the relations.

The fits by BRIDGES show that is is possible to succeed in this task without re-representing the stimuli. As discussed in Simulation 2, learners prefer the least general level of abstraction that is able to still predict the data. The infants in Gerken (2006) did not extend an AAB relation to the stimuli when they only followed an AA'di' pattern. Similarly, it is unlikely that the children develop a domain-general abstract representation of both high order symmetry and monotonic change over the course of only 4 trials

when it is unnecessary to perform the task.

One caveat is that it is possible that the 4-year-old children already had a representation of a domain-general symmetry and monotonic-change relation, only it had so little attention that it did not provide the necessary support for the correct mapping. In this case, during learning the children would shift attention to that relation in addition to the others and could presumably then succeed with even more featurally disparate stimuli than would be predicted if they did not have this relation.

Chapter 3

Experimental support for an interaction between comparison and category structure

The previous simulations support the idea that relation-based categories may be learned in the same manner as feature-based categories (through relational similarity to stored exemplars), which is consistent with the notion of family resemblance. Simulation 2 suggests that a similarity-based account is more parsimonious in regards to questions of the degree of generalizations as many rule-based models do not well specify when a rule should be formed and to what degree of abstraction (Gerken, 2006). Experiment 3 offers an example that cannot be fit by a rule-based model of learning because responses to the exemplars are affected by the typicality of the exemplar. Simulation 4 demonstrates the advantages of an attention-based account to learning because it correctly accounts for the interference of alternative predictors. Lastly, simulation 5 suggests that re-representation, the commonly used learning approach discussed in the analogy literature and progressive alignment, may be re-construed as a result of attention shifting and similarity comparisons. These findings present a coherent story of the role of similarity-based learning in relational tasks. Below, we look at how relational comparisons affect similarity-based learning.

Lassaline and Murphy (1998) provided an initial test of the role of alignment in category learning. They found that the number of Matches In Place (MIPS, e.g. two birds having the same color head) and Matches Out of Place (MOPS, e.g. a bird having the same color tail as another bird's head) relate to the difficulty of learning a category. Categories are easier to acquire when members have many MIPS in common compared to when they have many MOPS in common. Lassaline and Murphy's results provide support for the notion that dimensional correspondences (Goldstone, 1994) are critical in category learning. However, their results do not speak to the role of relational commonalities (i.e., relationships across dimension values). For example, two birds could share no MIPS or MOPS, yet share the relational commonality of having their heads brighter than their tails.

Below, we consider how people learn categories defined by such relational matches. We suggest that people use relational comparison processes (such as alignment) during category learning to match the current stimulus to stored category examples. Such processes may explain people's ability to readily learn purely relational categories from a small set of examples (e.g., Rehder & Ross, 2001. As evidenced above, computational models that incorporate these processes successfully explain how infants, adults, and animals learn seemingly abstract concepts based on a small set of training examples.

A number of predictions fall out of the view that category learning involves online relational comparisons to stored examples. One prediction that we test here is that online relational processing can benefit or hinder

learning depending on the relationship between presented stimuli and previous exemplars. When the preferred mapping between the current stimulus and stored stimuli serves to increase the coherence of a category, learners should benefit from relational processing. However, in other cases, relational matching processes can actually reduce category coherence by enabling discovery of non-obvious similarity relations among members of contrasting categories. In such cases, learning should be retarded and error patterns indicative of interference arising from relational comparison processes should be observed.

3.0.3 Stimulus Design

The stimuli used in our experiments consist of simple scenes that were designed to provide an informative comparison of relation- and feature-based category learning. To this end, both categories were defined over, and required processing, the same perceptual attributes (i.e. size, luminance) and the same number of perceptual attributes (two attributes).

Because features and relations are psychologically distinct and are associated with different processes (see Goldstone et al., 1991), in principle, one cannot construct a stimulus set that does not bias in favor of featural or relational learning. Relations fundamentally are defined over distinct entities, whereas features integrate over some range. For example, determining the overall luminance (a feature) of a scene requires integrating over the entire scene. It stands to reason that complex scenes consisting of multiple entities would make such a computation more difficult. Indeed, Duncan (1984) finds

that specifying features across multiple objects increases the time required to identify the features. Conversely, relational regularities are likely to be easier to detect when entities are readily individuated. To appreciate A causes B, one must be able to clearly discern A and B as distinct entities. In summary, stimuli that contain a single entity are likely to favor featural processing, whereas stimuli that contain multiple entities are likely to favor relational processing.

The norm in the field (e.g., Lassaline & Murphy, 1998) is for each stimulus to consist of a single entity, thus favoring featural processing. Because our primary interest is in exploring relational influences on category learning, we used a stimulus set consisting of scenes composed of two entities. Although it is not our focus, one interesting question is whether relational category learning is actually favored under such conditions compared to featural category learning. To foreshadow, our results indicate that it is. Although not discussed in this contribution, we find that same overall pattern of results, with a bias in favor of featural learning, when the present studies are conducted using stimuli that consist of a single entity.

In the studies reported here, each stimulus consisted of two circles appearing side-by-side. Across trials, these two circles varied in their size (small, medium, large) and brightness (light, moderate, dark). These circles were combined to give two overall relation attributes (which side was bigger and which was brighter) and two feature attributes (overall size and overall brightness). The medium and moderate values were always manifested once in a scene (See Figure 3.1).

In summary, in order to balance the perceptual requirements the relational attributes defined for these scenes, which side is bigger and which side is brighter, require the processing of the relationship between the two objects in the scene to correctly identify the attribute. In contrast, the evaluation of featural attributes does not require any consideration of the relational role a specific circle plays in determining an attribute's value. The features and relations differ only in how the participant combines the information about the two circles. For the relations, consistent with definitions of relational categories (Markman & Silwell, 2001), information from each circle plays a distinct role in determining the value of each relation, whereas for the features, information from each circle is combined independently in determining the value of each feature. Additionally, the requirement that the features are separated across the two circles should bias learning rates towards the relations because features become more difficult to process when spread across multiple objects (Duncan, 1984).

To foreshadow our results, in Experiment 1, we find that simple relational categories are learned faster than simple feature-based categories. Although this result may be surprising on the surface, it follows from our stimulus design (two entities as opposed to one, same perceptual substrate for featural and relational information) and previous findings in relative vs. absolute stimulus judgments. Using more complex category structures, Experiments 2 and 3 consider conditions under which online relational comparisons should promote or hinder learning. In Experiment 2, we find that relation-based category

learning is advantaged over feature-based category learning under conditions in which relational comparisons allow learners to increase category coherency. In Experiment 3, using category structures in which relational comparisons should not increase coherency, we find no advantage for learning relation-based categories over feature-based categories. Indeed, we observe error patterns indicative of comparison processes interfering with relation-based category learning. In Experiment 4, pairwise similarity ratings are collected for the stimuli used in Experiments 1-3. The similarity rating data were non-metrical in a manner consistent with online comparison processes and provide additional support for our interpretation of Experiment 1-3's results. While Experiments 1-4 support the notion that relational comparison can shape learning, Experiment 5 finds support for the complementary position that learning can shape relational comparison.

3.1 Experiment 1

Experiment 1 examines participants' ability to learn simple category structures defined by a single attribute. Participants learned to classify stimuli as members of one of two contrasting categories based on a single featural attribute (e.g., items with big circles are in one category, whereas items with small circles are in the other category), or learned to classify based on a single relational attribute (e.g., items in which circle on the right is bigger are in one category, whereas items in which the circle on the left is bigger are in the other category). Based on the ease with which people make relative judgments and

the increased difficulty of making feature judgments across multiple objects, we predict that participants will learn the relation-based categories faster than the feature-based categories.

3.1.1 Methods

3.1.1.1 Participants

Fifty-three undergraduate students from the University of Texas at Austin participated for course credit. Participants were randomly assigned to the brightness-relation relevant (n=13), the size-relation relevant (n=13), the brightness-feature relevant (n=13), or the size-feature relevant conditions (n=14).

3.1.1.2 Stimuli

The stimuli were the same as those described above. They varied along 4 binary attributes: 2 relational attributes, which side was bigger and which side was brighter; and 2 featural attributes, overall brightness (both circles combined) and overall size (both circles combined).

3.1.1.3 Procedure

Participants were presented with a screen of detailed instructions informing them that they were going to learn to categorize pairs of circles into two categories, A and B. Participants were instructed that each stimulus varied along 4 attributes: overall brightness, overall size, which circle was brighter, and which circle was bigger. They were told to look for a rule involving one

of those attributes. For each participant, the labels A and B were randomly assigned to the two categories.

On each learning trial, two circles were presented in the center of the computer screen. The stimulus was accompanied by the text prompt "Category A or B?". Participants freely responded with an A or B key press and immediately received either a brief low (wrong) or high (right) pitched auditory tone concurrent with text containing "WRONG" or "RIGHT" and the correct category label for the stimulus. The correct category label and the stimulus were presented for 1250 ms followed by a blank screen. After 500 ms, the next trial began.

The trials were blocked in groups of 16. Each block consisted of a random ordering of the 16 stimuli. Participants were not made aware of transition between blocks. Category training terminated when participants reached a learning criterion of correctly classifying 12 stimuli in a row or completed 18 blocks (288 trials) without reaching the criterion.

3.1.1.4 Results and Discussion

The proportion of trials correct for each participant was calculated. Remaining trials for participants reaching the learning criterion were scored as correct. Statistical tests found no significant differences between size and brightness for learning feature- or relation-based categories. Therefore, analyses collapse across size and brightness sub-conditions and focus on the distinction between feature- and relation-relevant category learning.

The results are displayed in Table 3.2. As predicted, participants in a relation-based category were significantly more accurate (.95 vs. .76) than those in a feature-based condition, $t(51) = 4.56, p < .001$. All 26 participants in the relation-based condition reached the criterion, while only 21 of the 27 participants in the feature condition did so; this difference is significant, $\chi^2(1, N = 53) = 4.49, p < .05$. These results confirm the hypothesis that relation-based categories are easier to learn than feature-based categories when the perceptual attributes are spread across multiple stimuli.

Table 3.1: Category Structures

Attr. 1	Attr. 2	Attr. 3	Attr. 4	XOR	Four-Category
0	0	0 *or* 1	0 *or* 1	A	A
0	1	0 *or* 1	0 *or* 1	B	B
1	0	0 *or* 1	0 *or* 1	B	C
1	1	0 *or* 1	0 *or* 1	A	D

Table 3.2: Summary of Results from Experiments 1, 2 and 3

		Relations Relevant		Features Relevant		Accuracy
	Rule	Accuracy	Criterion	Accuracy	Criterion	Differences
$Exp.1$	$1 - D$	.95	26/26	.76	21/27	$.19 * **$
$Exp.2$	XOR	.73	14/25	.54	4/27	$.19 * **$
$Exp.3$	$Four$	.78	22/26	.74	22/27	.04

3.2 Experiment 2 - Learning a relation- or feature-based XOR

In Experiment 1, categories defined by a single relational attribute were learned faster than categories defined by a single featural attribute. Experi-

ment 2 employs a nonlinear category structure in which items that are opposite in every respect are members of the same category. We hypothesize that the flexibility afforded by relational processes will enable learners to regularize relation-based category structures and represent the differences between the stimuli in a way that facilitates learning of the category. This flexibility, combined with advantages for relative judgments, should result in categories defined by relations being acquired more readily than comparable categories defined by features. This avenue for boosting coherency should not be available to learners of feature-based categories. Finally, overall performance should be lower in Experiment 2 than in Experiment 1 because the categories in Experiment 2 are defined by two attributes, whereas categories in Experiment 1 are defined by a single attribute.

3.2.1 Methods

3.2.1.1 Participants

Fifty-two undergraduate students from the University of Texas at Austin participated for course credit. Participants were randomly assigned to the feature- (n=27) or relation-relevant (n=25) condition.

3.2.1.2 Stimuli and Category Structure

. The stimuli were the same used in Experiment 1: pairs of circles varying along two relational and featural attributes. Two of the four stimulus attributes were relevant to determining category membership. An exclusive

disjunction (XOR) rule involving the two relevant stimulus attributes defined the category structures. XOR is a nonlinear classification rule that requires attention to both of the relevant attributes (see Table 3.1). Stimuli that are opposite one another on both relevant attributes are placed in the same category. In the relation condition, the two relation-based attributes were relevant and the features were irrelevant. The opposite was true for the feature condition.

3.2.1.3 Procedure

Training followed the same pattern as in Experiment 1.

3.2.2 Results and Discussion

The proportion of trials correct for each participant was calculated. Remaining trials for participants reaching the learning criterion were scored as correct. The results are summarized in Table 3.2. As expected, inspection of Table 3.2 reveals that Experiment 2's more complex (defined by two attributes) category structures were more difficult to acquire than Experiment 1's simple (defined by one attribute) category structures.

Participants were significantly more accurate (.73 vs. .54) in the relation-relevant than in feature-relevant condition, $t(50) = 5.19, p < .001$. A significantly greater proportion of participants (14/25 vs. 4/27) reached the learning criterion in the relation-relevant condition than in the feature-relevant condition, $\chi^2(1, N = 52) = 8.00, p < .01$. Experiment 2's results demonstrate that more complex categories defined by the relations are easier to learn than

categories defined by features.

3.3 Experiment 2 - Learning a relation- or feature-based four category structure

The results from Experiment 1 supports the idea that relation-based categories may be easier to learn than feature-based ones when the categories are balanced on the number of perceptual attributes and those attributes are spread across multiple objects. Experiment 2 further suggests that the flexibility afforded by online, relational comparisons benefit relational learners, particularly those who acquired the complex XOR category structure used in Experiment 2. Experiment 3 tests this hypothesis more fully by attempting to match the speed of relation-based learning and feature-based learning through utilizing a category structure that is not amenable to relational alignment.

In Experiment 2's relation-relevant condition, the online nature of relational comparisons enables the learner to establish stimulus comparisons that reduce the difficulty of the learning task by increasing within-category member similarity. In the XOR category structure used in Experiment 2, the most dissimilar items, those that differ on both relations or both features (See Figure 3.1), are placed within the same category. However, it is easy to see that the initial application of a swapping transformation, or by mapping the stimuli based on role instead of position, during the comparison process makes the top and bottom stimuli identical. This means that the within-category similarity for the relational XOR is probably much higher than it is for the feature-based

equivalent, and this should speed learning (Lassaline & Murphy, 1998; Rosch & Mervis, 1975). Swapping transformations and cross-mapping operations are not available for the feature-based categories.

Experiment 3 tests this online comparison account by training participants on a category structure in which such processes should not be advantageous to relational learners. Compared to Experiments 1 and 2, differences in performance between featural and relational learners are predicted to compress under these conditions. The category structure used in Experiment 3 is the four-category structure specified in Table 3.1. Unlike the XOR category rule used in Experiment 2, in the four-category structure items that differ on both relevant attributes are members of different categories.

Unlike Experiment 2, regularizing relational differences in Experiment 3 through online relational processes will not increase within category similarity because stimuli that differ on both relations are now in separate categories. As a consequence of this, relational learners engaging in such relational processes may in fact increase confusions between categories that differ on both relations, leading to opposite classification errors (e.g., confusing members of categories A and D, or B and C in Table 3.1).

While it is difficult to make cross-experimental comparisons between the three experiments because they contain a different number of categories, the key predictions for Experiment 3 are that the difficulty of featural and relational learning should converge. Instead of boosting performance as in Experiment 2, online relational comparison in Experiment 3 should manifest

itself in more errors to the opposite category for relational learners.

3.3.1 Methods

3.3.1.1 Participants

Fifty-three undergraduate students from the University of Texas at Austin participated for course credit. Participants were randomly assigned to the relation- (n=26) or feature-relevant (n=27) condition.

3.3.1.2 Stimuli and Category Structure

The stimuli were the same as those used in Experiments 1 and 2. Participants learned to classify each stimulus as a member of 1 of 4 different categories. The categories were the 4 unique combinations of the 2 values of the 2 relation attributes in the relation condition and the 2 feature attributes in the feature condition (see Table 3.1).

3.3.1.3 Procedure

The procedure was identical to that used in Experiments 1 and 2, except that participants had to learn to classify the circles as belonging to 1 of 4 categories (A, B, C, or D) by pressing the corresponding key.

3.3.2 Results and Discussion

The proportion of trials correct for each participant was calculated. Remaining trials for participants reaching the learning criterion were scored as correct. The results are summarized in Table 3.2. Accuracies were compa-

rable (.78 vs. .74) in the relation- and feature-relevant conditions, $t < 1$. A comparable proportion of participants (22/26 vs. 22/27) reached the learning criterion in the relation- and feature-relevant conditions, $\chi^2(1, N = 53) = .0039, p = .95$.

The pattern of participants' errors was also analyzed. Each incorrect response was classified as either a mistake to an adjacent category (e.g., $A \rightarrow B$ or C) or as a mistake to the opposite category (e.g., $A \rightarrow D$). As predicted, learners in the relation condition made proportionally more errors (34% vs. 27%) to the opposite category than did learners in the feature condition, $\chi^2(1, N = 3636) = 23.96, p < .001$. Relational learners who tended to make a higher proportion of opposite category errors relative to adjacent category errors had lower overall accuracy levels, $R^2 = .30, F(1, 24) = 10.49, p < .01$, whereas no such relationship held for feature learners, $R^2 = 0.00$.

As predicted, the relational advantage observed in Experiment 2 was not observed in Experiment 3. When a category structure is used in which online comparison processes are not beneficial to learning relation-based categories, featural and relational learning are of equal difficulty. Indeed, online comparison processes were manifested as a greater proportion of opposite category errors in the relation relevant condition and relational learners who showed stronger markers of online comparison processes were less accurate overall.

3.4 Experiment 4

A number of our claims in regards to online comparison processes center on how participants align stimulus elements to maximize perceived similarity. Such operations can boost (e.g., Experiment 2) or reduce (e.g., Experiment 3) category coherence. We proposed that the effects of these operations can be indirectly observed in category learning performance.

In Experiment 4, we directly investigate the nature of the comparison process by looking at similarity rating data for the stimuli. According to relational accounts of processing, the stimuli should exhibit the non-metrical effects of an online comparison process. In contrast to metrical views of comparison and similarity, where distances between stimuli follow a set of strict axioms and similarity decreases with each difference between the stimuli, transformation- and alignment-based approaches suggest that stimuli that differ along both relations might be rated as highly similar. These non-metrical effects could help explain the relational advantage observed in Experiment 2 and the error patterns in Experiment 3.

To test this explanation for the relation-based category advantage observed in Experiment 2 and the error pattern found in Experiment 3, similarity ratings for all stimulus pairs were collected in Experiment 4. We predict that ratings will be non-metrical, in that similarity will not substantially decrease (and may in fact increase) for stimulus pairs mismatching on both relations compared to stimulus pairs mismatching on only one relation.

Additionally, overall influences of featural and relational attribute matches can be assessed in the similarity rating data. We proposed that the advantage observed for relation-based categories in Experiment 1 primarily arose from the ease of relative judgments compared to absolute judgments. An alternative explanation is that relational attributes were somehow more salient. Experiment 4's similarity ratings allow for assessment of attribute salience and remove the memory retrieval processing component present in Experiments 1-3, which might have favored relative over absolute stimulus decisions.

3.4.1 Methods

3.4.1.1 Participants

Twenty-two undergraduate students from the University of Texas at Austin participated for course credit.

3.4.1.2 Stimuli

The stimuli were the same as those used in Experiments 1-3.

3.4.1.3 Procedure

Participants were instructed to rate the similarity of two presented stimuli on a scale from 1-9. As in Experiments 1-3, participants were instructed that each stimulus varied along 4 binary-valued attributes (overall brightness, overall size, which circle was brighter, which circle was bigger). On each trial, two stimuli were simultaneously presented on screen with text designating pair 1 and pair 2, as well as text asking for their similarity on a scale of 1-9. One

pair was displayed on the top of the screen and the other on the bottom. A line separated the pairs. Participants responded by pressing key 1 through 9. Following the participant's response, the screen blanked for 500 ms and the next trial began. Each participant rated 136 pairs of stimuli [(16 * 15)/ 2 + 16]: each stimulus paired with every other stimulus, plus each stimulus paired with itself. The overall order of the trials and the assignment of pairs to the top or bottom of the screen were randomized.

3.4.2 Results and Discussion

For the purposes of analyses, the similarity ratings were grouped according to how many features or relations were different within the comparison. Figure 3.2 illustrates the nine means resulting from this aggregation. A 3 (0, 1, or 2 relation differences) X 3 (0, 1, or 2 feature differences) within-participant ANOVA revealed a main effect of both the number of different relations, $F(2, 42) = 33.68, p < .001$, and the number of different features, $F(2, 42) = 204.81, p < .001$, as well as a significant interaction between the number of feature and relational differences, $F(4, 84) = 36.47, p < .001$.

The above interaction is indicative of a non-metrical similarity space arising from relational processes. To test the predictions of the alignment account more precisely, a 2 (relation or feature) X 2 (one or two differences) ANOVA was performed to compare the effects of mismatching on one or both relations (with both features matching) with the effects of mismatching on one or both features (with both relations matching). The strong interaction

predicted is shown in Figure 3.3. An ANOVA revealed a significant main effect for feature or relation difference, $F(1, 21) = 26.32, p < .001$, and a main effect for the number of differences, $F(1, 21) = 23.85, p < .001$, as well as a significant interaction between the type of difference and the number of differences, $F(1, 21) = 96.05, p < .001$. Planned t-tests revealed that rated similarity was higher when both relations differed than when only one relation differed, $t(21) = 2.13, p < .05$, whereas rated similarity was lower when both features differed than when only one feature differed, $t(21) = 13.68, p < .001$. As predicted by the online relational processing hypothesis, similarity ratings were non-metrical in that stimulus pairs differing on both relations were rated as more similar than stimulus pairs differing on only one relation. These similarity data are consistent with the learning results from Experiments 1 and 2.

To test for differences in salience between the relations and the features, a regression model was fit to each participant's similarity ratings, with the number of relational differences and the number of featural differences as independent predictors. A paired t-test was then conducted on the fitted weights for the relational and featural terms across the participants. This test showed a significantly larger effect for feature differences on rated similarity, mean coefficient of 1.68, compared to relational differences, mean of .77, $t(21) = 5.01, p < .001$. This test suggests that the advantage for relational learning observed in Experiment 1 was not due to the relations being more salient.

3.5 Evidence for learning-induced alignment

The previous studies support the notion that relational comparison processes can affect category learning by increasing (Experiment 1) or decreasing (Experiment 2) category coherency. We hypothesized that these effects arise because 2 stimuli differing in both relational attributes can be made more similar by a relational alignment that puts circles in correspondence that differ in left/right spatial position. In Experiment 3, the similarity rating data supported these conclusions. The experiments focused on directly testing the assumption that relation-based similarity affected category learning.

An attention-based theory of relational category learning also predicts the opposite direction of effect, that the category structure can affect the mapping. This is a fundamental assumption underlying BRIDGES. The category structure determines which dimensions are attended, those that minimize categorization errors. Because we posit that the utilized mapping is the one that maximizes similarity between the exemplars, different mappings should be chosen depending on the learned category structure.

3.6 Experiment 5

To investigate if learning a particular category structure can affect the way in which learners map the stimuli, we trained participants on either a feature-based or relation-based categories with an XOR structure (as in Experiment 1). Following training, participants compared two stimuli and reported which circles corresponded across the two stimuli. We predict an interaction

such that participants in relation-relevant condition will be more likely than participants in the feature-relevant condition to put circles in correspondence that differ in spatial position when stimuli differ in both relations compared to when the stimuli match on both relations. This prediction can be viewed as testing whether attention shifting-like phenomena (e.g., Kruschke, 1992) extend to relational stimuli (e.g., Tomlinson & Love, 2006).

3.6.1 Methods

3.6.1.1 Participants

Twenty-one undergraduate students from the University of Texas at Austin participated for course credit. Participants were randomly assigned to the feature- (n=11) or relation-relevant (n=10) condition.

3.6.1.2 Stimuli

The stimuli were the same as those used in Experiments 1-4.

3.6.1.3 Procedure

The learning phase of this experiment was conducted exactly as Experiment 2, with the same instructions detailing the manner in which the stimuli varied. Participants learned either a relation-based XOR or the feature-based XOR category structure. After the participants could correctly categorize 12 stimuli in a row, or after 288 trials, the participants were transferred to the second phase of the experiment.

In the second phase, the participants were instructed that they would see two stimuli and that one of the circles in one of the stimuli would be highlighted in red. The participant's task was to determine which circle in the other stimulus corresponded to the highlighted circle. Following this judgment, participants were instructed that they would then be asked to rate the similarity of the two presented stimuli on a scale from 1-9. Participants were alerted that they could take short breaks between responses to help maintain concentration.

On each trial, two stimuli were simultaneously presented on screen with text designating pair 1 and pair 2, along with text asking them to pick the circle that went with the highlighted circle. One pair was displayed on the top of the screen and the other on the bottom. A line separated the pairs. One of the circles (randomly determined each trial) was displayed with a red box around it. The participant used the mouse to select which of the two circles from the other stimulus corresponded to the highlighted circle. Immediately following this judgment, text appeared asking the participant to rate the similarity of the two stimuli on a scale of 1-9. Participants responded by pressing a key, 1 through 9. Following the participant's response, the screen blanked for 500 ms and the next trial began. The stimuli were presented in a random order and no participant saw a stimulus paired with itself, nor the same pairing twice. Unfortunately, due to a coding error, each participant saw only 105 of the 120 possible stimulus pairs (randomly determined for each participant).

3.6.2 Results and Discussion

For the learning phase, the proportion of trials correct for each participant was calculated. Remaining trials for participants reaching the learning criterion were scored as correct. The results replicated those from Experiment 2. Participants were significantly more accurate (.89 vs. .57) in the relation-relevant than in feature-relevant condition, $t(19) = 4.52, p < .001$. Similarly, a greater proportion of participants (9/10 vs. 4/11) reached the learning criterion in the relation-relevant condition than in the feature-relevant condition, although this difference did not reach significance, $\chi^2(1, N = 19) = .78, p = .38$.

Turning our focus to the correspondence judgments, we calculated the proportion of times each participant selected circles as corresponding that matched in spatial position (left or right). The overall proportions (.45 vs. .46) for the relation- and feature-based participants was not significantly different, $t < 1$.

To test our hypothesis that participants in the relation-based condition should favor positional correspondences when stimuli are relationally similar, but disfavor such correspondences when stimuli are relationally dissimilar, we calculated the proportion of positional correspondences for stimulus pairs that matched or mismatched on both relations for both the relation- and feature-based condition participants. These four means are shown in Figure 5. A 2 (relational match/mismatch) x 2 (Condition) mixed ANOVA found that the predicted interaction was significant, $F(1, 19) = 11.05, p < .01$. Additionally, there was a main effect of relational match/mismatch on the probability of a

participant aligning the circles based on position, $F(1, 19) = 48.77, p < .001$. There was not a significant main effect of condition, $F(1, 19) = 2.36, p = .14$.

The previous analysis confirms our predictions, but two questions remain in regards to the feature-based condition: whether participants are choosing correspondences at random and whether shared category membership affects correspondence judgments. To evaluate these, a supplementary ANOVA was conducted using the correspondence judgments from the feature-based participants. A 2 (relational match/mismatch) x 2 (category match/mismatch) within-subject ANOVA was conducted and found no interaction,$F < 1$, nor a main effect of the category variable, $F(1, 10) = 2.00, p = .19$. Although it is dangerous to over interpret null effects, the lack of any significant effects involving the category variable does ease concerns over our interpretation of the data shown in Figure 5. Providing further reassurance, rather than determining correspondences randomly, feature-based condition participants were more likely (60 vs. .40) to prefer position matches when stimuli matched than mismatched in their relations, $F(1, 10) = 13.67, p < .01$

3.7 Discussion

Contrary to accepted wisdom, the results of Experiment 1 demonstrated that learning to classify by relations can be easier than by features when the categories are balanced on the number of perceptual attributes and those attributes are spread across multiple objects. Experiment 2 showed that this effect persists in a more complicated category structure, when the category

structure is supported by the preferred alignments between the stimuli based on the relations. Experiment 3 employed a category structure for which online comparisons could be detrimental to relational learners and this structure tempered any inherent advantage for relations, as relation- and feature-based categories were acquired at the same rate. Indeed, relational learners who showed stronger markers of online comparison processes were less accurate overall. Experiment 4 further supported the notion that people make online comparisons of stimuli that aim to maximize perceived similarity. Experiment 5, supported this conclusion by demonstrating that learners in the relation-based condition preferred different correspondences for stimulus pairs than those in the feature-based condition.

The combined results from Experiments 1-5 advance our understanding of the role of online comparison processes during learning and preclude alternative explanations based on general biases in favor of features or relations. These results are important because they suggest revisiting findings demonstrating relational deficits, not just in adults and children, but in special populations such as those suffering from schizophrenia (S. Johnson, Lowery, Kohler, & Turetsky, 2005) and Alzheimer's Disease (Waltz et al., 2004), using the methods and well-matched stimulus set developed here.

Response times were also collected during category learning in Experiments 1, 2 and 3. Unfortunately, large differences in learning accuracy and in the proportion of participants reaching the learning criterion for Experiment 1 and 2's relation and feature conditions rendered a meaningful analysis

of response times untenable. In Experiment 3, accuracy levels for the two conditions were roughly equal, thus, comparing response times is meaningful. The mean of each participant's correct median response time was significantly greater (2501 ms vs. 1892 ms) in the relation than in the feature relevant condition, $t(42) = 2.72, p < .01$. According to the alignment view, this difference follows from the complexity of determining relational correspondences.

Another interesting finding related to the response times it that the response time in Experiment 4s similarity rating task was highly correlated (r=-.93) with rated similarity (see Figure 3.5). According to the alignment view, more readily aligned stimuli result in a feeling of fluency, which influences rated similarity (cf. W. A. Johnson, Dark, & Jacoby, 1985).

These experiments strengthen the link between work in relational processing and category learning. Earlier work, such as Lassaline and Murphy (1998), demonstrated the importance of considering feature matches across attributes (through alignment) in determining the difficulty of learning feature-based categories. Good alignments led to faster learning. We extend Lassaline and Murphy's findings to multi-place predicates, which allows for exploration of relation-based category learning. Our results support the view that the flexibility of relational processing can help or hinder learning with benefits observed when online comparison processes boost category coherency.

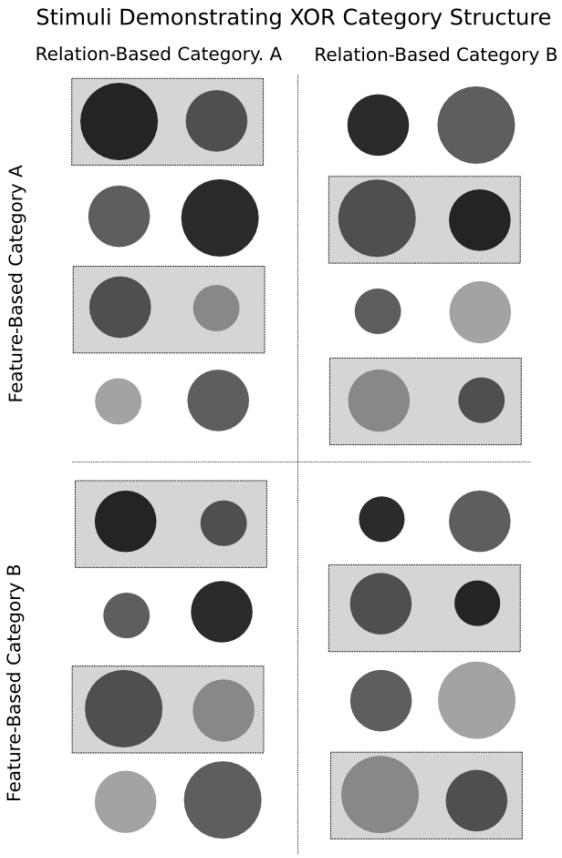

Figure 3.1: The 16 stimuli are arranged according to the relational (left vs. right) and featural (top vs. bottom) XOR category structure. The circles vary on 4 attributes: 2 features and 2 relations. The features are overall size and overall brightness (defined over both circles). The relations are which circle (by left/right spatial position) is bigger and which circle is brighter. The size-relation based 1-dimensional category groups stimuli where the larger circle is on the left in category A, and those with the larger circle on the right in B. For the size-feature, stimuli with large circles are in one category, while those with small circles are in the other. The feature-based XOR groups large dark stimuli with small light stimuli, while the relation-based XOR groups stimuli with darker circles on the left and smaller circles on the right with stimuli that have lighter circles on the left and larger circles on the right. The grey boxes are not part of any stimulus and are intended to promote the clarity of the figure by grouping constituent stimulus elements together.

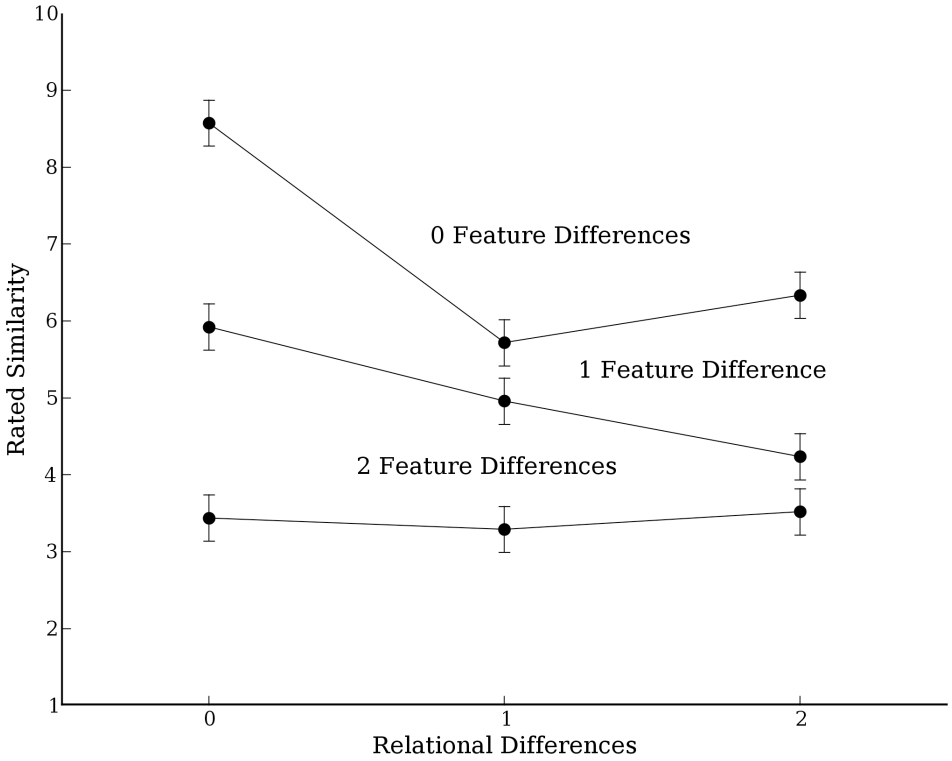

Figure 3.2: Experiment 4's mean similarity ratings as function of number of feature and relation differences. Error bars represent approximate 95% confidence intervals.

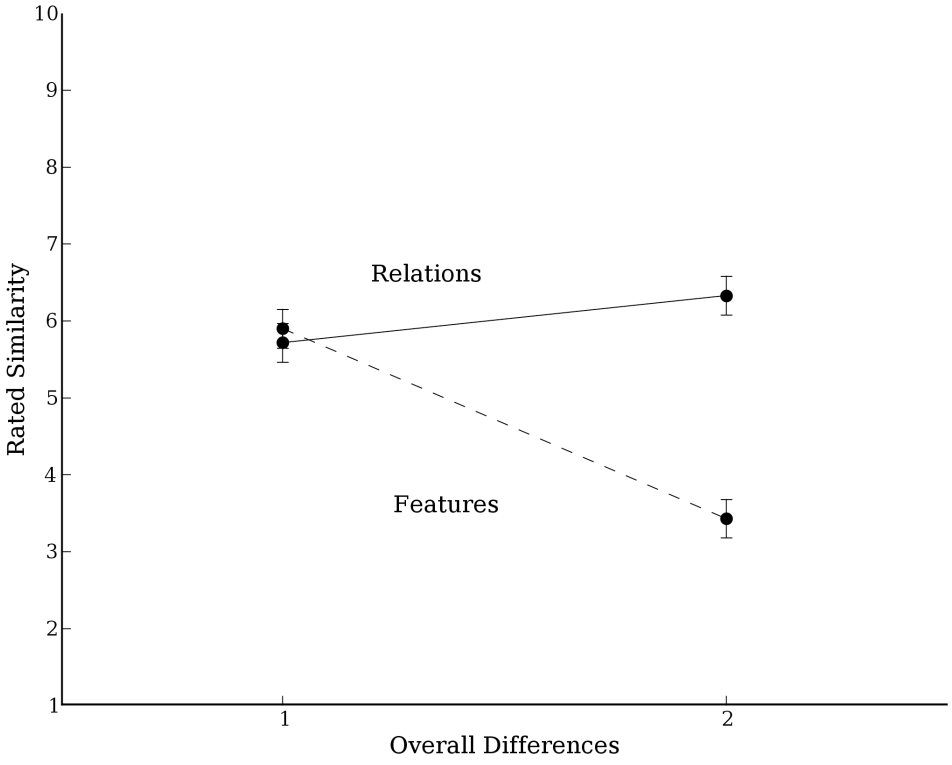

Figure 3.3: A subset of Experiment 4's mean similarity ratings reveals the strong interaction consistent with relational flexibility. Mismatching on both relations (with both features matching) increases similarity, whereas mismatching on both features (with both relations matching) decreases similarity. Error bars represent approximate 95% confidence intervals.

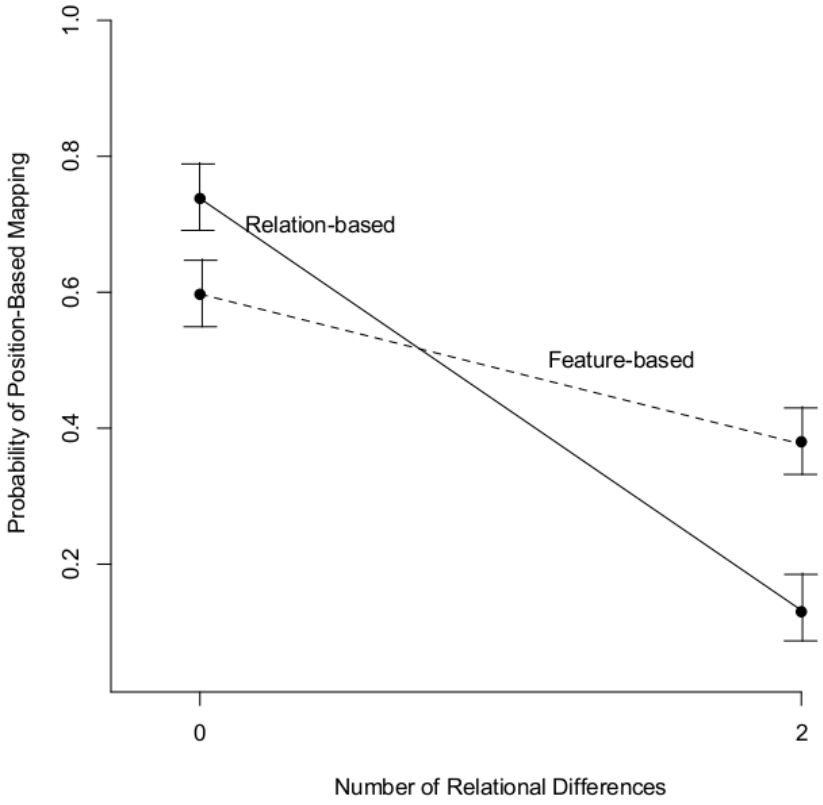

Figure 3.4: The mean probability that participants in the feature- and relation-based learning conditions establish stimulus correspondences based on position as a function of whether there were 0 or 2 relational differences between the stimuli. Error bars represent approximate 95% confidence intervals.

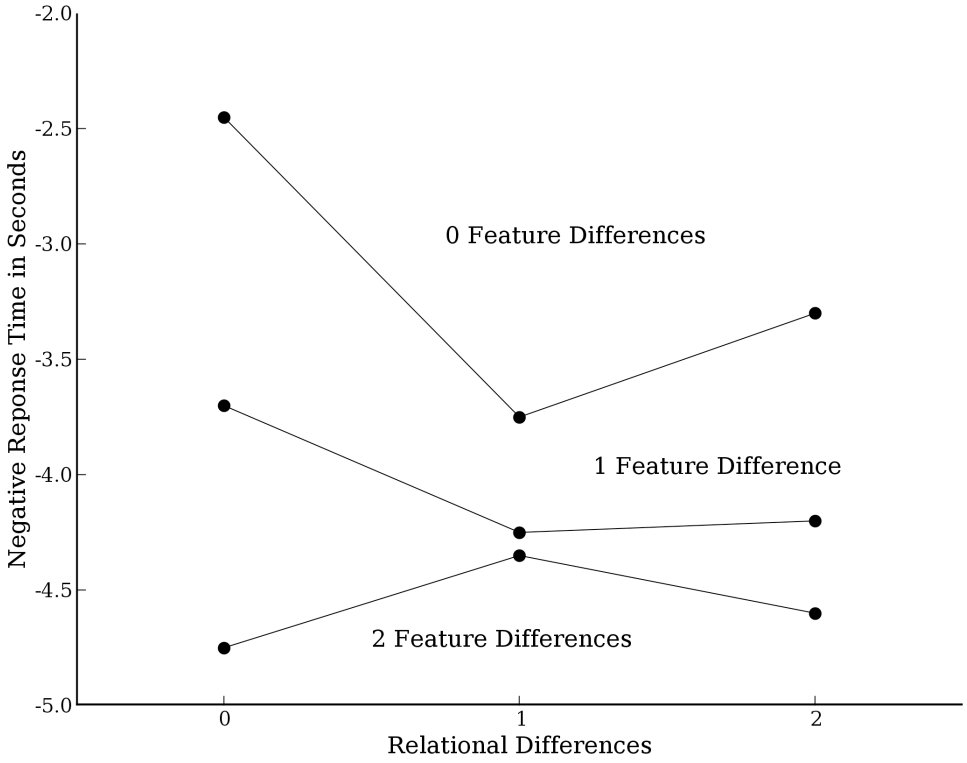

Figure 3.5: The mean of each participants median response time was calculated for each cell and its negative is displayed to ease comparison with the strongly correlated similarity rating data shown in Figure 2. To reduce visual cluster, error bars are not included, but 95% confidence intervals on the means are approximately +-.55 seconds.

Chapter 4

General Discussion

In the preceding sections we developed a theory of category learning that utilizes the representational frameworks provided by the analogy literature to learn categories that may only be described through the appreciation of the structured relations within their members. This theory provided an account of how learners gain abstract, relational understandings of domains based solely on comparing concrete instances. Its success arose from bridging two established fields: category learning and analogy.

In the supportive simulations, BRIDGES offered a model-based description of the course of generalizations from feature-based to relation-based. The simulations detailed how this progression is driven by the constraints of the task. When the outcome can be accounted for by the features, relational abstraction will not occur. In contrast, when the features are irregular and the outcome is predicted by the relations, then the model will weight relational similarity higher when comparing exemplars and form generalizations over the features.

The theory behind BRIDGES suggested that the difficulty of learning a relation-based category should depend on the congruence between a learner's

preferred alignment for category member pairs and the category structure. This effect played out in a series of experiments showing that learning is facilitated when the category structure aligns with the learner's preferred representation of the stimuli, and hindered when it is not. The data further supported the idea that learners change the way they align the category members and non-members based on the category structure. Whereas Rosch and Mervis (1975) focused on how the structure of the environment biases acquisition toward categories that have high within- and low between-category similarity, our findings suggest the cognitive machinery provided for online relational comparisons can exert a strong influence in regularizing categories to conform to the Rosch and Mervis ideal.

4.1 Relations and Roles

In this discussion of relational categories we have described a relational category as any category where successful categorization requires an appreciation of the structure within the exemplars and where category members are identified by their similar structure. Markman and Stilwell (2001) suggest a more refined description. They stipulate that structure can be used to identify two distinct types of categories: relational categories and role-based categories. Relational categories are those that describe relations between elements, such as *give* or *visit*, where the label describes the entire relational structure. In contrast, role-based categories are those that describe only the role a particular element of a relation plays within the larger relational structure, such as

predator or *visitor*. Role-based categories make up a significant portion of category-types; approximately 6% of the most frequent nouns describe role-based categories (Fan, Barker, Porter, & Clark, 2001). Additionally, people are also highly sensitive to role information. For instance, (Jones & Love, 2007) found that objects that play the same role are rated as more similar.

Our theory does not distinguish between relation-based and role-based categories. Because both types are supported by relational commonalities, they are learnable by BRIDGES, however it is not possible to discriminate an instance of a role from its sibling roles or its parent relation without something constraining the alignment mechanism. For example, in a scene showing a shark eating a tuna, *predator*, *prey* and *preying* are all equally as valid labels. Additional ambiguity is provided by the fact that tuna are also *predators*, in other situations.

One way to disambiguate the roles from the relations would be if support was provided in the form of a highlighting relation, such as a role tag (see Fan et al., 2001 for a comparable example) or a *focus* relation. These act to provide additional support for focusing on the correct part of the relational structure. During learning they provide a way to ensure proper alignment between scenes representing the roles. For instance, if the above example were supported by a phrase, such as "the tuna is also prey", which highlighted the entity as exhibiting a role alignment of the entity with respect to the role would be facilitated. This could occur even in the face of overwhelming feature evidence to the contrary.

An alternative way in which BRIDGES could learn to discriminate roles and relations would be if the training examples exhibited variability supportive of learning the role. For instance, exemplars labeled as *steal* might exhibit variability over all of the features, whereas exemplars which are labeled *thief* might exhibit greater stability in what is playing the role of thief. Assuming category specific attention or abstraction, this would cause attention within the *thief* category to be higher for features or relations supportive of the role (e.g. animacy or sinister), in contrast to the relation which would have more diffuse attention. This idea rests on the idea that such role specific delineators exist, in some form.

4.2 Rerepresentation and progressive alignment

DORA and other re-representational explanations (e.g., Yan, Forbus, & Gentner, 2003 and Kotovsky & Gentner, 1996) support the idea that children acquire new relations during the course of a short experiment. Our theory expresses the idea that these short experiments teach learners which aspects of the stimuli are important to succeeding in the task and that re-representation is not needed. We are not suggesting that re-representation does not happen; clearly re-representation occurs in scientific discovery (Gentner et al., 1997). Instead, it is important to establish differences between results that are explainable using attention shifting (e.g., Kotovsky & Gentner, 1996) and actual changes in representation as the result of a slow and possibly deliberate process.

In contrast to the short experiments of Kotovsky and Gentner (1996), experiments such as Gomez, Bootzin, and Nadel (In Press), while not conclusive of re-representation, seem more supportive. In this study, the researchers had infants learn an artificial grammar of the form aXb where a and b were contingent (a always predicted b), but X was one of 24 different words. Half of the infants were allowed to nap after training, while the remaining were not. The researchers found that when tested four hours after training the learners who napped showed evidence of abstraction, while those who did not nap showed a familiarity effect for the exemplars

Tied to the issue of re-representation is progressive alignment. A learned selective-attention theory of progressive alignment provides a more nuanced description of the phenomena. However, these approaches are not necessarily germane to progressive alignments that happen over actual re-representations. Importantly though, the model provides insight into what mechanisms and processes support progressive alignment on a shorter time-scale. Recall that BRIDGES suggests that progressive alignment is primarily influenced by two factors. One factor is that learning is error-based: If the outcome is too easy and perfectly predicted by the exemplars, then no learning will take place. The other factor is that attended attributes interfere with making a relation-based mapping; therefore, attention needs to be shifted away from those attributes. Learning theory provides a suite of models that suggest the possible ways for accomplishing this task.

4.3 Analogy across phylogeny and ontology

There is a large and involved debate across many literatures as to how well animals can reason with and utilize relational information. It is undeniable that many animals can succeed at simple relational tasks that involve relations between features, (Thompson et al., 1997; Young & Wasserman, 1997; Vonk, 2003), as well as demonstrate an understanding of relationships between phonemes (Hauser & Weiss, 2002). However, there are clearly limitations. For instance, there is no compelling evidence showing that animals can reason about high-order relations (see Penn et al., 2008a). Much of this work parallels the debate on the relational abilities of children (see Gentner & Ratterman, 1991).

While BRIDGES is capable of modeling both high-order relational commonality and low-order commonalities, and uses both in this thesis, we do not suggest that animals nor infants can necessarily do the same. Simulations 1-4 showed that relational behavior may be captured only using a *type* relation, supposing categorization of the objects within the scene by the learner. As pointed out by Penn, Holyoak, and Povinelli (2008b), these simulation do not require that learners reason about the relations-between relations, only that they can make flexible mappings between scenes involving relations between features and shift attention to these mappings.

As suggested by Penn et al. (2008a) the key differences might lay in the representation that the animals can support. Importantly our theory provides a formal framework where these investigations can take place in an environ-

ment that is supportive of learning and easily modified to account for different assumptions (such as the inclusion of only low-order relations, or limits to the number of mappings).

4.4 Comparison of alternative models and BRIDGES

BRIDGES is not the first model to use analogical alignment to support category learning. SEQL can acquire category structures through a process of repeated abstraction of a structured category representation, and has been successfully applied to the infant grammar learning studies considered here (Kuehne et al., 2000). While SEQL stresses building abstract representations of the category through the formation of a prototype, abstraction in BRIDGES arises from learned selective-attention. Both SEQL and BRIDGES utilize structure mapping, which raises the interesting possibility of a hybrid approach using prototype formation and selective attention to capture a wider-range of effects.

DORA (Discovery of Relations by Analogy, Doumas et al., 2008) is another model that seeks to explain how people acquire novel relational structure. DORA is patterned after the LISA model (Hummel & Holyoak, 1997) of analogy and incorporates LISA as a special case. DORA supports the use of structure mapping, like SEQL, but uses a distributed representation to encode features, relations, and the bindings between these while accounting for neural constraints.

Learning in DORA makes use of a set-intersection approach. DORA

creates new predicates by discovering consistency between scenes. For example, through comparing trucks and elephants, a single place predicate is created for *big*; analogous comparisons could lead to a predicate for *little*. Finally, by analyzing scenes with *big* and *little* things, a multi-place predicate is extracted for $bigger-than$, or $big\&little$. Importantly, as DORA learns new predicates and more complex representations for exemplars, it keeps copies of the old exemplars using the old representation in memory. This allows DORA to account for some instances of feature-based interference in relational learning. DORA's approach has been shown to work for modeling numerous experiments demonstrating relational learning (see Doumas et al., 2008).

The advantage of BRIDGES over both SEQL and DORA lies in its reliance on the theories developed within the learning and categorization literatures. These theories allow BRIDGES to make predictions about learning effects in relation-based categories, such as the learnability of different category structures (Medin & Schwanenflugel, 1981; Shepard et al., 1961) and blocking (Kruschke & Blair, 2000), which are inaccessible to the above models. Although, as evidenced in Experiments 1-5, care is required when designing these experiments, the presence or absence of these effects from well designed studies would provide a valuable piece of data.

Additionally, changes to BRIDGES are straightforward because of its generality. Integrating effects such as rapid attention shifting (Kruschke & Johansen, 1999) or exemplar specific attention (Kruschke, 2001) are trivial and well defined. These integrations allow BRIDGES to easily account for

effects such as the inverse-base-rate effect (Kruschke, 2003; Medin & Edelson, 1988) and highlighting (Kruschke et al., 2005). Due to the complexity of some of the competing models it is unclear how such mechanisms would be incorporated.

BRIDGES's internal representation is also easily changed. One idea is to allow for the formation of clusters, or prototypes within its hidden-layer. By utilizing the tested formulations behind SUSTAIN (Love et al., 2004), BRIDGES may be extended to create prototypes, similar to SEQL, when necessary. This allows the model to predict enhanced recognition for differentiated exemplars of relational categories, as well as removing some of the issues with regards to positing a complete mapping of every exemplar in memory at every time-step. Other solutions for this latter problem exist in the form of MAC/FAC (Forbus et al., 1994), which posits initial feature-based mappings, or EBRW (Nosofsky & Palmeri, 1997) which suggests a similar approach wherein exemplars race to take place in the model's decision.

Analogous to the notion of surprising exemplars creating novel clusters, is the idea of well learned categories could spur conceptual change, and an incorporation of the relational category into BRIDGES exemplar representation. Approaches to learning relational structure, such as DORA, create new predicates for those parts of the exemplars that overlap between examples, essentially forming a prototype of the relation (Doumas et al., 2008). In contrast, BRIDGES would need to create a representation of the relation as a function of the exemplars, category activation, and learned attention. Within BRIDGES

all of these parts would need to be utilized to form new relations. For example, BRIDGES would suggest that attention serves as a way to identify those pieces of the structure that should be connected by the new relation. Interestingly, because BRIDGES uses a family-resemblance structure, the model postulates that the same relation might be represented by multiple, distinct, structured representations.

While we have taken a domain general approach to a family-resemblance theory of relational category learning, other theorists have proposed domain general theories of analogical learning, such as for learning games (Linhares & Brum, 2007) or language (Bod, 2009). Initial steps have been taken to test the idea of language through analogy. For instance, (Goldwater, Tomlinson, Echols, & Love, In Press) presented evidence that children use structure-mapping to guide them in forming sentence structure. The study showed that young children are better at mapping more complex structures when aided by shared surface similarity. This tantalizing evidence suggests the possibility of a larger role for an analogical theory of learning outside of analogy and category learning.

References

Ahn, W. (1999). Effect of causal structure on category construction. *Memory & Cognition*, *27*, 1008-1023.

Allen, S. W., & Brooks, L. R. (1991). Specializing the operation of an explicit rule. *Journal of Experimental Psychology: General*, *120*, 3-19.

Altmann, G., & Dienes, Z. (1999). Rule learning by seven-month-old infants and neural networks. *Science*, *284*, 875a.

Anderson, J. (1991). The adaptive nature of human categorization. *Psychological Review*, *98*, 409-429.

Ashby, F., & Alfonso-Reese, L. (1995). Categorization as probability density estimation. *Journal of Mathematical Psychology*, *39*, 216-233.

Barsalou, L. W. (1985). Ideals, central tendency, and frequency of instantiation as determinants of graded structure of categories. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *11*, 629-654.

Bod, R. (2009). From exemplar to grammar: A probabilistic ananlogy-based model of language learning. *Cognitive Science*, *33*(5), 752-793.

Chi, M. T., Feltovich, P. J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science*, *5*(2), 121-152.

Clark, P., & Porter, B. (1997). Building concept representations from reusable components. In *Fourteenth national conference on artificial intelligence.*

Doumas, L. A. A., Hummel, J. E., & Sandhofer, C. M. (2008). A theory of the discovery and predication of relational concepts. *Psychological Review*, *115*, 1-43.

Duncan, J. (1984). Selective attention and the organization of visual information. *Journal of Experimental Psychology: General*, *113*(4), 501-517.

Eimas, P., Siqueland, E., Jusczyk, P., & Vigorito, J. (1971). Infant speech perception. *Science*, *171*(3968), 303-306.

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

Fan, J., Barker, K., Porter, B., & Clark, P. (2001). Representing roles and purpose. In *First international conference on knowledge capture.* ACM.

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

Genter, D., & Kurtz, K. J. (2005). Relational categories. In W. Ahn, R. Goldstone, B. Love, A. Markman, & P. Wolff (Eds.), *Categorization inside and outside the lab. washington.* American Psychological Association.

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

Gentner, D., Brem, S., Ferguson, R. W., Markman, A. B., Levidow, B. B., Wolff, P., et al. (1997). Analogical reasoning and conceptual change: A case study of johannes kepler. *Journal of the Learning Sciences*, *6*(1), 3-40.

Gentner, D., & Kurtz, K. J. (2006). Relations, objects, and the composition of analogies. *Cognitive Science*, *30*, 609-642.

Gentner, D., Loewenstein, J., & Hung, B. (2007). Comparison facilitates children's learning of names for parts. *Journal of Cognition and Development*, *8*, 285-307.

Gentner, D., & Markman, A. B. (1994). Structural alignment in comparison: No difference without similarity. *Psychological Science*, *5*(3), 152-158.

Gentner, D., & Markman, A. B. (1997). Structure mapping in analogy and similarity. *American Psychologist*, *52*, 45-56.

Gentner, D., & Ratterman, M. J. (1991). Language and the career of similarity. In S. A. Gelman & J. P. Byrnes (Eds.), *Perspectives on thought and language: Interrelations in development* (p. 225-277). London: Cambridge University Press.

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

Gerken, L. A. (2006). Decisions, decisions: infant language learning when multiple generalizations are possible. *Cognition*, *98*, B67-B74.

Gibson, B. M., & Wasserman, E. A. (2004). Time-course of control by specific stimulus features and relational cues during same-different discrimination training. *Learning & Behavior*, *32*(2), 183-189.

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

Gick, M. L., & Holyoak, K. J. (1983). Schema induction and analogical transfer. *Cognitive Psychology*, *15*, 1-38.

Goldstone, R. L. (1994a). Influences of categorization on perceptual discrimination. *Journal of Experimental Psychology: General*, *123*, 178-200.

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

Goldstone, R. L. (1996). Alignment-based nonmonotonicities in similarity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(4), 988-1001.

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

Goldstone, R. L., Medin, D., & Gentner, D. (1991). Relational similarity and the nonindependence of features in similarity judgments. *Cognitive Psychology*, *23*, 222-262.

Goldwater, M. B., Tomlinson, M. T., Echols, C. H., & Love, B. C. (In Press). Structural priming as structure-mapping: Children use analogies from previous utterances to guide sentence production. *Cognitive Science*.

Gomez, R. (2002). Variability and detection of invariant structure. *Psychological Science*, *13*, 431-436.

Gomez, R., Bootzin, R., & Nadel, L. (In Press). Naps promote abstraction in language learning infants. *Psychological Science*.

Hahn, U., Chater, N., & Richardson, L. B. (2003). Similarity as transformation. *Cognition*, *87*, 1-32.

Halford, G. (1984). Can young children integrate premises in transitivity and serial order tasks? *Cognitive Psychology*.

Halford, G. (1993). *Children's understanding: The development of mental models*. Hillsdale, NJ: Erlbaum.

Halford, G., Wilson, W., Guo, J., Gayler, R., Wiles, J., & Stewart, J. (1994). Connectionist implications for processing capacity limitations in analogies. In K. H. . J. Barnden (Ed.), *Advances in connectionist and neural computation theory, vol. 2, analogical connections.* Norwood, NJ: Ablex.

Hauser, M., & Weiss, D. (2002). Rule learning by cotton-top tamarins. *Cognition*, *86*, B15-22.

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

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

Huttenlocher, J., Duffy, S., & Levine, S. (2002). Infants and toddlers discriminate amount: Are they measuring? *Psychological Science*, *13*, 224-249.

Johnson, S., Lowery, N., Kohler, C., & Turetsky, B. (2005). Global-local visual processing in schizophrenia: Evidence for an early visual processing deficit. *Biological Psychiatry*, *58*(12), 937-946.

Johnson, W. A., Dark, V. J., & Jacoby, L. L. (1985). Perceptual fluency and recognition judgments. *Journal of Experiment Psychology: Learning, Memory, and Cognition*, *11*, 3-11.

Jones, M., & Love, B. C. (2007). Beyond common features: The role of roles in determining similarity. *Cognitive Psychology*, *55*, 196-231.

Kamin, L. J. (1969). Selective association and conditioning. In N. J. Mackintosh (Ed.), *Fundamental issues in instrumental learning* (p. 42-64). Halifax, CA: Dalhousie University Press.

Kokinov, B. (1994). A hybrid model of reasoning by analogy. In K. Holyoak & J. Barnden (Eds.), *Advances in connectionist and neural computation theory: Vol.2. analogical connections.* Ablex.

Kotovsky, L., & Gentner, D. (1996). Comparison and categorization in the development of relational similarity. *Child Development*, *67*, 2797-2822.

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, *99*, 22–44.

Kruschke, J. K. (2001). Toward a unified model of attention in associative learning. *Journal of Mathematical Psychology*, *45*, 812-863.

Kruschke, J. K. (2003). Attentional theory is a viable explanation of the inverse base rate effect: A reply to winman, wennerholm, and juslin (2003). *Journal of Experiment Psychology: Learning, Memory, and Cognition*, *29*(6), 1396-1400.

Kruschke, J. K., & Blair, N. J. (2000). Blocking and backward blocking involved in learned attention. *Psychonomic Bulletin & Review*, *7*(4), 636-645.

Kruschke, J. K., & Johansen, M. K. (1999). A model of probabilistic category learning. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *25*, 1083-1119.

Kruschke, J. K., Kappenman, E. S., & Hetrick, W. P. (2005). Eye gaze and individual differences consistent with learned attention in associative blocking and highlighting. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *31*, 830–845.

Kuehne, S., Gentner, D., & Forbus, K. (2000). Modeling infant learning via symbolic structural alignment. In *Proceedings of the twenty-second annual conference of the cognitive science society.* Hillsdale, NJ: Lawrence

Erlbaum Associates.

Kurtz, K., & Loewenstein, J. (2007). Converging on a new role for analogy in problem solving and retrieval: When two problems are better than one. *Memory & Cognition*, *35*, 334-341.

Larkey, B., L., & Love, B. C. (2003). CAB: Connectionist Analogy Builder. *Cognitive Science*, *27*, 781-794.

Lassaline, M. E., & Murphy, G. L. (1998). Alignment and category learning. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *24*(1), 144-160.

Lee, M. D., & Navarro, D. J. (2002). Extending the ALCOVE model of category learning to featural stimulus domains. *Psychonomic Bulletin & Review*, *9*, 43-58.

Linhares, A., & Brum, P. (2007). Understanding our understanding of strategic scenarios: what role do chunks play. *Cognitive Science*, *31*, 989-1007.

Lippmann, R. P. (1989). Pattern classification using neural networks. *IEEE Communications Magazine*, *27*(11), 47-64.

Love, B. C., Medin, D. L., & Gureckis, T. (2004). SUSTAIN: A network model of human category learning. *Psychological Review*, *111*, 309-332.

Lynch, E. B., Coley, J. B., & Medin, D. L. (2000). Tall is typical: Central tendency, ideal dimensions, and graded category structure among tree experts and novices. *Memory & Cognition*, *28*, 41-50.

Marcus, G. F. (1999). Do infants learn grammar with algebra or statistics? response to seidenberg & elman, eimas, and negishi. *Science*, *284*, 436-437.

Marcus, G. F., Vijayan, S., Bandi Rao, S., & Vishton, V. (1999). Rule learning by seven-month-old infants. *Science*, *283*(5298), 77-80.

Markman, A. B. (1999). *Knowledge representation*. Mahwah, NJ: Lawrence Erlbaum Associates.

Markman, A. B., & Gentner, D. (1993). Structural alignment during similarity comparisons. *Cognitive Psychology*, *23*, 431-467.

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

Markman, A. B., & Gentner, D. (1997). The effects of alignability on memory. *Psychological Science*, *8*, 363-367.

Markman, A. B., & Stilwell, C. H. (2001). Role-governed categories. *Journal of Experimental and Theoretical Artificial Intelligence*, *13*(4), 329-358.

Medin, D. L., & Edelson, S. M. (1988). Problem structure and the use of base-

rate information from experience. *Journal of Experimental Psychology: General*, *117*, 68-85.

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, *85*, 207-238.

Medin, D. L., & Schwanenflugel, P. J. (1981). Linear separability in classification learning. *Journal of Experimental Psychology: Human Learning & Memory*, *7*, 355-368.

Nosofsky, R. M. (1986). Attention, similairty, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, *115*, 39-57.

Nosofsky, R. M. (1988). Similarity, frequency, and category representations. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *14*, 54-65.

Nosofsky, R. M., & Palmeri, T. J. (1997). An exemplar-based random walk model of speeded classification. *Psychological Review*, *104*, 266-300.

Nosofsky, R. M., Palmeri, T. J., & McKinley, S. C. (1994, January). Rule-plus-exception model of classification learning. *Psychological Review*, *101*(1), 53-79.

Nosofsky, R. M., & Zaki, S. F. (2002). Exemplar and prototype models revisited: Response strategies, selective attention, and stimulus generalization. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *28*, 924-940.

Palmeri, T. J., & Nosofsky, R. M. (1995). Recognition memory for exceptions to the category rule. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *21*, 548-568.

Penn, D. C., Holyoak, K. J., & Povinelli, D. J. (2008a). Darwin's mistake: Explaining the discontinuity between human and nonhuman minds. *Behavioral and Brain Sciences*, *31*, 109-178.

Penn, D. C., Holyoak, K. J., & Povinelli, D. J. (2008b). Darwin's triumph: Explaining the uniqueness of the human mind without a deus ex machina. *Behavioral and Brain Sciences*, *31*(2), 153-169.

Posner, M. I., & Keele, S. W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology*, *77*, 241-248.

Ramscar, M. J. A., & Yarlett, D. G. (in press). Semantic grounding in models of analogy: An environmental approach. *Cognitive Science*.

Reed, S. K. (1978). Category vs. item learning: Iimplications for categorization models. *Memory & Cognition*, *612-621*, 6.

Rehder, B., & Hoffman, A. B. (2003). Eyetracking and selective attention in category learning. In *Proceedings of the annual conference of the cognitive science society.* Mahwah, NJ: Lawrence Erlbaum Associates.

Rehder, B., & Ross, B. H. (2001). Abstract coherent categories. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *27*(5), 1261-1275.

Rescorla, R., & Wagner, A. (1972). A theory of pavlovian conditioning: Variations in the ef- fectiveness of reinforcement and non-rein- forcement. In A. Black & W. Prokasy (Eds.), *Classical conditioning: Ii. current research and theory* (p. 64-99). New York: Appleton- Century-Crofts.

Richland, L. E., Morrison, R. G., & Holyoak, K. J. (2006). Children's development of analogical reasoning: Insights from scene analogy problems. *Journal of Experimental Child Psychology*, *94*, 249-273.

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

Rosch, E., & Mervis, C. B. (1975). Family resemblences: Studies in the internal structure of categories. *Cognitive Psychology*, *7*, 573-605.

Sakamoto, Y., & Love, B. C. (2004). Schematic influences on category learning and recognition memory. *Journal of Experimental Psychology: General*, *33*, 534-553.

Sakamoto, Y., & Love, B. C. (2006). ancouver, toronto, montreal, austin: Enhanced oddball memory through differentiation, not isolation. *Psychonomic Bulletin & Review*, *13*, 474-479.

Seidenberg, M., & Elman, J. (1999). Do infants learn grammars with algebra or statistics? *Science*, *284*(5413), 433.

Seidenberg, M. S., Marcus, G. F., Elman, J. L., Negishi, M., & Eimas, P. D. (1999, April). Do Infants Learn Grammar with Algebra or Statistics? *Science*, *284*, 433-+.

Shepard, R. N. (1964). Attention and the metric structure of the stimulus space. *Journal of Mathematical Psychology*, *1*, 54-87.

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

Shepard, R. N., Hovland, C. L., & Jenkins, H. M. (1961). Learning and memorization of classifications. *Psychological Monographs*, *75*(13, Whole No. 517).

Sheya, A., & Smith, L. B. (2006). Perceptual features and the development

of conceptual knowledge. *Journal of Cognition and Development*, *7*(4), 455-476.

Shultz, T., & Bale, A. (2001). Neural network simulation of infant familiarization to artificial sentences: Rule-like behavior without explicit rules and variables. *Infancy*, *2*, pp. 501-536.

Sloman, S. A., Love, B. C., & Ahn, W. (1998). Feature centrality and conceptual coherence. *Cognitive Science*, *22*, 189-228.

Smith, E. E., & Sloman, S. A. (1994). Similarity- versus rule-based categorization. *Memory & Cognition*, *22*, 377-386.

Smith, J. (2002). Exemplar theory's predicted typicality gradient can be tested and disconfirmed. *Psychological Science*, *13*, 437-442.

Smith, J. D., & Minda, J. P. (1998). Prototypes in the mist: The early epochs of category learning. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *24*, 1411–1430.

Thompson, R. K. R., Oden, D. L., & Boysen, S. T. (1997). Language-naive chimpanzees (pan troglodytes) judge relations between relations in a conceptual matching-to-sample task. *Journal of Experimental Psychology: Animal Beha- vior Processes*, *23*, 31-43.

Tomlinson, M. T., & Love, B. C. (2006). From pigeons to humans: Grounding relational learning in concrete examplars. In *Proceedings of the twenty-first national conference on artificial intelligence* (p. 199-204). Cambridge, MA: MIT Press.

Turney, P., & Littman, M. (2005). Corpus-based learning of analogies and semantic relations. *Machine Learning*, *60*, 251-278.

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

Vonk, J. (2003). Gorilla (gorilla gorilla gorilla) and orangutan (pongo abelii) understanding of first and second order relations. *Animal Cognition*, *6*, 77-86.

Waltz, J. A., Knowlton, B., Holyoak, K., Boone, K., Back-Madruga, C., McPherson, S., et al. (2004). Relational integration and executive function in alzheimer's disease. *Neuropsycholgia*, *18*(2), 296-305.

Wills, A., Reimers, S., Stewart, N., Suret, M., & Mclaren, I. (2000). Tests of the ratio rule in categorization. *The Quarterly Journal of Experimental Psychology*, *53A*(4), 983-1011.

Wittgenstein, L. (1953). *Philosophical investigations*. Oxford, England: Blackwell. (G. E. M. Anscombe, trans.)

Yan, J., Forbus, K., & Gentner, D. (2003). A theory of rerepresentation in ana-

logical matching. In *Proceedings of the twenty-fifth annual meeting of the cogntive science society.* Mahwah, NJ: Lawrence Erlbaum Associates.

Young, M. E., Ellefson, M. R., & Wasserman, E. A. (2003). Toward a theory of variability discrimination: finding differences. *Behavioral Processes*, *62*, 145-155.

Young, M. E., & Wasserman, E. A. (1997). Entropy detection by pigeons: Response to mixed visual displays after same-different discrimination training. *Journal of Experimental Psychology: Animal Behavior Processes*, *23*, 157-170.

Young, M. E., & Wasserman, E. A. (2001). Entropy and variability discrimination. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *27*(1), 278-293.

Younger, B. (1985). The segregation of items into categories by ten-month-old infants. *Child Development*, *55*(6), 1574-1583.

Vita

Marc Thomas Tomlinson graduated from Peoples Academy High School in Morrisville VT and the Vermont Academy of Science and Technology at Vermont Technical College in Randolph VT in the Spring of 1997. He pursued his Bachelor's degree in Computer Science with a major in Psychology at Rensselaer Polytechnic Institute, from where he graduated in 2000, after a year abroad at the University of Sussex in Brighton, UK.

Marc worked for Answerthink inc. for several years as a consultant working on database management and designing accounting tools for government regulated industries while working to receive a Masters degree in Applied Cognition from the University of Texas at Dallas in 2004. Marc also spent time working as a consultant for Humanalytics from 2003 until 2005. While there he designed text comparison software and assisted in designing a distributed database application for use in human resources departments.

Marc was accepted into the Doctoral program in Psychology at the University of Texas at Austin in the fall of 2004.

Permanent address: 211 Taylor Rd, Elgin TX, 78758

This dissertation was typeset with LaTeX[†] by the author.

[†]LaTeX is a document preparation system developed by Leslie Lamport as a special version of Donald Knuth's TeX Program.