Ties That Bind: Reconciling Discrepancies Between Categorization and Naming

Kenneth R. Livingston (livingst@vassar.edu**)**
Department of Psychology and Program in Cognitive Science
Vassar College, 124 Raymond Avenue, Box 479
Poughkeepsie, New York 12604

Janet K. Andrews (andrewsj@vassar.edu**)**
Department of Psychology and Program in Cognitive Science
Vassar College, 124 Raymond Avenue, Box 146
Poughkeepsie, New York 12604

Patrick Dwyer (padwyer@vassar.edu**)**
Department of Psychology
Vassar College, 124 Raymond Avenue, Box 3532
Poughkeepsie, New York 12604

Abstract

We present the results of a study designed to show that dissociations between lexical and similarity-based boundary partitions for a set of items can be produced in the laboratory. This is achieved by an incremental process of learning to assign a category label to items increasingly far removed (in similarity space) from the center of that category and approaching a different category. This process occurs in parallel with a compression effect in psychological similarity space such that increasingly distant items labeled as members of category A nonetheless come to be viewed as more similar to category B (the category to which they are in fact closer in pre-category learning similarity space) than they are by people who have not learned the category distinction.

Introduction

Although patterns of similarity are related in a more complex way to categorization and concept learning than was once thought, the evidence is mounting that such relationships are crucial to the partitioning of a set of items into categories (Medin, 1989; Medin, Goldstone, and Gentner, 1993; Goldstone, 1994a; 1994b). Among the most interesting recent findings is the discovery that the psychological similarity space usually assumed in the effort to measure relationships among category members is not static, but may actually undergo a change in its metric properties during the process of category learning. Thus, for example, Goldstone and his colleagues have found repeatedly that people who have learned to categorize a set of items make more reliable discriminations between pairs of items that cross the category boundary than people who have not learned to categorize them (Goldstone, 1994a; 1994b) This expansion or acquired distinctiveness effect seems to involve a stretching of the psychological similarity space in the region of the category boundary, with the result that smaller changes along the category relevant dimensions are sufficient to produce a just noticeable difference (JND).

Category clusters might also be formed by a process of compression in the region of similarity space that contains a set of items, with the result that a larger change is required to produce a JND. A number of researchers have found evidence for this process as well (Kurtz, 1996; Livingston, Andrews, and Harnad, 1998). Either of these processes, expansion or compression of the similarity space, is sufficient to produce a categorical distinction between sets of items, and thus looks promising as a general mechanism for concept learning. On such an account, a concept is formed when regions of psychological similarity space are warped so as to create a relatively more compact representational structure that can be more readily manipulated as a unit.

This description of a general mechanism for category learning is called into question, however, by recent findings concerning the relationship of language labels to category similarity structures. Malt, Sloman, Gennari, Shi, and Wang (1999) asked native speakers of English, Spanish, and Chinese to name each member of a set of sixty containers. They then asked these same people to sort these sixty items into groups based on their overall similarity. Perhaps not surprisingly, language communities differ in their lexical partitioning of the set of items. Chinese speakers partition the set using five words, English speakers use 7, and Spanish speakers use 15. The variation in grain is not the result of simply forming a greater number of subcategories in,

say, Spanish, of the same larger category boundaries found in Chinese; there is substantial non-shared variance in these lexical groupings.

The real surprise comes when one examines the clustering of items that occurs during the sorting task. Malt, et al., found that the sorts were very similar across language communities, in spite of the disparity in lexical boundaries for the set. There seems to be a dissociation of lexical boundaries from category boundaries based on similarity, and this creates something of a theoretical problem for the account of category learning given above. Indeed, it constitutes a puzzle for any theory of conceptual structure that suggests that there is a central tendency in the representation of the members of a category. The widely held belief that terms derive their meanings from their links to coherent concepts seems inconsistent with this result.

Unfortunately, we have no information about the kinds of judgments that people might have made of Malt, et al.'s set of containers *prior* to learning to name and categorize them. We therefore have no way of knowing how this odd state of affairs came to pass, nor whether it really constitutes a disconfirmation of the claim that psychological similarity space warps in a coherent fashion during category learning.

How might an item come to have a name other than that of its nearest neighbors with whom it shares a region in similarity space? One possibility discussed by Malt, et al. (1999) is that a series of intermediate cases could be introduced, one at a time, each inheriting the name of its nearest neighbor, and each more remote from the category to which the name was initially attached. If the chain of items spans the boundary between two categories, one could wind up with instances that are labeled as members of category A, the point of origin for the chain, even while having more in common perceptually with the members of category B. The dissociation of lexical and similarity groupings occurs because the former is based on nearest exemplar pairings introduced incrementally, while the latter is based on the warping of similarity space described above.

In order to test this hypothesis, we constructed a set of stimuli whose distribution in similarity space allowed for partitioning into two categories while leaving a set of items in the space between these groupings to allow the building of a naming chain from one to the other. Success at building such a chain would constitute an existence proof for this process, and comparison of data from people who learned to categorize the set with data from people without category or name-learning experience would allow us to determine whether this process alters the character of any warping of the similarity space.

Method

Participants

Participants were seventy-eight Vassar College undergraduates who were given course credit in an introductory psychology course. Twenty people participated in preliminary research to select an appropriate stimulus set. The remaining fifty-eight people participated in the experiment reported here, twenty-two of them in a control group, and eighteen in each of two experimental groups.

Stimuli

Ten members of the Nemipteridae family of fish (threadfin breams) and ten members of the Labridae family (wrasses) were selected for preliminary analysis from Burgess, Axelrod, and Hunziker (1997). Each stimulus was color photocopied, glued onto a blank card (5.1 by 10.8 cm), laminated, and then randomly assigned a number from 1 to 20, which was printed on the back of the card. Following examination of the two-dimensional solution to a multidimensional scaling analysis (MDS) of the similarity judgments of twenty people on all 190 possible pairs (see Procedure below), the set was culled to remove outlying cases. These were then replaced with items that occupied the central region of the 2D space, which appeared to be defined by the dimensions of degree of body striping, and the ratio of body width to length. These replacement items were selected without regard to membership in the two original categories.

Procedure

Participants were assigned either to a control condition (N=22) or to one of two learning conditions (N=18 each group). Participants in the control condition were asked to judge the degree of similarity between all possible pairs of the twenty stimuli (190 pairs). The participant was seated at a table upon which the stimuli had been placed face up, and was allowed to inspect the entire stimulus set for one minute. The stimuli were then turned over, revealing the numbers on the back, and the experimenter began naming pairs in a previously determined random order. The participant was instructed to turn over each corresponding pair, look at the two items, and verbally rate the similarity of the stimuli on a 9-point scale from 1 (most similar) to 9 (most different). Ratings were provided to one decimal place.

For control group participants, the similarity judgment task was the only task required. All control group participants were run through the procedure in a block before work began with the experimental group. This allowed us to complete an MDS analysis of the data from this group to confirm the pattern of similarity relationships in the set and the choice of stimuli for building the chain between categories. As expected, the MDS analysis of the control group data revealed that

stimuli clustered in two core groups consisting of eight and seven items, respectively. The five remaining stimuli were intermediate cases, four of which formed a chain between the two larger groups. One intermediate stimulus that was not part of the chain was treated as a neutral stimulus, and did not appear in the training task. It served as the means to a test for demand effects in the data set (see Results, below).

This analysis of the pre-categorization similarity space allowed us to design the stimulus sets for the learning group. Learning participants were first taught to categorize the core set of fifteen stimuli. Stimuli were presented individually, in blocks of fifteen that included all members of both categories. The participant was asked to label each picture as either *gracilia* or *aurora*. The experimenter gave immediate feedback, recorded the response, and then presented the next picture. Order of presentation within each trial block was random. Training continued until the participant met the criterion of two consecutive errorless trial blocks (30 stimuli), or a total of 20 trial blocks had passed, whichever came first.

Once category training on the core stimuli was complete, the first stimulus in the chain was introduced into the subsequent trial block. Which stimulus this was depended on the direction in which the chain was being built. For half of the study participants, the chain was built from *gracilia* to *aurora* (the G-root group) and for half it was built in the opposite direction (the A-root group). Assignment to these subgroups was random. The first chaining stimulus introduced was the one closest to the root category in the MDS space derived from the control group data.

Once this new stimulus was introduced into the set, training continued as before, except that trial blocks now contained one additional stimulus. Order of presentation was still random, except that the new chaining stimulus was always presented immediately after its nearest neighbor in the root category. Training using this newly expanded set continued until the participant met the criterion of two consecutive trial blocks in which all chaining stimuli presented up to that point had been correctly categorized, or a total of 5 trial blocks (after introduction of the new item), whichever came first. Once the participant met criterion for one chaining stimulus, the next stimulus in the chain was introduced into the subsequent trial block. The new chaining stimulus always followed the previously learned one, which appeared randomly in the trial block along with the 15 core stimuli and any other chaining stimuli. Training was halted when the participant met the training criterion after the introduction of the fourth and final item in the chain.

Once training was completed, the participant took a short break before completing the similarity judgment task. Procedures for this task were exactly as for the people in the control group, and are described above.

Results

Six people in the learning group did not meet the criterion for successful learning by the conclusion of training. All six of these people were in the G-root group. Data for those who failed to reach criterion were excluded from all analyses. Mean similarity ratings were calculated separately for the control and learning groups, and, within group, mean similarities were calculated for *aurora-aurora* (A-A) pairs, *for aurora-gracilia* (A-G) pairs, and for *gracilia-gracilia* (G-G) pairs. Separate analyses were then performed for the two different chaining groups, using the same control group data for comparison in both cases.

In order to determine whether category learning produced compression and/or expansion effects in the A-root group, we performed a 2 (group: control vs. learning) by 3 (pair type: A-A, A-G, G-G) analysis of variance (ANOVA) on mean similarity ratings with repeated measures on the second variable. The analysis revealed a significant main effect of pair type, $F(2,76) = 149.506$, $MSE = 63.152$, $p < .0001$; and a significant interaction effect, $F(2,76) = 9.952$, $MSE = 4.202$, $p < .001$. (See Figure 1). For the G-root group, the same 2 (group: control vs. learning) by 3 (pair type: A-A, A-G, G-G) ANOVA with repeated measures on the second variable revealed significant main effects of group, $F(1,32) = 9.488$, $MSE = 22.437$, $p < .005$, and pair type, $F(2,64) = 57.551$, $MSE = 20.376$, $p < .0001$. The interaction effect was not significant. (See Figure 2.)

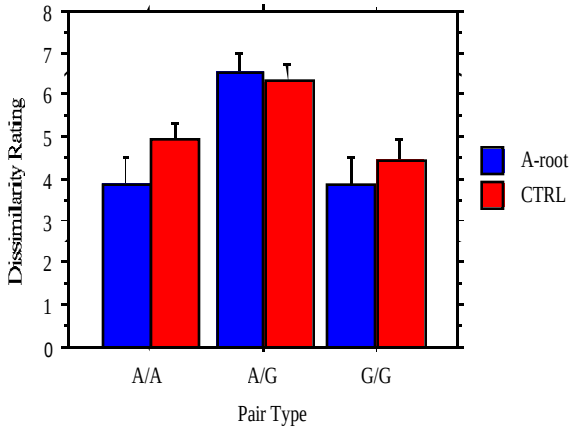

Figure 1. Comparison of the mean similarity ratings of the control group and the group who learned a chain of items rooted in the *aurora* category. Interaction of group with pair-type is shown.

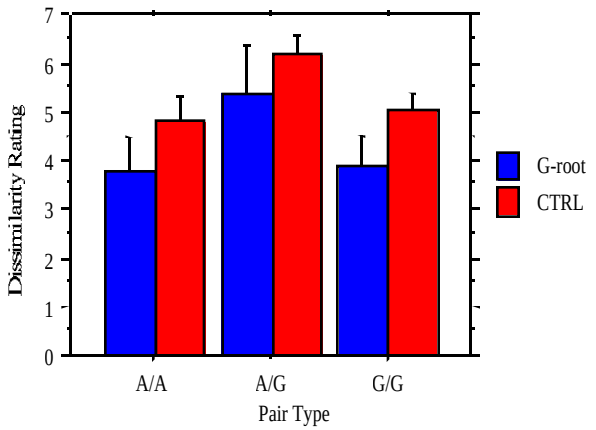

Figure 2. Comparison of the mean similarity ratings of the control group and the group who learned a chain of items rooted in the *gracilia* category. Interaction of group with pair-type is shown.

Thus, compression occurred in both learning groups, i.e., item pairs were judged to be more similar relative to control group ratings, particularly for the within-category (A-A and G-G) pairs. No expansion occurred, i.e., between-category pairs (A-G) were not judged to be less similar by the learning groups than by the control group.

Neutral Stimulus Analysis

Goldstone, Lippa, and Shiffrin (2001) have devised an ingenious technique for detecting the presence of demand effects in data of this kind. The analysis works by comparing the pattern of changes in similarity relationships among items in a learning set relative to a neutral item not included in training. If the similarity judgments of pairs including the neutral item change in ways that are predictable from the compression or expansion effects that occur for categorized items, this must be due to actual changes in the underlying similarity space and not demand effects, since the neutral item was never categorized.

For our data set, the absolute difference for each participant between all possible within-group and between-group pairs of pairs involving the neutral stimulus was calculated, averaging separately across each group. Separate 2 (group: learning vs. control) by 2 (pair of pair type: within vs. between) ANOVAs with repeated measures on the second variable were conducted for both the A-root and the G-root groups. The pattern of results was the same as that reported above. There was thus no evidence that the observed compression was due to a demand effect for either group.

MDS Analysis

In order to better understand the nature of the changes taking place in psychological similarity space we performed a multi-dimensional scaling analysis of the mean similarity ratings of all three groups (control, A-root, and G-root). The full matrix of mean similarities from each of the three groups was entered into an INDSCAL analysis. The two-dimensional solution provides a relatively good fit to the data (R-squared = .872), and is plotted in Figure 3. The locations of all twenty stimuli for each of the three groups (control, A-root, and G-root) are depicted, and the pattern of changes in similarity is shown by arrows connecting three points. The point at the tail end of the arrow represents the location in the space of the stimuli as judged by the control group. The middle point shows the locations of the twenty stimuli as judged by the G-root group, while the points at the tips of the arrow heads show the locations of the twenty stimuli as judged by the A-root group. The graph gives a sense of the compression that occurs in the similarity space as categories are learned. Note that the greater relative proximity of items following compression of the similarity space has the effect of making them a more easily identified grouping, even though there is no statistically significant increase in mean inter-item differences across the category boundary. Most importantly, the graph shows how the chained items move toward the central tendency of their nearest neighbor cluster, even when the label training ties them to the more remote cluster.

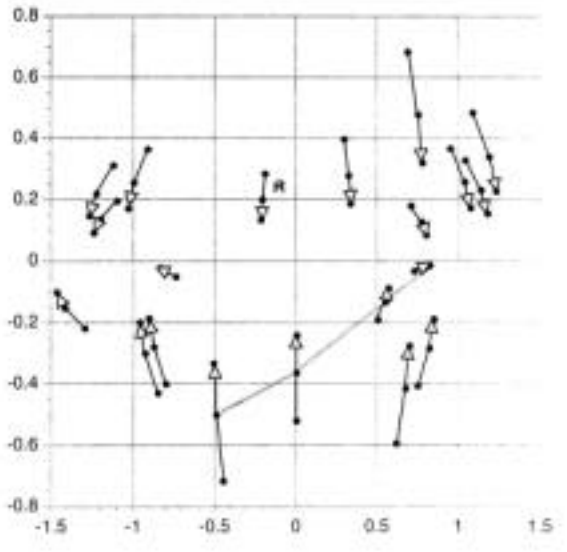

Figure 3. Shows the two-dimensional MDS analysis (INDSCAL) of similarity relationships among the twenty stimulus items for all three groups. The arrow labeled R is for the neutral item (see text). The four arrows linked by a shaded line are the chaining items.

The remaining points on the right side of the graph are the core items in the *aurora* category, while those on the left are the core items in the *gracilia* category. The space of similarities for the control group is represented by closed circles found at the tails of the arrows. The circles at the tips of the arrow heads represent the space of similarities for the experimental group that learned a chain that begins in the *aurora* category and extends toward the *gracilia* category. The filled circles found roughly in the middle regions of the arrows represent the space of similarities for the group that learned a chain that begins in the *gracilia* category and extends toward the *aurora* category.

Discussion

The domain of words must map in some predictable way to the domain of concepts if there is to be anything to the story that says words have meaning in virtue of the concepts to which they are linked. This is why Malt, et al.'s (1999) finding of an apparent dissociation between lexical boundaries and category boundaries in similarity space is so provocative. The experiment reported here demonstrates one process by which names might become attached to items that are remote in similarity space from the core of the concept to which the name typically refers. Note that although we did not measure typicality in this study, one further prediction would be that chained items should be seen as atypical of the category they name.

More important than the demonstration of a procedure for producing this dissociation are the data showing how this effect is related to the warping of similarity space previously shown to occur during category learning. This effect was clear in comparisons of the A-root group (people who learned a chain that begins with an item well within the region of the *aurora* group) with the control group. Analysis of variance revealed the same pattern of within-category compression found in previous research (e.g., see Kurtz, 1996; Livingston, et al., 1998). The effect is less clear in the G-root group. Compression occurs in this case, but it occurs for between category pairs as well as for within-category pairs. Obviously, the direction in which we tried to build chains made a difference, an observation further confirmed by the fact that all of the study participants who failed to reach our learning criterion were in the G-root chaining group. The nature of the difference between our two experimental groups, and its consequences for similarity judgments, can be seen in Figure 3.

First, notice that the chains differ in how deeply rooted they are in their originating categories. The *aurora*-based chain begins with an item well inside the region of similarity space that encompasses the category, and it does not extend very deeply into the region occupied by the *gracilia* category. Exactly the opposite is true for the G-root chaining group. Thus, even when people are learning to label the last two items in the chain leading deeply into *aurora* territory as *gracilia*, the region of space that they occupy is being compressed still further around the central tendency of the *aurora* category. It is therefore not surprising that one sees evidence of what counts as between-category compression for this group, because those last two items in the chain are considered *gracilia* for purposes of this analysis.

Looked at from this lexical perspective, the result seems straightforward enough, but the effect is far more interesting for what it tells us about how category learning warps similarity space. Notice that the lexical chaining effect seems to have relatively little effect on the direction in which similarity space is compressed during learning. The overall magnitude of the effect is different in the two cases, an effect that is likely the result of the fact that the task is more difficult and so produces slower and less robust learning in the G-root group, but the space warps in the same way in both cases. This warping, based on the pattern of structure and variation in the whole set of items, overrides local tendencies associated with rogue items chained in from elsewhere. As can be seen in Figure 3, the result is that the chained items move in the direction of the nearest region of compression, even while they are being labeled as members of the more remote cluster. The similarity-based warping of the representational space occurs independently of labeling. Thus are lexical and similarity-based category dissociations produced.

If this account is correct, several further predictions follow. We have already mentioned the predictions for typicality. Over long periods of time, we would also expect chains that stretch far from their roots would become unstable and break, with items at the ends of those chains receiving new labels more in keeping with those of their similarity-space neighbors. Linguistic analysis of patterns of lexical evolution should reveal such phenomena in the history of any language. We would also expect this pattern to be most common for artifact categories, where genuinely new instances are introduced with some frequency, and the perceptual and functional feature landscape is quite fluid. We are currently conducting a replication and extension of the study reported here using artifact categories and additional control groups. Finally, these results have deeper implications for the nature of the relationship between systems for representing lexicons, at least and systems for representing category information. These two systems must remain in register to some extent, but it is clear that each is sufficiently modular with respect to the other to permit some rather remarkable dissociations to develop in very short order.

Acknowledgments

Our thanks to the Undergraduate Research Summer Institute at Vassar College, and to Maria Jalbrzikowski and Paul Francaviglia for their assistance in the conduct of this experiment.

References

Burgess, W.E., Axelrod, H.R., & Hunziker, R. (1997). Dr. Burgess's *Mini-Atlas of Marine Aquarium Fishes Mini-Edition*. Neptune City, NJ: TFH Publications, Inc.

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

Goldstone, R. L. (1994-b). The role of similarity in categorization: Providing a groundwork. *Cognition, 52*, 125-157.

Goldstone, R. L., Lippa, Y., & Shiffrin, R. (2001). Altering object representations through category learning. *Cognition, 78*, 27-43.

Kurtz, K. J. (1996). Category-based similarity. In G. W. Cottrell (Ed.) *Proceedings of the Eighteenth Annual Conference of the Cognitive Science Society*, 290.

Livingston, K. R., Andrews, J. K., & Harnad, S. (1998). Categorical perception effects induced by category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24*, 732-753.

Malt, B.C., Sloman, S.A., Gennari, S., Shi, M., & Wang, Y. (1999). Knowing versus naming: Similarity and the linguistic categorization of artifacts. *Journal of Memory and Language*, *40*, 230-262.

Medin, D.L. (1989). Concepts and conceptual structure. *American Psychologist, 44*, 1469-1482.

Medin, D.L., Goldstone, R.L., & Gentner, D. (1993). Respects for similarity. *Psychological Review, 100*, 254-279.