Machine Learning 3: 343–372, 1989

Conceptual Clustering, Categorization, and Polymorphy

STEPHEN JOSÉ HANSON[†] (JOSE@BELLCORE.COM)
Bell Communications Research, 435 South Street, Morristown, NJ 07960, U.S.A.

MALCOLM BAUER (MALCOLM@CONFIDENCE.PRINCETON.EDU)
Cognitive Science Laboratory, Princeton University, 221 Nassau Street,
Princeton, NJ 08542, U.S.A.

(Received: October 1, 1986)
(Revised: October 31, 1988)

Keywords: Conceptual clustering, categorization, correlation, polymorphy, knowledge structures, coherence

Abstract. In this paper we describe WITT, a computational model of categorization and conceptual clustering that has been motivated and guided by research on human categorization. Properties of categories to which humans are sensitive include best or prototypical members, relative contrasts between categories, and polymorphy (neither necessary nor sufficient feature rules). The system uses pairwise feature correlations to determine the "similarity" between objects and clusters of objects, allowing the system a flexible representation scheme that can model common-feature categories and polymorphous categories. This intercorrelation measure is cast in terms of an information-theoretic evaluation function that directs WITT's search through the space of clusterings. This information-theoretic similarity metric also can be used to explain basic-level and typicality effects that occur in humans. WITT has been tested on both artificial domains and on data from the 1985 *World Almanac*, and we have examined the effect of various system parameters on the quality of the model's behavior.

1. Introduction

The majority of machine learning research has followed AI in using logic or predicate calculus for the representation of knowledge. Such logical formalisms have many advantages: they are precise, general, consistent, powerful, and extensible, and they seem distantly related to natural language. Early research on learning from examples (Winston, 1975) successfully used logical definitions for concepts, and recent work on conceptual clustering (Michalski, 1980) has done so as well. Mitchell's (1978) work on version spaces and the more recent

[†] Also with the Cognitive Science Laboratory, Princeton University, 221 Nassau Street, Princeton, NJ 08542, U.S.A.

results with explanation-based methods (Mitchell, Keller, & Kedar-Cabelli, 1986; DeJong & Mooney, 1986) have also seemed to validate this framework.

Nonetheless, there are at least four reasons why representations employing logic or rules tend to be disadvantageous for concept formation. First, rule-based approaches have generally adopted an "Aristotelian" view of categories, assuming a membership rule that requires necessary and sufficient conditions for the members of the category. Second, concepts represented by logic tend to have precise, fixed boundaries and little or no category structure or gradient of membership (see Zadeh, 1965, for a general weakening of this assumption). Third. logical approaches have focused on common feature mechanisms (e.g., "conjunctive clustering"), which can misrepresent and ignore relations among features. Finally, logical descriptions of categories emphasize the absolute properties of individual categories and mitigate the use of relative category properties or between-category comparisons.

In contrast, we will make several assumptions that are consistent with knowledge about human categorization, but at the same time run counter to the majority of categorization models employing logical representations:

(1) Categories tend to possess members which have features that are neither necessary nor sufficient (*polymorphy*).

(2) Categories have a *distribution* of members, some more representative and some less (Posner & Keele, 1968; Homa, 1978).

(3) Categories can be represented by the *intercorrelation* between feature sets or relations between features; such intercorrelations can be used as a measure of the coherence of the concept underlying a category and provide a first cut at the causal structure underlying the category (Murphy & Medin, 1985; Medin, Wattenmaker, & Hampson, 1987; Estes, 1986).

(4) Categories arise as *contrasts* between one another; in other words, categorization is relative to the existing context of other putative categories (cf. Rosch & Lloyd, 1978; Smith & Medin, 1981).

Central to the present approach is the concept of *polymorphy.* This notion intersects with many of the important heuristics about category formation that we will employ. Polymorphy can be defined in many mutually reinforcing ways. For example, in logic it is known as "majority logic" (M out of N logic) and falls somewhere between a conjunction and a disjunction operator.

Figure 1 shows two examples of polymorphous categories. Both are cases of "two out of three" polymorphy where two out of the three features COLOR, SHAPE, and SYMMETRY were systematically varied to produce exemplars. There is evidence that people can learn to sort these stimuli, but that they cannot generally give a verbal account of their sorting strategy (Dennis, Hampton, & Lea, 1973). This sort of dissociation between verbal reporting and categorization indicates that the ability to *describe* "rules" is not necessarily related to the ability to *use* "rules."

Another view of polymorphy comes from the philosopher Wittgenstein (1953) For Wittgenstein, categories were composed of entities that possessed "family resemblance" and belonged together, not because of necessary or sufficient

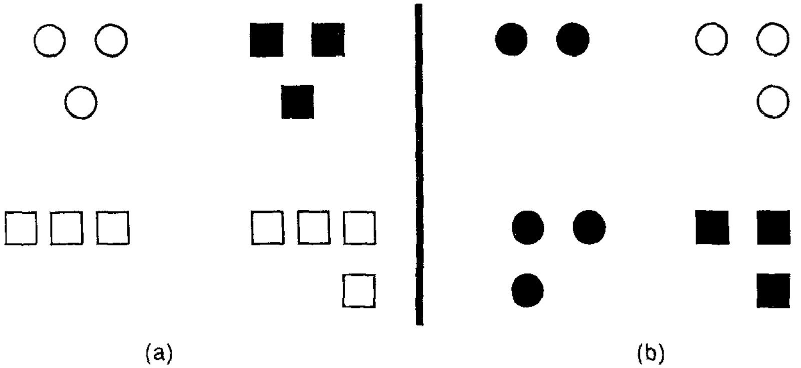

Figure 1. An example of two out of three polymorphy, adapted from Dennis, Hampton, and Lea (1973). Each "object" consists of two out of three possible features from SHAPE (CIRCLE or SQUARE), COLOR (BLACK or WHITE), or SYMMETRY around a 90 degree line (SYMMETRIC or ASYMMETRIC). NUMBER OF SYMBOLS per object is an arbitrary feature. Groups (a) and (b) each define a category that includes at least two out of three of the previously described features.

features, but because of polymorphous or "polythetic" sets of features. Hence, "game" is a coherent category because it includes sets of features that link various kinds of games together.

Finally, another view critical to our approach is the way in which feature intercorrelations and polymorphy interact. In fact, polymorphy *requires* some reference to feature correlations or, more generally, feature *relations*. The absence of common features in a category (polymorphy) encourages analysis of the feature relations in order to account for structural aspects of the category. Feature relations are important in human categorization because they may indicate something about the causal or functional nature of the category in which they appear (Murphy & Medin, 1985; Medin, Wattenmaker, & Hampson, 1987). Feature relations also have been implicated in the recognition of *basic-level* categories (Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976).

The remainder of the paper describes WITT, a computational model of human categorization that represents and acquires concepts using feature relations. Although the facts concerning human categorization are complex, there are a few generalizations that can be drawn from the literature. We presented several above that have directly influenced the design of our model. After describing the system's representation and its learning algorithm, we evaluate its behavior along three dimensions. We first report some experiments that examine the effect of varying system parameters in both artificial and natural domains. Next we consider the computational complexity of the learning scheme. Finally, we consider the model's ability to account for two well-established psychological phenomena, basic levels and typicality. We close with a discussion of the program's relation to earlier clustering work, along with some suggestions for future research.

2. An overview of WITT

In this section we present an overview of WITT, a conceptual clustering system that incorporates the assumptions listed earlier. Michalski and Stepp (1983a) provide a definition of the conceptual clustering task: given a set of instances, one must place those instances into disjoint clusters and formulate descriptions for each category. In addition, they argue that one should take the quality of the concept descriptions into account when evaluating alternative clusters of instances. We will see that WITT follows this strategy, though it does so in a quite different manner than Michalski and Stepp's CLUSTER/2 system. Below we describe the model's representation of instances and clusters, its metric for evaluating cluster quality, and its overall algorithm for clustering objects.

2.1 Representing instances and concepts

Like most conceptual clustering systems, WITT represents each instance as a set of attribute-value pairs. The current version limits itself to nominal (symbolic) attributes, but attributes can take on a *set* of values. For instance, the representation for a specific person might be:

PERSON-1	NAME	{MARY}
	PROFESSION	{LAWYER}
	LOCATION	{NEW-YORK-CITY}
	OWNS-CAR	{PORSCHE}
	MARITAL-STATUS	{DIVORCED}
	DOMICILE	{APARTMENT-IN-NEW-YORK, CONDO-ON-LONG-ISLAND}
	CHILDREN	{BILLY, SALLY, EDDIE}
	VACATIONS	{BAHAMAS, COLORADO, EUROPE}

Thus, all three of Mary's children are listed under the CHILDREN attribute, and the attributes DOMICILE and VACATIONS similarly have multiple values. One can view each instance as a point in a multidimensional space, and this leads naturally to the notion of distance between instances. This idea will play an important role in our clustering algorithm.

Although WITT's representation of instances is fairly standard, the system's representation of concepts diverges widely from traditional approaches. The system focuses on the intercorrelation among features, or what we will refer to as the *relational structure* implicit in the features. Thus, in WITT, concepts are represented in terms of co-occurrences between pairs of features (attribute-value pairs).[1] *Contingency tables* provide a natural way of representing such correlations; these are two-dimensional matrices that show – for a given pair of attributes – the frequency of co-occurrence for each possible pair of values. WITT represents each category as a *set* of contingency tables, one for each possible pair of attributes.

[1] In principle, one could also incorporate higher-order correlations between three or even more features. However, for unsupervised learning techniques the memory and computational requirements of such an approach would be prohibitive.

Table 1. Instances and contingency tables for category (a) from Figure 1.

Color	Orientation	Shape
white	symmetric	circle
white	symmetric	square
white	asymmetric	square
black	symmetric	square

	Color× *Orientation*	
	black	white
symmetric	1	2
asymmetric	0	1

	Color× *Shape*	
	black	white
circle	0	1
square	1	2

	Orientation× *Shape*	
	symmetric	asymmetric
circle	1	0
square	2	1

Table 1 presents the three contingency tables associated with category (a) from Figure 1. The top of the table lists the attribute values for the four objects in the category. Because there are three attributes, three pairwise tables result – one for COLOR and ORIENTATION, another for COLOR and SHAPE, and a third for ORIENTATION and SHAPE. Because each attribute takes on two possible values, there are four cells in each table. Note that if a particular combination of values never occurs, the associated cell has a score of zero. Given N attributes, WITT requires $N(N-1)/2$ contingency tables to represent each concept. Such a correlational representation has considerable storage costs, but it is required if feature correlations are of interest for categorization. In our case, feature correlations/relations will be central to the processing stages of our model.

2.2 Evaluating alternative clusters

Natural categories do not exist in isolation. Instead, they are best viewed as contrasts; one category can be used to define another. For example the antonym "X is not a Y" defines a contrast between two categories. In general, people seem to construct categories that maximize similarity within categories and that concurrently minimize similarity between categories. In a similar spirit, WITT employs an information-theoretic metric called *cohesion* to evaluate potential clusters in terms of their between-group and within-group similarity. We define the cohesion C_c of a category c as

$$C_c = \frac{W_c}{O_c},$$

where W_c represents the within-category cohesion of the category and O_c represents the average cohesiveness between c and all other categories. Conceptually, cohesiveness (C_c) is a measure of the average distance between objects within a cluster (in a multi-dimensional space) relative to the average distance between that cluster and other clusters. This metric is not a Euclidean distance metric in which all features are summed independently, but rather a measure of distance in terms of relations between features as represented in contingency tables.

The within-category term W_c is the easiest to explain, so let us begin there. This term measures the average variance in the co-occurrences of all possible attribute-value pairs for a given category. We can state it formally as

$$W_c = \frac{\sum_{i=1}^{N-1} \sum_{j=i+1}^{N} D_{ij}}{N(N-1)/2},$$

where N is the number of attributes and D_{ij} stands for the co-occurrence distribution associated with the contingency table for attributes i and j. We define this term as

$$D_{ij} = \frac{\sum_{m=1}^{i} \sum_{n=1}^{j} f_{mn} log(f_{mn})}{(\sum_{m=1}^{i} \sum_{n=1}^{j} f_{mn})(log(\sum_{m=1}^{i} \sum_{n=1}^{j} f_{mn}))},$$

where f_{mn} is the frequency with which value m of attribute i and value n of attribute j co-occur. Recall that each contingency table consists of an $n \times m$ matrix, in which one attribute has n values and the other attribute has m values. The D_{ij} term involves summing over all $n \times m$ cells, giving an overall score for the table.

As an example, consider the frequencies shown in Table 2, which describes four different categories and their associated summary data. The same attributes (SIZE and SHAPE) are involved in each case, but the distributions are quite different. Table 2 (a) shows a situation in which all instances are both LARGE and SQUARE; thus, all cells but this one have frequencies of zero. This is an example of a conjunctive concept; in this case, the overall D score is one. Table 2 (b) portrays a different situation, in which all instances are LARGE, but three are SQUARE and one is a CIRCLE. This category can also be described in logical terms, but its D score is only 0.59.

The frequencies in Table 2 (c) present a case of polymorphy: two instances are large squares, one is a SMALL SQUARE, and one is a LARGE CIRCLE, but none are SMALL CIRCLES. The score in this situation is 0.25. Finally, consider the completely disjunctive data in Table 2 (d); in this case the D score is zero. Note that the D score decreases as one moves from the conjunctive category in (a) to the category in (b), where there is less dependency between the features. The score drops again as one moves to a polymorphous concept, and it

Table 2. Contingency tables illustrating the grading of the D metric for different types of concepts (conjunction, polymorphy, and disjunction).

(a) Two Common Features

		Size	
		small	large
Shape	circle	0	0
	square	0	4

D=1.00

(c) One Out of Two Polymorphy

		Size	
		small	large
Shape	circle	0	1
	square	1	2

D=0.25

(b) One Common Feature

		Size	
		small	large
Shape	circle	0	1
	square	0	3

D=0.59

(d) Total Disjunction

		Size	
		small	large
Shape	circle	1	1
	square	1	1

D=0.00

drops still further when one reaches the extreme case of a disjunctive category, where the features are independent. In summary, the D metric *prefers* logical, conjunctive descriptions if they exist, but it can still identify useful regularities when conjunctive descriptions are absent.

Recall that finding the within-group cohesion W_c involves computing the average D score for all contingency tables that are associated with category c. The between-category term O_c also requires averaging, but in this case across the cohesion between the current category and every other category. We are interested in the relative cohesion between two categories c and k

$$B_{ck} = \frac{1}{W_c + W_k - 2W_{c \cup k}} ,$$

which we will assume is related to the intersection of the two categories. The term B_{ck} measures the variance in the attribute-value co-occurrences in the union of the categories, relative to the variance in the isolated categories. Given this metric, we can define the average 'other-group' cohesion for category c as

$$O_c = \frac{\sum_{i \neq k=1}^{L} B_{ck}}{L - 1} ,$$

where L is the total number of categories. This term becomes the denominator in our overall measure of the cohesion $C_c = W_c / B_c$ of category c.

A geometric analogy may clarify the effect of these functions. As illustrated in Figure 2, imagine a set of objects that form natural clusters in a two-dimensional space. As more objects are added to a given cluster, the variance within the cluster increases (i.e., W_c decreases) and the relative variance between it and the other clusters decreases (i.e., O_c increases). There exists an inherent tradeoff between these two metrics, and the optimal clustering

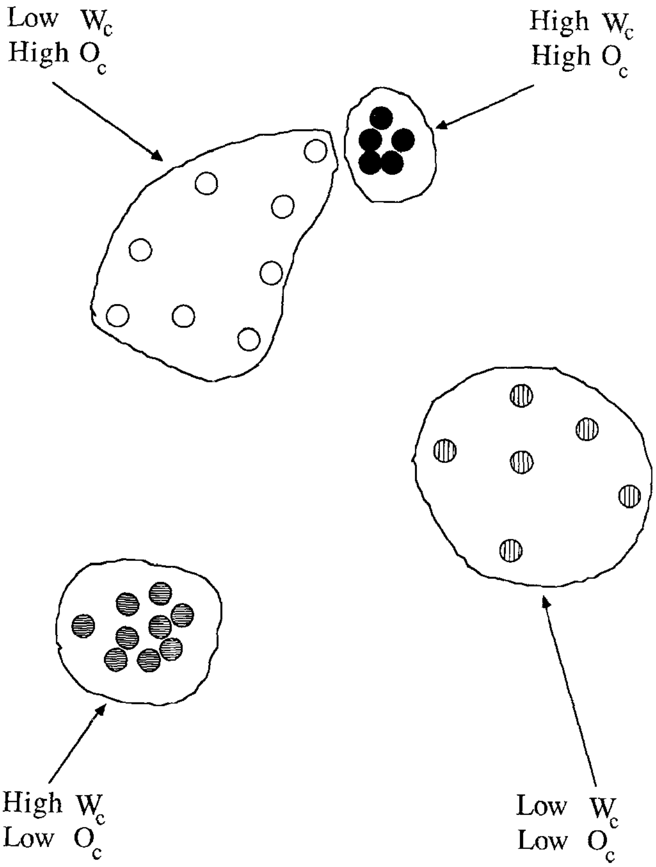

Figure 2. A geometric example indicating the possible category relations that might be encountered in a feature space. Dots shaded similarly show different objects belonging to the same categories. Categories are enclosed by a solid line indicating cluster boundary, spread, and shape. Four possible relations of within (W_c) and between (O_c) category similarity are depicted.

typically exists somewhere between the extreme values of these two functions. Visually, an ideal set of categories consists of several dense clusters far away from each other. This amounts to finding large sets of necessary and sufficient features for each category. Thus, our approach will find necessary and sufficient categorizations if they exist; however, because it considers "weaker" descriptions of possible categories in terms of feature intercorrelations, it can also find other interesting structure that may be useful for category formation.

2.3 Witt's clustering algorithm

In principle, given a set of N instances, one could exhaustively consider all partitions of those instances, compute the quality of each partition, and select the optimal one. However, the size of this space is exponential with

Table 3. Witt's preclustering algorithm.

```
1. Compute the distance between all objects in memory; set the smallest distance
   to $D$, and set the threshold $T_1 = F \times D$.
2. Select the closest pair of objects in memory;
   2.1 If their distance is greater than $T_1$, then halt;
   2.2 Else combine them into a cluster, replacing both objects with that cluster.
3. Compute the distance between the new objects and all remaining objects in
   memory and go to 2.
```

N. This makes an exhaustive approach impractical, and neither would it be psychologically plausible. Instead, we have employed a heuristic approach that is not guaranteed to find the optimal clustering, but which finds reasonable clusterings in most situations.

Witt's learning processes divide naturally into two stages – an initialization phase and a refinement phase. The first of these operates much like a standard numerical taxonomy method, as described by the algorithm in Table 3. This phase is necessary because the category evaluation functions described above operate upon categories, not upon single objects. Therefore one must first select a small set of tentative categories using what we call a *preclustering* algorithm. The algorithm forms these categories based on the densest regions of the instance space, forming groups that consist of the most similar objects. The preclustering algorithm uses an *information loss* measure that determines how much feature uncertainty is lost by removing objects from a set (cf. Lance & Williams, 1967; Orloci, 1969; Wallace & Boulton, 1968). The system first finds the two closest instances in the N-dimensional feature space and groups them together to form a cluster. It then finds the next two closest objects, either by creating a new cluster or by adding another instance to the existing cluster. If multiple clusters are created, Witt considers joining them as well.

Note that this 'preclustering' phase does not use the evaluation metric described in the last section; it considers only the distances between instances and clusters in terms of information loss. Because the information loss measure does not consider relations between features nor take into account the effect of adding an object to a cluster on the distances between that cluster and other existing clusters, this metric is not desirable as a category evaluation metric. However, it is useful for choosing an initial set of categories to which our cohesion metric can be applied.

The preclustering algorithm differs from numerical taxonomy schemes by employing a distance threshold (T_1) beyond which combinations are not made. This term is a function of two numbers – the distance D between the two nearest objects in the space multipied by F, a user-specified parameter. Thus, Witt continues to combine objects until the distance between all remaining objects is greater than $T_1 = F \times D$. If F is one, no objects further apart than the distance D will be joined. In the runs described below, F was set to

Table 4. WITT's refinement algorithm.

```
1. Compute the cohesion score $C$ for all pairs of unclassified instances and existing categories.
2. Select the instance-category pair with the highest cohesion score $S$.
3. If $S$ is greater than $T_2$, then add the instance to the category and go to 1;
4. Else call the preclustering algorithm on the unclassified instances to generate additional categories.
   4.1 For each new category $c$, if $W_{i\cup c}$ less than $T_3$ for all i in the set of existing categories then add $c$ to memory.
   4.2 If at least one category is added to memory, then go to 1.
5. Else compute the within-category cohesion score $W_{i\cup j}$ for all pairs of existing categories and select the pair with the highest score;
6. If this score is greater than $T_3$, then merge the two categories and go to 1; else halt.
```

1.5; this means the preclustering stage continued until the distance between all remaining objects was at least 1.5 times the smallest observed distance.

Once preclustering has finished, the memory contains some initial categories and many instances that have not yet been clustered. WITT then invokes its refinement algorithm, and this is where the cohesion metric and feature co-occurrences come into play. We have summarized this refinement stage in Table 4. The program begins by examining all unclassified objects, in each case computing the cohesion score that would result from adding that instance to each of the existing categories. The system selects the instance-category pair that maximizes the cohesion measure, and if the resulting score exceeds another threshold (T_2), it adds the object to the cluster. WITT then repeats this process, looking for the next best instance-category pair, adding the instance to that class, and so forth.

This process continues until the cohesion score for the best choice fails to exceed the T_2 threshold. This suggests that the existing structure of memory is inadequate, so WITT attempts to create one or more new categories to supplement the existing ones. In this case the system reinvokes the preclustering algorithm to formulate the best clusters from the unclassified instances. However, before adding these categories to memory, WITT ensures that these new categories occupy new parts of the object space to avoid creating a "new" category that really exists within an established category. It does this by calculating $W_{i\cup j}$ for all i in the set of previously existing categories and all j in the set of new categories, and then comparing each $W_{i\cup j}$ to another threshold, T_3. If $W_{i\cup j}$ is greater than T_3, it means that the jth new category does not occupy a new part of the space, but overlaps to some extent with the existing ith category. If one or more of the new clusters are sufficiently distinct, then the program returns to its normal mode of adding instances to existing categories.

However, if all of the proposed categories overlap with existing categories, WITT rejects the new categories and considers merging two existing categories instead. The system compares all pairs of clusters and attempts to combine the pair with the largest within-category cohesion, $W_{i \cup j}$. If this measure is above the T_3 threshold, the program merges the categories and returns to its normal mode of adding individual instances. On the other hand, if the largest $W_{i \cup j}$ is less than T_3, then none of the categories overlap. Hence, WITT tries no further merging. Any change in the category structure at this point would violate one of the thresholds; consequently the program halts and reports about the categories it has formed. It lists important feature relations, prints category members, and points out prototypical and atypical members of each category.

One can view WITT's clustering method in terms of a two-stage search. The initial preclustering stage employs a greedy algorithm to form initial categories; the evaluation function in this case is the distance between objects, and the termination condition involves the parameter F. The refinement stage is more complex, incorporating three separate operators: (1) adding an object to an existing cluster; (2) creating new clusters; and (3) merging two existing clusters. The evaluation function for this stage is the cohesion metric described earlier.

The parameters T_2 and T_3 are cohesion thresholds that, in part, determine which operator WITT will apply at each point. As T_2 is made larger and T_3 smaller (or as the ratio $R = T_2/T_3$ is made larger), the system requires the categories to have a higher degree of cohesiveness in order for the operators to be applied. In the extreme case, this means that all the categories' features must be necessary and sufficient. Conversely, as their ratio decreases, WITT's criterion for building categories becomes less stringent, allowing for varying degrees of polymorphy and finally complete disjunction. Thus T_2 and T_3 specify the levels of cohesiveness at which objects can be added to categories, new categories can be created, and categories can be merged.[2] The parameter T_3 also specifies when the clustering process should halt.

WITT's learning strategy embodies an important principle: that major reorganizations of category structure are rare events and should be undertaken with reluctance. Consequently, its control structure reflects this principle by progressively moving through a series of operators that in turn represent successively more category reorganization. Thus the second operator (creating a new category) is invoked when the existing category structure fails to accommodate a new object. Merging two or more categories requires an even greater reorganization and is only invoked when the first two operators fail. The thresholds T_2 and T_3 determine the sensitivity of each operator to the amount of reorganization.

Another possible control structure would involve applying the operator that maximizes the average C_c for the current set of categories. This is really a standard hill-climbing algorithm. However, because it makes no distinction about the amount of reorganization required for each operation, strict hill climbing is less psychologically plausible. We assume that complex operators should not be considered if simpler ones are sufficient, and the thresholds essentially

[2] In Section 3.2 we describe experiments that examine the effect of varying this ratio.

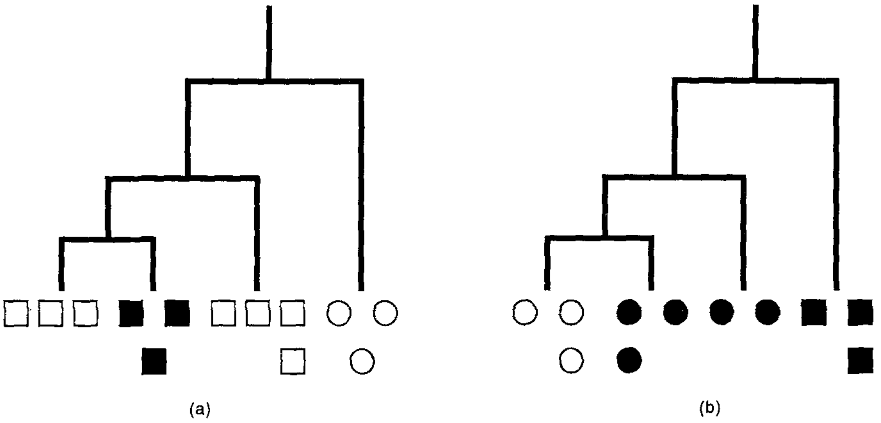

Figure 3. A trace of WITT's clustering behavior on the polymorphy example from Figure 1. The dendrogram shows the merge history of each object with other objects. Merges lower in the tree have smaller W_c and O_c ratios than those higher in the tree.

let the program make these judgements. This bias towards avoiding reorganizations until simpler operators are determined to be inadequate distinguishes our model from both Fisher's (1987) COBWEB and Lebowitz's (1987) UNIMEM, which give clustering and reorganization the same priority.

2.4 A simple polymorphy example

In order to demonstrate some aspects of WITT's behavior, we will describe a run on the two out of three polymorphy example shown earlier in Figure 1. The two dendrograms[3] in Figure 3 show the resulting clusters and the order in which they were constructed. WITT begins by isolating two contrasting groups, each consisting of two tokens. In this case the common features are SQUARE and SYMMETRIC in the first group (a) and CIRCLE and ASYMMETRIC in the second group (b), independent of the COLOR (BLACK/WHITE) attribute. This represents one of the largest possible Hamming distances that WITT can select, though other starting points are also possible. Next, the system places the SQUARE, WHITE, and ASYMMETRIC instance into group (a), after which it adds the BLACK, ASYMMETRIC CIRCLE in group (b). This balances and, in fact, increases the intercorrelation contrast between the two groups. Finally, the WHITE, SYMMETRIC CIRCLE is placed in group (a) and the BLACK, ASYMMETRIC SQUARE is placed in group (b), decreasing the cohesion within each group (W_c), but decreasing the cohesion between the two groups (B_{ck}) much more. Because there are no more objects to be clustered, the system halts at this point.

[3] A *dendrogram* is a rooted tree showing the order in which clusters are created. One can create a dendrogram for any type of clustering algorithm, whether agglomerative or divisive, including conceptual clustering algorithms.

By systematically increasing the ratio of T_2 to T_3 from 2.5 to 40, of the 4035 possible partitions of this data set, we found that the two groups shown in Figure 3 tended to occur for input threshold values near .9 for T_2 and .1 for T_3. This solution has the largest within-category cohesiveness and the smallest between-category cohesiveness of all the solutions found by WITT as we varied the two threshold parameters. The fact that the two out of three polymorphy solution has the highest cohesion suggests that it is not just one particular feature relation (feature to feature correlation) that supports the partitioning as shown in Figure 1, but rather is supported by all of the features (albeit with moderate strength) within each group, since that is what polymorphy requires.

Notice that in the clustering trace there is a strict path dependence; the presence of one instance in a category allows another instance to be added. The final structure is a set of intercorrelations that support each group. Although in this case the result is similar to that obtained by a common-features approach, for data that possess a complex intercorrelation structure, WITT will reveal many properties that a common-features approach like UNIMEM (Lebowitz, 1987) would miss.

2.5 An incremental algorithm

The above algorithm is nonincremental in that it requires all instances to be present at the outset. Because a nonincremental algorithm requires all instances to be held in memory simultaneously, it violates usual intuitions about the nature of human concept formation. Consequently, we have included a parameter in WITT that specifies the number of objects the system holds in memory during processing. If the parameter is set to the entire set of objects available for clustering, then the behavior described above emerges as a special case of this more general incremental version.

The procedure works by filling a short-term memory buffer by a random selection of K objects from a larger pool of objects available in the environment. WITT then operates as usual on the K objects that are presently available in the memory, ignoring objects not yet available for clustering. Once it has attempted to classify these objects, whether or not all have been successfully clustered, the system randomly selects another K objects from the $N - K$ left. These new objects are placed in the buffer with those remaining from the previous pass and WITT then proceeds as before. This process continues until all objects in the environment are incorporated.

In general, one would expect this memory-limited algorithm to produce clusters of lower quality than the nonincremental version because of biased sampling, and we will see some evidence to this effect later in the paper. However, we also expect that humans are subject to similar kinds of sampling errors; thus, they must either carry out a sampling procedure similar to that used by WITT or restructure their conceptual organization during clustering. This type of incremental restructuring would be similar to that used in Fisher's (1987) COBWEB system.

Table 5. Attributes and values extracted from the 1985 *World Almanac* concerning the nations of the world.

ATTRIBUTE	POSSIBLE VALUES
AREA	(HI, MID, LOW)
LOCATION	(NORTH AFRICA, INDOCHINA, ...)
INDUSTRIES	(IRON, CARS, ELECTRONICS, ...)
DEFENSE	(HI, MID, LOW)
CURRENCY	(DOLLAR, RIEL, KYAT, ...)
LITERACY	(HI, MID, LOW)
CHIEF CROPS	(GRAINS, WINE, POTATOES, ...)
MINERALS	(OIL, IRON, COAL, ...)
IMPORTS	(USA, FRANCE, ...)
EXPORTS	(USA, WEST GERMANY, ...)
TYPE	(REPUBLIC, COMMUNIST, ...)
LANGUAGE	(ENGLISH, FRENCH, ...)
RELIGIONS	(HINDU, CHRISTIAN, ...)
TELEVISION SETS	(HI, MID, LOW)
NATIONAL BUDGET	(HI, MID, LOW)
PER CAPITA INCOME	(HI, MID, LOW)
INFANT MORTALITY	(HI, MID, LOW)

3. Evaluating WITT

Having described WITT in some detail, we can now evaluate the model both in terms of its clustering behavior and its psychological plausibility. We have already examined the system's behavior in a simple, artificial domain, but below we consider its response to a more complex naturalistic data set. We then report some empirical studies of WITT's behavior with different settings of its parameters, and we follow this with a discussion of the learning algorithm's computational complexity. Finally, we consider the model's ability to account for two well-established psychological effects – typicality and the existence of basic-level categories.

3.1 Clustering nations

Although WITT behaves quite well on simple, artificial data such as those shown in Figure 1, we felt the need to test the system on a larger, more complex domain with a real-world flavor. We decided on data involving the nations of the world as described in the 1985 *World Almanac.* From this source we arbitrarily selected 37 countries, with the constraint that they covered most continents and that there existed a wide range of variation. This set was large enough to make the clustering problem nontrivial, but still small enough that we could examine the system's behavior in detail.

We should stress that we were not looking for one particular "right" set of categories, but rather for any organization that was comprehensible and provided a sensible hypothesis about the group membership based on feature sets. Higher-order descriptions such as "super-powers," "third world," and "poor but technologically advanced" are unlikely to arise by chance organizations alone.

To create the data set, we extracted triples of the form (COUNTRY ATTRIBUTE VALUE) were extracted from a machine-readable version of the 1985 *World Almanac.* Quantitative attributes such as "percent literacy of population" were automatically transformed to ordinal values by examining the frequency histogram of the attribute and looking for modes in the data that would suggest breaks into group such as "high," "middle," and "low." Approximately half the variables were quantitative and transformed accordingly.[4] WITT uses only the nominal value for such attributes, ignoring possible order information, although the value is reported along with the feature correlations. For each country, we extracted 17 multivalued features, which are shown in Table 5 along with their descriptions. They consist of attributes such as DEFENSE, RELIGIONS, and INFANT MORTALITY.

Figure 4 shows the dendrogram for the 37 nations of the world generated by the nonincremental version of WITT. For each category, the dendrogram shows the order in which countries were added, the order in which clusters were merged to make up the category, and the distance between the objects at the time of the join. For example, for the category at the top of the page, the program first created the three clusters SPAIN and ITALY, UNITED KINGDOM and FRANCE, and USA and CANADA. Next it merged SPAIN and ITALY with UNITED KINGDOM and FRANCE, and then started adding countries to this merged group. EAST GERMANY was added first, followed by BELGIUM, WEST GERMANY, and DENMARK. Finally the system merged USA and CANADA with the rest of the group. The distance between two objects or groups at the time of their joining is represented by the distance of their connection from the right-hand margin. Thus, USA and CANADA are much closer to each other than they are to the the rest of the cluster. Similarly, USSR, POLAND, CONGO, and JAPAN comprise a category that has much lower within-category cohesion (W_c) than the European cluster, so their connection is much further from the right-hand margin.

Examination of the dendrogram reveals that WITT did discover reasonable and comprehensible groupings. At the top of the dendrogram, the first subcluster of two countries includes USA and CANADA, and this set seems to be separated from a European subcluster that includes ITALY, SPAIN, UNITED KINGDOM, and FRANCE. This cluster is distinct from a larger cluster that starts near the bottom of the dendrogram with CAMBODIA, VIETNAM, and BANGLADESH.

The structure of the dendrogram can be broken into seven or eight groups. At the highest level we see a split between countries that may be described as "third world" and technologically advanced countries with a relatively high quality of life. Finer distinctions emerge as we examine the clusters indi-

[4] The technique we used to do this is based on a unidimensional clustering method that seeks modes or clumps in a single dimension. Although in general more sophisticated techniques may be used to determine whether categorical breaks exist, in this case we forced the data into three categories, possibly producing some misrepresentation of the continuous variation. A better method would involve clustering the data and finding clumps in the continuous variables simultaneously, but we did not examine this approach.

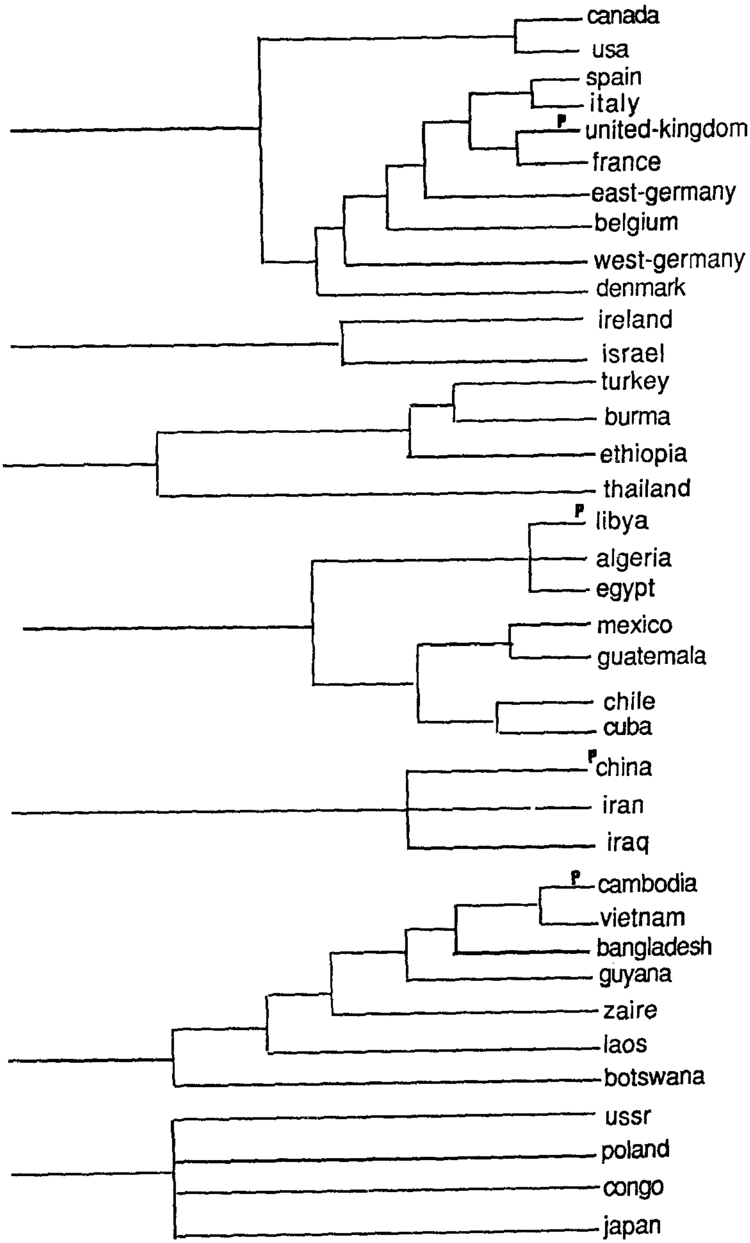

Figure 4. A trace of WITT's clustering behavior on the nations of the world data. The dendrogram shows the merge history of each nation with other nations. Merges lower in the tree have smaller W_c and O_c ratios than those higher in the tree. The letter 'P' in the dendrogram indicates the country chosen by WITT to be closest to the prototype of the category.

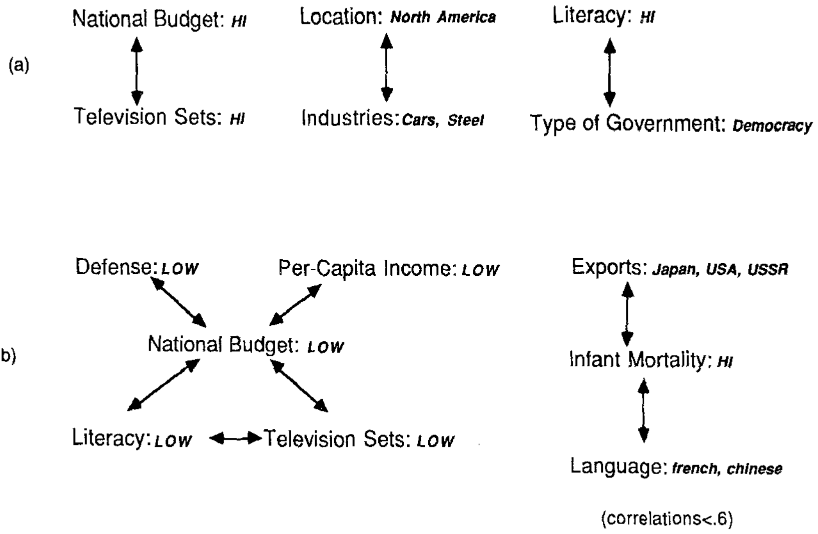

Figure 5. Correlations within the two strongest clusters from Figure 4. Cluster group (a), including USA and Canada, has strong correlations between NATIONAL BUDGET and TELEVISION SETS and between LOCATION and INDUSTRIES. A correlation less than 0.6 is shown between LITERACY and TYPE OF GOVERNMENT. Cluster group (b), including Vietnam, Cambodia, Bangladesh, and Guyana, has strong correlations between NATIONAL BUDGET, TELEVISION SETS, LITERACY, DEFENSE, and PER CAPITA INCOME. A correlation less than 0.6 is shown between EXPORTS, INFANT MORTALITY, and LANGUAGE; all others are greater than this amount.

vidually, including a cluster of European countries and another composed of Southeast Asian and African nations. However, geography seems less central to these finer groupings than the abstract qualities of economy, quality of life, and industries. We turn next to a specific look at some selected clusters.

Figure 5 shows two groups at the lowest level of the tree – (a) the North American cluster and (b) the Southeast Asian cluster. The figure specifies some of the feature relations Witt used to characterize each cluster.[5] Three sets of correlations summarize the North American group, each related to a high quality of life. The first of these is the high correlation between the number of television sets and the size of the national budget. The second is between location in the world and industries. Finally, there is a high correlation between the literacy of the population and the type of government.

In contrast, the Southeast Asian group contains many features that might indicate a low quality of life. Economic factors seem to be an important attribute of the cluster. A low national budget is correlated with low per capita income. These are both correlated with other variables like literacy, defense,

[5] For the sake of clarity, the figure highlights those pairs of features with correlations greater that 0.6, showing only a few with lower correlations on the bottom right. However, Witt represents and uses all correlations throughout all stages of clustering.

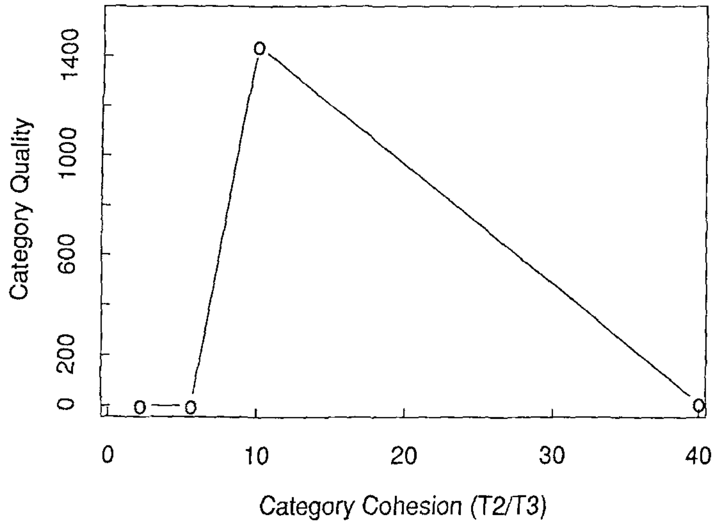

Figure 6. The quality of WITT's clusterings as a function of the category cohesion parameter (T_2/T_3) on the polymorphy example from Figure 1. Note the U-shaped curve, which suggests that there may be an optimal setting for clustering parameters.

and number of televisions. The interrelations among this mix of economic, governmental, educational, and technological variables help characterize the group. Another significant set of correlations involves infant mortality, exports, and language. Such covariates may implicate not only quality of life but hospital care, local health conditions, and education.

In each of these configurations of correlations, there may be a causal account or "story" that explains the indicated correlations. We call this the *proto-explanation hypothesis.* We hypothesize that discovering the correlational configuration among features constitutes an initial step towards constructing such causal explanations and further elaboration of categories.

3.2 Sensitivity of WITT's parameters

WITT contains three main parameters: the factor F used in the preclustering stage and the thresholds T_2 and T_3 used in the refinement stage. The incremental version includes another parameter, the size of memory. The values of these parameters clearly affect the system's behavior, and we have carried out experiments in order to better understand this effect. In the first study, we varied the ratio of T_2 to T_3 (R), and examined the quality of the resulting categories. We measured category quality as the ratio of mean category cohesion (C) for each category and the variance of these cohesions across the categories. Intuitively, this measure is sensitive to both the within-category and between-category structure, relative to each category. High values indicate cohesion scores for each category that are relatively large and similar. Figure 6 shows the quality scores obtained when we varied the ratio R and ran WITT on the

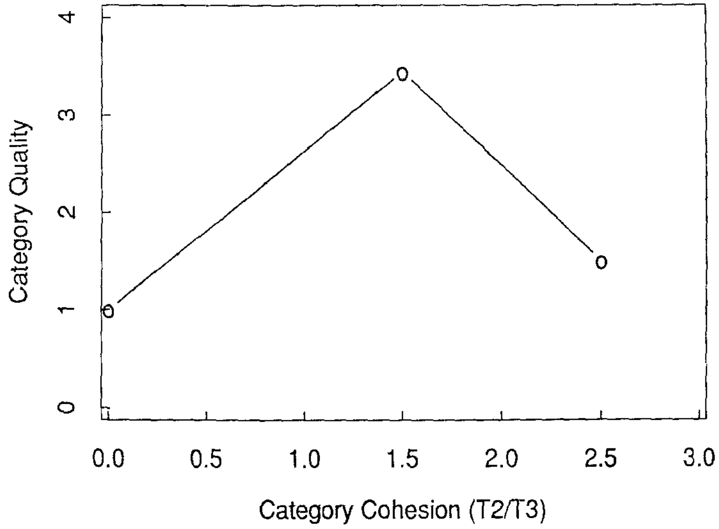

Figure 7. The quality of Witt's clusterings as a function of the category cohesion parameter (T_2/T_3) on the nations of the world data. Again, the presence of a U-shaped curve suggests an optimal setting for clustering parameters.

polymorphy data from Figure 1. The curve is more or less U-shaped, with the highest score (1440) occurring when the ratio was set to ten. In fact, this two-cluster partition is the same as that shown in Figure 1; it constitutes one of four optimal clusterings for these data. Lower settings of R led to partitions composed of three clusters with lower cohesion scores, whereas higher settings also led to lower scores.

We also carried out an analogous study using the nation data described in the previous section. Figure 7 graphs the ratio R values against the quality scores that resulted in this case. Again the curve is roughly U-shaped, with the best cohesion scores over categories occurring in the middle range. The peaks in Figures 6 and 7 suggest that for both data sets there is an optimal setting for the ratio R that results in an optimal set of categories. This setting will probably depend on the feature relations, but this is an empirical question. Future experiments should examine how the peak value varies as a function of the *a priori* feature structure of a data set.

The final experiment examined the effect of limiting the memory for instances of the incremental version of Witt. Figure 8 shows the effect of varying the memory size while holding the ratio R constant (1.0) on the clusters produced for the nation data. As one would expect, severe limits on memory cause significant degradations in the category quality. However, the system does not appear to need *all* instances present in memory to respond adequately: a memory size of ten led to cohesion scores close to those obtained in nonincremental mode, and we expect further increases in memory size would produce clusterings similar to the nonincremental version.

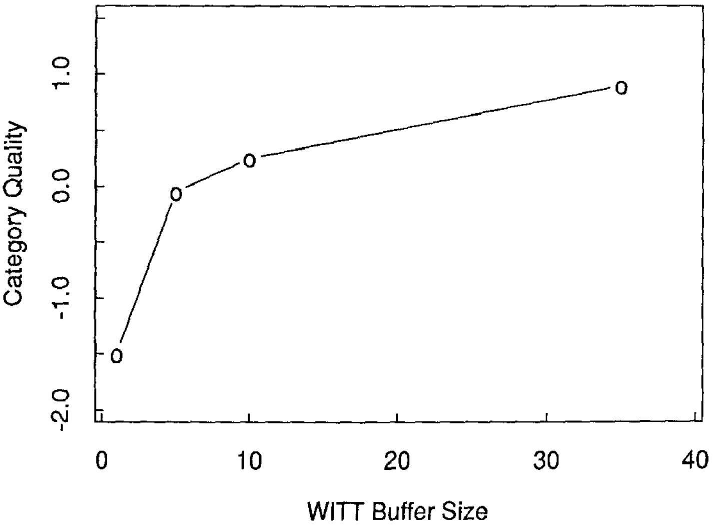

Figure 8. The quality of WITT's clusterings as a function of buffer size. As the size of memory is increased from one to 35 items, category quality improves significantly.

Examination of WITT's traces clarified the reason small memory sizes (from one to three) produced such poor clusterings. Because only a few objects were available for inspection at any given time, the system created categories with little difference in their within-category and between-category cohesiveness. This massive overlap in the initial categories kept WITT from applying operators in ways that would create reasonable contrasts between categories.

3.3 Computational complexity of WITT

Most unsupervised learning schemes tend to fare poorly on time or space requirements, due to the large number of possible partitions they must consider. Furthermore, WITT's use of feature correlations introduces additional complexity. As a result, the complexity of the algorithm is an important concern. In this section we examine the worst-case behavior of WITT and its component processes, along with the average-case behavior we have observed in actual runs.

If the preclustering algorithm were allowed to run to completion (i.e., continue running the procedure until it had clustered all the objects), its worst case asymptotic time complexity would be $O(N^3 f)$, where N is the number of objects and f is the number of attributes. Space complexity for preclustering is $O(Nf)$, because all the objects are held in memory at once. However, for most possible distributions of objects in feature space, the algorithm makes a small number of passes before satisfying the first threshold, T_1, and time complexity for preclustering is typically $O(N^2 f)$. Space complexity for the rest of the procedures is $O(Nf + m^2)$, where m is the maximum number of feature values of any attribute. Although all the objects are held in memory, the correlations are calculated iteratively for each category-object and category-category comparison, so that the algorithm need never hold all possible correlations in memory at once. Empirically, the space WITT requires grows linearly with the number of objects to be categorized.

In terms of time complexity, WITT is really no worse than most nonincremental clustering schemes. Let us consider each operator in turn. First, the operator for adding an object has a time complexity of $O(NCf^2)$, where N is the number of objects under consideration, C is the average number of categories, and f is the number of features. Next, the operator for creating a new category has time complexity $O(N^3f + CC_nf^2)$, where C is the average number of existing categories and C_n is the average number of categories created by the operator. Like the preclustering algorithm, this makes only a few passes in creating new categories and thus typically takes $O(N^2f + CC_nf^2)$. Finally, the operator for merging categories has a time complexity of $O(C^2f^2)$.

Because the functions are additive for time complexity, worst-case behavior of the algorithm is $O(N^2Cf^2 + NC^2f^2)$, where C is a function of the regularities in the data set and the input thresholds. Worst-case behavior is not particularly revealing, because empirically the time complexity varies widely with the correlational structure of the input. WITT's bias towards adding objects to existing categories tends to make the algorithm conservative about creating new categories and merging existing categories. Thus, if the category structure is apparent from a relatively small fraction of the objects, the effective complexity is $O(N^2Cf^2)$ because there is little need for the second and third operators.

The incremental version considers only a constant number of objects at a time, thus reducing time complexity to $O(NCf^2)$, which is linear with the number of objects. This method is faster than most numerical taxonomy algorithms (e.g., complete-linkage clustering is $O(N^3f)$) and comparable to Fisher's COBWEB (1987). WITT requires a factor of f^2, whereas most other algorithms require f. This results from calculating correlations across features, whereas most other metrics make only single-feature comparisons.

3.4 Relation to human categorization

An important finding in cognitive psychology is that people tend to partition groups of objects such that within-group instances have more common features then between-group instances (Rosch & Lloyd, 1978). Of many possible partitions, people tend to prefer the one that produces the greatest predictability of categories from features, the greatest between-group discriminability, and the greatest group density of features. In a hierarchy, this partition set corresponds to a level that is more fundamental or more *basic* (Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976) in terms of reference and organization. For instance, the category *robin* is more specific than *bird*, and *bird* in turn is more specific than *animal*. For many inferences (such as 'Does X fly?'), the concept *animal* is too abstract and the concept *robin* is overly specific. The intermediate level (*bird*) provides most of the information needed for everyday reasoning without including more than necessary.

The metric and the control structure used by a conceptual clustering algorithm to determine category quality can be consistent or inconsistent with this basic-level effect. For example, if the metric predicts high between-category discriminability at the middle of the taxonomy, then it is consistent with basic-level effects in humans. Since WITT halts once it detects a significant increase in between-group discriminability, it has no way of knowing if there are larger values of the metric further into the taxonomy. Nonetheless, this assumption is consistent with the basic-level hypothesis, which requires a *non-monotonic* category discriminability metric throughout the taxonomy.

Such metrics can be contrasted with measures based on the conditional probabilities of features given categories or the conditional probabilities of categories given features, which one can show to be monotonic over taxonomic level (Murphy, 1982). Consequently, frequency-based or probabilistic approaches like Lebowitz's UNIMEM (1987), Cheeseman, Kelly, Self, Stutz, Taylor, and Freeman's AUTOCLASS (1988), and Quinlan's (1986) decision-tree clusterers will not be consistent with basic-level effects, unless their metrics are sensitive to within-category and between-category variance as a function of the level in the hierarchy.

Recent explanations of basic-level effects involve combinations of conditional probabilities (Gluck & Corter, 1985) that are sensitive to feature predictability both within and between categories. Fisher's (1987) COBWEB system uses such a measure, making it sensitive to the basic level in the taxonomy. WITT's correlation measure (in particular its category cohesion score) is also non-monotonic over taxonomic levels, and tends to be more sensitive to the density of feature-to-category predictability within levels (even applying it independently of its control structure). At present there are not enough data to distinguish these various measures, but it is important that conceptual clustering metrics be sensitive to feature predictability over taxonomic level.

A simple way to assess basic level within a set of instances is to ask subjects to perform a free sort of the instances into a set of disjoint categories. This type of partitioning is equivalent to a cut through the concept hierarchy, thus producing a preferred level of categorization. The output of any conceptual clustering system that produces a set of disjoint categories can be compared directly to such free sorts, providing a straightforward test of its ability to generate the basic level.

To provide this comparison for WITT's behavior, we asked a group of human subjects to sort the nations of the world into comprehensible categories. Both the number of categories and the overall group structure is similar to WITT's clustering. Figure 9 shows the 'averaged' partitions of ten subjects who were given either just the nation labels (e.g., FRANCE), the attribute values for that nation, or both. We first analyzed subjects individually, finding few differences across the three experimental conditions. We then collapsed their partitions to obtain a co-occurrence matrix, in which each cell specified the number of subjects who agreed that nation i belonged with nation j. We then used a single-linkage clustering method to construct a dendrogram from this co-occurrence matrix, giving a tree structure that represented the agreement of all the subjects on the pairwise similarity of the nations (cf. Miller, 1971).

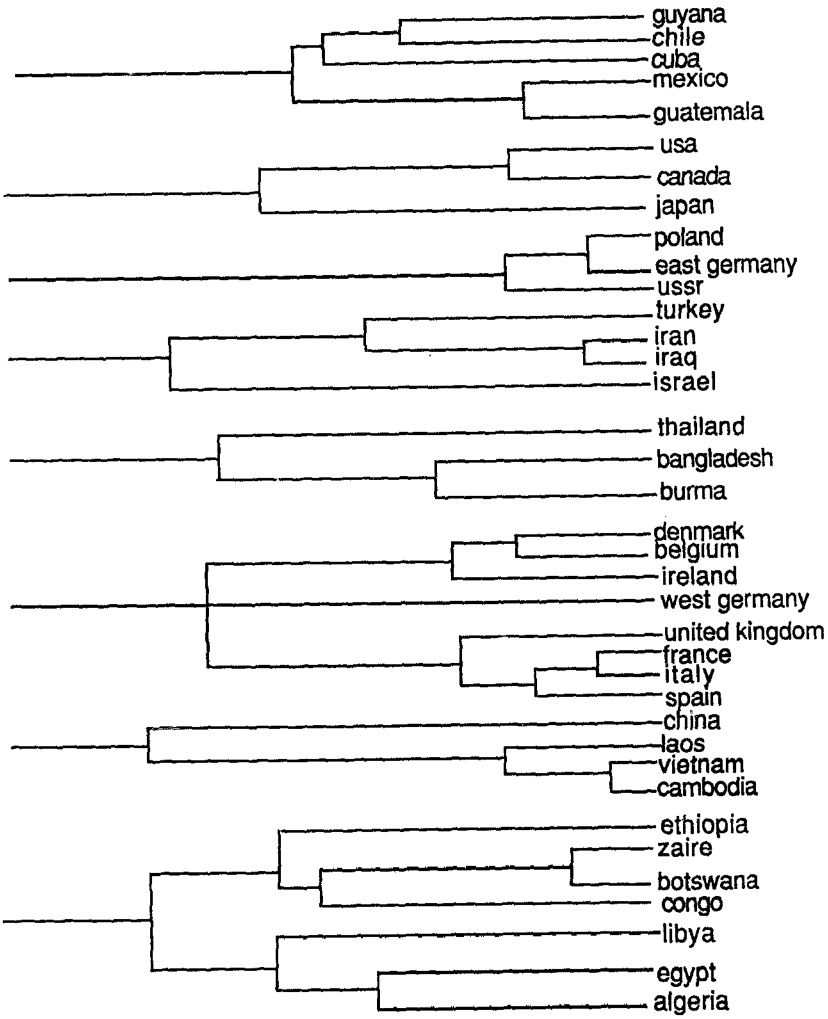

Figure 9. A dendrogram based on ten subjects' sorts of world nations. Note the similarities to WITT's clustering of nations shown in Figure 4.

On the average, the subjects sorted nations into eleven categories, whereas the tree shown in Figure 9 suggests a cut at about seven to nine clusters. Overall, the main partitions seem very similar to WITT's groups, including a category for Europe, another for USA and CANADA, and many of the same third-world groups. Some clusters that WITT missed involved more complicated political features (USSR, POLAND, EAST GERMANY) and simpler clusters motivated by geography (ZAIRE, BOTSWANA, CONGO).

As noted before, the *number* of clusters actually reflects a judgement about the basic level for the data. For example, if we assume the concept hierarchy is a standard tree structure, then using only two large clusters to characterize the instances suggests the basic-level cut is near the top of this concept hierarchy.

In contrast, if subjects used thirty small clusters to characterize the data, this suggests the basic level is deep into the concept hierarchy. In the present case, the number of clusters suggested by the subjects' dendrogram was about eight. Recall that WITT halted its clustering at seven clusters. Since both subjects and WITT sorted the objects into roughly the same number of clusters, they have taken a similar cut in the concept hierarchy, giving similar basic levels.

For humans, a category is more than a set of criteria that determine whether an object is a member. People can also decide how representative a particular member is of a category. In other words, there is structure within a category, with some exemplars being judged as more representative than others. For example, most people would say that a robin is a more typical bird than is an ostrich. This grading within a category is usually referred to as *typicality.*

By using contingency tables to represent categories, one can define metrics that determine the typicality of their members. One way to measure a member's typicality is to see the effect on category structure when it is removed. Within WITT's framework, we define typicality as the change in the category's cohesion when the object is included in the category and when it is not. An object that strongly supports the category's featural relations will cause a large drop in cohesion when removed and will thus have high typicality. In contrast, an object that does not change the cohesion when removed or actually causes it to increase will have low typicality. One can use this scheme to identify the *prototype* (the most typical instance) for each category, as shown in Figure 4.

4. Discussion

In this paper, we have focused on human categorization abilities and the constraints they impose on conceptual clustering. We view these abilities not so much as limitations but as guidelines for a clustering system to follow. Existing tools for knowledge representation do not sufficiently constrain the structure of categories. It is not enough to allow disjunction or probability operators into the concept description space. As we have tried to demonstrate, the exact nature of the category structure is important to specify – the similarity between exemplars matters, the relation between clusters matters, the typicality of exemplars matters, and the aggregate measure (e.g., kind of prototype) within the clusters matters. Decisions concerning all of these structural properties of categories will affect the clustering process, the categories recovered, and the usefulness of the categories for accommodating new information while maintaining important relations that already exist between exemplars.

The choices we made in designing WITT are based on the four heuristics about categorization listed earlier. These are drawn from the psychological literature, and similar ones have been adopted by other researchers who model induction (e.g., Grossberg, 1976). We treat these heuristics as constraints on the possible metrics and control structures that might appear in a conceptual clustering program. Of course, it is possible to construct other metrics and control structures that use the heuristics. However, we do want to argue that category contrasts (competitive induction) and feature interrelations are very important for inducing category structure. We feel it is more important to intelligently constrain the possible categories than to create another representation language with novel and potentially powerful, but unconstrained, features.

4.1 Comparison to numerical taxonomy algorithms

Too often claims are made about conceptual clustering results without comparison to optimal or normative results. Michalski and Stepp (1983b) established a precedent for comparing conceptual clustering results with numerical taxonomy or statistical clustering models. Interestingly, although they found their CLUSTER/2 system produced clustering results that were more similar to human experts', a numerical clustering analysis run on the same data misclassified only two of the cases and also provided clusters similar to those of experts.[6]

Short of examining all the possible clusterings of objects, it is difficult to establish that an algorithm has found optimal categories, but it is worth knowing that a method is different from standard clustering methods. The latter use simple metrics and simple group membership rules that have been established as normative in the statistical and numerical taxonomy literature over the last 30 years. In fact, there are a large number of possible clustering models, including some that are based on correlations (cf. Sneath & Sokal, 1973). However, very few consider between-feature correlations and none use the correlation metric to contrast within-category and between-category coherence.

In order to provide a contrast with WITT's results, we ran the nation data through a standard single-linkage clustering method (Everitt, 1974), using a Hamming distance measure. This is one of the simplest types of agglomerative clustering methods. For a set of objects defined on binary vectors, the Hamming distance is computed for each pair of objects and the pair closest on this measure are joined. Distances are recalculated for this new group, with distance being determined by comparison to the nearest neighbor in a group.

Several aspects of the clusters resulting from the single-linkage run on the nation data are worth noting. First, there was what statisticians call a "chaining" effect (typical for single-linkage methods), in that nations were generally seen as very close to one another. In other words, it was difficult for the single-linkage technique to discriminate between sets of nations; it tended to overgeneralize. Also, if the dendrogram is cut in order to form eight groups similar to the subject sortings in Figure 9, the single-linkage clusters indicate one large group with thirty nations and seven singleton groups. Overall, its metric seems to behave very differently from that used by WITT.

Other standard clustering techniques could have been chosen for comparison, but single-linkage uses the least amount of aggregate (or "abstracted") information about clusters. Thus, it provides a plausible *baseline* to contrast conceptual clustering results that typically invoke hypotheses about complex aggregate information in each cluster. Alternative clustering methods may in fact produce clusters similar to WITT's, but to try the thousand or so possible

[6] However, the statistical technique did not provide a summary description of each cluster, which could provide a measure of comprehensibility.

clustering schemes in order to find a similar outcome for this particular data set does not seem very useful. Various parameters could be also added to statistical clustering techniques for determining the number of clusters and cluster density, but this would begin to resemble a conceptual clustering approach and, in any case, such parameter additions cannot be made arbitrarily.

4.2 Relation to other conceptual clustering research

In this section we will compare our approach to several other conceptual clustering methods, including CLUSTER/2 (Michalski & Stepp, 1983a), UNIMEM (Lebowitz, 1987), and COBWEB (Fisher, 1987). There are at least three dimensions on which conceptual clustering approaches can differ, and these will help highlight some of the heuristics upon which we had earlier predicated our model. The first dimension we consider is the similarity metric used to represent the "closeness" of the objects to be clustered (heuristic 3). The second dimension involves the way one represents category structure, i.e., the intensional description of the category (heuristics 1, 2, and 4). Finally, a third dimension concerns the relation of categories to each other; that is, whether concepts arise due to the within-category similarity or due to a tradeoff of within-category and between-category similarity (heuristic 4).

Presently, most conceptual clustering programs use either a common-features measure or a probabilistic measure over common features. For example, Michalski and Stepp (1983a) calculate a metric similar to an information measure based on the probability that feature Booleans do not intersect (related to their "sparseness measure") while covering a set of events or objects. This kind of metric trades off within-category similarity and between-category similarity (heuristic 4), but it does so by maximizing the measure in terms of features that are common to the cover of a set objects. Their model does not refer to prototypes or category structure, although one could establish such measures within their multi-variable logic representation.

Lebowitz's (1987) UNIMEM is similar to statistical clustering approaches that employ a Hamming distance measure of similarity and use cluster centroids to represent group membership. Categories in UNIMEM are established as within-group similarity structures and have graded group membership status. The system incorporates a notion of central tendency or prototypicality, based on a frequency count of features in examples it has seen most recently. Consequently, this type of clustering method is a close relative to statistical clustering methods, and it conflicts with many of the heuristics we have argued are important for conceptual clustering. In particular, there is no explicit mechanism to cause a tradeoff of within-cluster and between-cluster similarity, and there is no attempt to reference feature relations.

Finally, Fisher's (1987) COBWEB model employs a more sophisticated metric that is based on work by Gluck and Corter (1985) concerning the nature of basic-level categories. This similarity metric, called *category utility*, can be viewed as a common-features measure that weights features relative to how well they discriminate one category from others. This is definitely related to the notion of "category contrast" in our fourth heuristic. COBWEB's measure also has the laudable effect of choosing objects that will tend to group together with the highest feature discriminability, producing (if the hierarchy is cut) a basic-level effect. The system does not directly reference feature relations, nor does it seem to represent category structure directly (e.g., prototypes), but it could given the nature of its similarity metric.

In general, it is worth noting there is a lot of similarity between all three of these approaches and the present model. One distinctive aspect of our approach is our explicit focus on feature relations and the use of such relations for category formation, retrieval, and explanation. Nonetheless, each of the above approaches incorporates at least one of the heuristics we laid out as potential constraints on conceptual clustering.

4.3 Future research

Although we believe WITT has advantages over most earlier models of the category formation process, more work remains to be done. There are at least four areas we feel need to be explored in future research. The first area involves the use of feature correlations in representing category structure and the relationship of category structure to its function or use. For example, some AI researchers have argued that clustering systems are not useful because they are not associated with goals, plans, or expectations about the world (cf. Schank, Collins, & Hunter, 1986). They claim that background knowledge is necessary for creating and using categories, since such knowledge can help determine which features are relevant in a given situation.

For example, feature selection that is motivated by the *function* of the category can produce very different categories and category structure than does a clustering scheme that uses unweighted features. However, in humans, background knowledge does not magically appear; it too must be learned. Selecting feature weights through goals or domain knowledge should first be based on the intrinsic feature structure available in the objects; further weights must be learned from experience with the categories in such domains. The approach we have taken leads naturally to different weights on features, and these could be used both in retrieval and in the creation of typicality judgements. They would also allow future versions of WITT to make predictions about the values of missing attributes and eventually lead to methods for the construction of schemas.

A second area of research involves the nature of explanation. Throughout this paper, we have claimed that discovering correlational structure is one important mechanism for creating conceptual structures. Of course, correlation is not causality, but we claim it is an important first step. Such correlational structure forms an initial set of hypotheses about possible explanations that are available in data. One might call this correlational structure a "proto-explanation," even if it is initially biased or incorrect. Feature-relation approaches can help discern important features and important constellations of features by determining which features are more involved in the structure of the concept. For example, the feature PER CAPITA INCOME in the "third-world country" group is more highly correlated with other features than is CURRENCY, and thus provides more support for the concept description.

This correlational knowledge will act in two ways. First, this feature will predict (or reduce the uncertainty about) other features and will have more effect than other features in future clustering choices. More importantly, the feature PER CAPITA INCOME will act with other features to make predictions about the concept space. If a new nation without a value for PER CAPITA INCOME is considered for inclusion, other features that are inter-correlated with it, such as INFANT MORTALITY, can act as probabilistic "indices" to let this new nation inherit the prototypical value for per capita income. In the other direction, the cluster will "expect" to find new nations for inclusion that possess a constellation of features that help reduce uncertainty about the existing inter-correlated features. One can imagine an activation mechanism that would activate several inter-correlated features once one of them had been seen.[7] For example, attributes of earthquakes could activate other attributes such as size, location, and deaths, through the correlational structure that exists among these features. Thus, correlational structure can be used as an important source of information in constructing conceptual representations, such as schemata and explanations about the domain.

A third direction for extending the current model is related to the assumption of non-overlapping categories. Clearly, not all human concepts obey this constraint, and there is no inherent reason why an instance could not be placed in multiple categories, provided this strategy led to improved cohesion scores. In this regard, we also plan to explore alternative metrics for cluster quality, evaluating them by their computational behavior and by their ability to explain psychological phenomena.

Finally, we intend to examine connectionist approaches to conceptual clustering and their relation to WITT. The connectionist framework supports both polymorphous concepts and the correlational representation that is central to our approach. As in the present model, this will require some judicious heuristics that introduce constraints on the possible outcomes of categorization.

4.4 Summary and conclusions

To summarize, WITT is a model of human concept formation that makes a number of important assumptions. Rather than relying on "logical" definitions of concepts, the system describes categories in terms of feature relations. This lets the model represent and acquire concepts that cannot be defined in terms of necessary and sufficient conditions. In discovering such polymorphous concepts, WITT employs an information-theoretic metric to direct its search through the space of clusters. During this search, the system prefers to add instances to existing clusters, resorting to more drastic measures like category creation and cluster merging only when the simpler strategy fails. The model is consistent with both basic-level and typicality effects that have been observed in humans, and we have tested the system in both artificial and natural domains, exploring the effect of various parameter settings. We believe that WITT embodies important heuristics for conceptual clustering that are closely related to the process of human categorization.

[7] This is also an important property of most connectionist approaches (Rumelhart & McClelland, 1986).

Acknowledgements

We would like to thank R. B. Allen, G. W. Furnas, G. A. Miller, K. Miller, and D. Walker for commenting on an earlier version of this paper, and R. A. Amsler for implementing a program that extracted the data on nations from machine-readable text. We would also like to thank Pat Langley and the reviewers for many useful comments.

References

Cheeseman, P., Kelly, J., Self, M., Stutz, J., Taylor, W., & Freeman, D. (1988). *Proceedings of the Fifth International Conference on Machine Learning* (pp. 54–64). Ann Arbor, MI: Morgan Kaufmann.

DeJong, G., & Mooney, R. (1986). Explanation-based learning: An alternative view. *Machine Learning*, *1*, 145–176.

Dennis, I., Hampton, J. A., & Lea, S. E. G. (1973). New problem in concept formation. *Nature*, *243*, 101–102.

Estes, W. K. (1986). Memory storage and retrieval processes in category learning. *Journal of Experimental Psychology: General*, *115*, 155–174.

Everitt, B. (1974). *Cluster analysis.* London: Heinemann Educational Books.

Gluck, M. A., & Corter, J. E. (1985). Information, uncertainty, and the utility of categories. *Proceedings of the Seventh Annual Conference of the Cognitive Science Society* (pp. 283–287). Irvine, CA: Lawrence Erlbaum.

Fisher, D. H. (1987). Knowledge acquisition via incremental conceptual clustering. *Machine Learning*, *2*, 139–172.

Grossberg, S. (1976). Adaptive pattern classification and universal recoding. Part I: Parallel development and coding of neural feature detectors. *Biological Cybernetics*, *23*, 121–134.

Homa, D. (1978). Abstraction of ill-defined form. *Journal of Experimental Psychology: Human Learning and Memory*, *4*, 407–416.

Lance, G. N., & Williams, W. T. (1967). Note on a new information-statistic classificatory program. *Computer Journal*, *9*, 373–380.

Lebowitz, M. (1987). Experiments with incremental concept formation: UNIMEM. *Machine Learning*, *2*, 103–138.

Medin, D. L., Wattenmaker, W. D., & Hampson, S. E. (1987). Family resemblance, conceptual cohesiveness and category construction. *Cognitive Psychology*, *19*, 242–279.

Michalski, R. S. (1980). Knowledge acquisition through conceptual clustering: A theoretical framework and an algorithm for partitioning data into conjunctive concepts. *International Journal of Policy Analysis and Information Systems*, *4*, 219–244.

Michalski, R. S., & Stepp, R. E. (1983a). Learning from observation: Conceptual clustering. In R. S. Michalski, J. G. Carbonell, & T. M. Mitchell (Eds.), *Machine learning: An artificial intelligence approach.* Los Altos, CA: Morgan Kaufmann.

Michalski, R. S., & Stepp, R. E. (1983b). Automated construction of classifications: Conceptual clustering verses numerical taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *5*, 396–410.

Miller, G. A. (1971). Empirical methods in the study of semantics. In D. D. Steinberg & L. A. Jakobovits (Eds.), *Semantics.* Cambridge: Cambridge University Press.

Mitchell, T. M. (1978). *Version spaces: An approach to concept learning.* Doctoral dissertation, Department of Electrical Engineering, Stanford University, Palo Alto, CA.

Mitchell, T. M., Keller, R. M., & Kedar-Cabelli, S. T. (1986). Explanation-based generalization: A unifying view. *Machine Learning*, *1*, 47–80.

Murphy, G. L. (1982). Cue validity and levels of categorization. *Psychological Bulletin*, *91*, 174–177.

Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review*, *92*, 289–316.

Orloci, L. (1969). Information analysis of structure in biological collections. *Nature*, *223*, 483–484.

Posner, M. I., & Keele, S. W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology*, *77*, 353–363.

Quinlan, J. R. (1986). Induction of decision trees, *Machine Learning*, *1*, 81–106.

Rosch, E., Mervis, C., Gray, W., Johnson, D., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, *7*, 382–439.

Rosch, E., & Lloyd, B. B. (Eds.). (1978). *Cognition and categorization.* Hillsdale, NJ: Lawrence Erlbaum.

Rumelhart, D. E., & McClelland, J. L. (1986). *Parallel distributed processing: Explorations in the micro-structure of cognition.* Cambridge, MA: MIT Press.

Schank, R. C., Collins, G. C., & Hunter, L. E. (1986). Transcending inductive category formation in learning. *Behavioral and Brain Sciences*, *9*, 639–686.

Smith, E. E., & Medin, D. L. (1981). *Categories and concepts.* Cambridge, MA: Harvard University Press.

Sneath, P. H., & Sokal, R. R. (1973). *Numerical taxonomy: The principles and practice of numerical classification.* San Francisco, CA: Freeman.

Wallace, C. S., & Boulton D. M. (1968). An information measure for classification. *Computer Journal*, *11*, 185–194.

Winston, P. H. (1975). Learning structural descriptions from examples. In P. H. Winston (Ed.), *The psychology of computer vision.* New York: McGraw-Hill.

Wittgenstein, L. (1953). *Philosophical investigations.* Oxford: Basil Blackwell.

Zadeh, L. A. (1965). Fuzzy sets. *Information and Control*, *8*, 338–353.