Natural Categories for More Natural Generation

Ben E. Cline and J. Terry Nutter

TR 90-22

NATURAL CATEGORIES FOR MORE NATURAL GENERATION

Ben E. Cline and J. Terry Nutter
Department of Computer Science
Blacksburg, VA 24061
nutter@vtopus.cs.vt.edu

Abstract

Psychological research has shown that natural taxonomies contain a distinguished or basic level. Adult speakers use the names of these categories most frequently and can list a large number of attributes for them. They typically cannot list many attributes for superordinate categories and list few additional attributes for subordinate categories. Because natural taxonomies are important to human language, their use in natural language processing systems appears well founded. In the past, however, most AI systems have been implemented around uniform taxonomies in which there is no distinguished level. It has recently been demonstrated that natural taxonomies enhance natural language processing systems by allowing selection of appropriate category names and by providing the means to handle implicit focus. In previous research, we have argued that benefits from the use of natural categories can be realized in multi-sentential connected text generation systems. We briefly summarize the psychological research on natural taxonomies that relates to natural language processing systems, the use of natural categorizations in current natural language processing systems, and the results of our previous research in which we show how natural categories can be used in multiple sentence generation systems to allow the selection of appropriate category names, to provide a mechanism to help determine salience, and to provide for the shallow modeling of audience expertise. We then describe additional benefits of natural categories in generation systems by demonstrating that natural categories provide a mechanism that aids selection of discourse schema and increase the efficiency of inheritance.

1. INTRODUCTION

People represent information about kinds in taxonomies which are not uniform [Rosch *et al.* 1976; Mervis & Rosch 1981]. In these natural taxonomies, one level of abstraction, called the *basic level*, is the most important and carries the most information. Adult speakers use basic level category names most frequently, and they are able to list large numbers of attributes for categories at this level. Since natural taxonomies form a fundamental basis underlying human language, it is important that natural language understanding and generation systems model them.

The use of natural categories in natural language understanding systems and in single sentence question and answer systems has been demonstrated [Peters and Shapiro 1987; Peters, Shapiro and Rapaport 1988]. Benefits include the ability to use appropriate category names and to handle implicit focus. In [Cline and Nutter 1989], we argued that the use of natural categories is also important in natural language generation systems that produce multi-sentence texts. In addition to allowing selec-

tion of appropriate category names, use of a natural taxonomy provides a mechanism to help determine salience and provides for shallow but potentially useful modeling of audience expertise. In this paper, we describe additional benefits: natural categories provide a mechanism to aid schema selection and provide a means to construct a taxonomy with efficient inheritance.

The structure of this report is as follows. Section 2 presents a brief overview of categorization theory results that relate to natural language generation. Section 3 reviews natural language understanding systems that use natural categories. Finally in Section 4, additional enhancements to natural language generation systems that can be derived from the use of natural categories are outlined.

2. THEORY OF NATURAL CATEGORIES

A category is a collection of nonidentical objects or events that an organism treats as equivalent for some given context. Organisms divide their environment into categories in order to deal efficiently with the vast amount of information presented to them. Taxonomies are collections of categories organized by class inclusion. In a uniform taxonomy, no level is distinguished and attributes are placed at the level of maximal coverage. Although most AI systems model categorizations using a uniform taxonomy, psychologists have argued that one level of natural taxonomies is distinguished [Rosch et al. 1976]. Categories at this basic level are the most cognitively efficient, carry the most information, and are those categories most differentiated from one another. Members of a basic level category have the most non-inherited attributes in common. In other words, a large number of attributes are introduced at the basic level that do not occur at a higher level in the hierarchy. Although subordinate categories inherit attributes from their basic level category, only a small number of additional distinguishing attributes are associated with the subordinate categories.

For example, a typical biological taxonomy has basic level categories for both cats and dogs. Superordinate categories for these basic level categories include *mammal* and *animal*. The basic level categories have subordinate categories for particular breeds. Since members of basic level categories have the most attributes in common, a manx and a Maine ring-tail coon cat will have more attributes in common than either one has with a collie. Two subordinate categories of a basic level category will share many features. In addition, they have some additional features that distinguish them. For example, the *manx* subordinate category has the attribute *has short fur*, while *maine coon* has the attribute *has long fur*. But both subordinate categories share all the common features associated with felines.

The most important results of category theory, relative to the topic at hand, relate basic level categories to human language. Research has shown that subjects list the greatest number of attributes for categories at the basic level. Few attributes are listed for superordinate categories, and few additional attributes are listed for subordinate categories [Rosch et al. 1976]. Regularities in classification across languages have been uncovered [Tversky and Hemenway 1984]. Although category cuts were originally thought to be arbitrary, these regularities appear to be linked to structure in the perceived world. Experiments by Rosch *et al.* [1976] have demonstrated that the names associated with the basic level categories are those most used by adults and first used by children. The basic level is the one at which adults spontaneously name objects.

Classically, it was thought that category membership was established by necessary and sufficient criteria. More recent research has focused on graded category membership [Mervis and Rosch 1981; Smith and Medlin 1981]. Some exemplars of a category are highly representative while others are less so. For example, most birds have feathers and fly. However, penguins are members of the basic level category *bird*, but they are atypical in their flying ability. One line of research claims that the most representative exemplars may be used as prototypes for determining class membership[Smith and Medlin 1981].

Finally, categorization research has pointed out that although principles by which we decide which categories are at the basic level are expected to be universal, for a given domain, the basic level category itself may not be universal [Mervis and Rosch 1981; Rosch et al. 1976]. Both expertise and cultural significance of the domain affect the selection. The level of expertise also affects the amount of information associated with the basic and subordinate levels.

3. Applications of Categories in Natural Language Systems

Peters and Shapiro [1987] have implemented a semantic network system for natural language understanding that models natural category systems. In this system, there is not a great deal of inheritance in the hierarchy. Instead, most inheritance occurs between basic level categories and members of these categories. One of the most important results of their system is that it is able to choose the most appropriate category name for an object in answers to questions.

Peters, Shapiro, and Rapaport [1988] describe an extended version of this system in which context affects the attributes associated with basic level categories. For example, in the context of *farm*, cows, horses, and pigs are more typical of the category *animal* than lions and elephants. The reverse is true in the context of *zoo*. The system uses the context-independent and context-dependent information associated with basic level categories to guide focus while processing English text input. This technique enhances text understanding and anaphora resolution.

This system uses default generalizations to represent typical attributes of members of a basic level category. These generalizations are based on category part-whole structure and image schematic structure, other perceptual structure, and functional attributes. This information is useful in determining category membership and is the knowledge that forms the context-independent structure of the basic level categories.

In [Cline and Nutter 1989], we argued that the use of natural taxonomies contributes to connected text generation in three ways. First, as in natural language understanding systems, the use of natural categories allows objects to be described in terms of their basic level category names. The basic level category provides the most appropriate name because adults spontaneously name objects using this name and can list a large number of attributes for a basic level category.

Second, the use of basic level categories provides a mechanism to help determine salience. Based on the idea of graded category membership [Smith and Medlin 1981], we attach to each basic level category a set of typical features. We identify the potentially salient features of a member of a basic level category as those features of

the individual that differ from the features labeled as typical at the basic level. This provides a first step toward the kind of dynamic determination of salience proposed in [Nutter 1983] and [Nutter 1985]. Although salience rules based on typicality of basic level categories were useful, they did not not identify all the salient features in our taxonomy. For example, rules to identify central features of an item being described or salient features in a particular context were needed.

Third, natural categories provide a method to provide a shallow model of audience expertise for a generation system. Human experts have a different set of basic level categories in their area of expertise than non-experts, and the categories representing this expertise contain more knowledge than for a non-expert. By modeling this type of categorization scheme, a generation system can tailor its output for an audience at a certain level of expertise. This type of modeling allows a generation system to use appropriate terminology for a particular audience; however, additional mechanisms are required to alter discourse structure used in descriptions for different audiences. As noted in [Paris 1988], novices require functional explanations of concepts that experts already understand.

These three benefits of natural categories for generation systems were demonstrated in a natural language generation system that we implemented using the SNePS-2.1 semantic network system [Shapiro and Rapaport 1987; Shapiro and SNIG 1989]. The system contains a knowledge base that represents a taxonomy of microcomputers, rules describing typical features of the basic level categories, and rules to determine salient features of the knowledge base. An ATN is used to produce natural language text of descriptions of items in the knowledge base.

4. Additional Benefits of Natural Categories in Connected Text Generation

We have argued that natural language generation systems benefit from the use of natural categories. In addition to the benefits outlined in the previous section, natural categories can be used to help select discourse schema and provide for increased efficiency of inheritance compared to uniform taxonomies. We describe these benefits and their use in a natural language generation system that produces descriptions of microcomputers.

4.1 Aiding Schema Selection

A natural language generation system can exploit knowledge about natural categories in selecting discourse schema. One way in which a generation system can select and organize the concepts to be converted to surface level text is by the use of schemata which represent standard patterns of discourse which a speaker or writer can use to accomplish some discourse purpose [McKeown 1985]. A schema guides decisions concerning what is to be said and in which order. McKeown's TEXT system uses four schemata: *identification*, *attributive*, *constituency*, and *compare and contrast*. *Identification* is used to identify entities or events. *Attributive* is used to illustrate a particular point about a concept or object. *Constituency* is used to describe an object in terms of its parts, while *compare and contrast* is used to describe an object by contrasting it to another object. In TEXT, a schema is selected based on the discourse goal (i.e., the question asked) and the availability of information required by the schema. Associated with each question type to TEXT accepts is a subset of the schemata from which a discourse strategy is selected. The association of relevant

schemata and question type is built into the system. Then based on the information available in the knowledge base, one of the schemata in the limited subset is selected to provide the discourse structure of the answer. For example, it the question type requires a definition and the knowledge base has little information on the object to be described but more information on its subclasses, the *constituency* schema is selected. However if there is more information on the object than its subclasses, the *identification* schema is selected. In a more general system, tying schema selection to predefined discourse goals is not adequate. Broader techniques that take into account full system knowledge are needed.

Subordinate categories in a natural taxonomy present opportunities to use the *compare and contrast* schema. When describing a subordinate category, there is a potential for comparing and contrasting the subordinate category to another subordinate category of the same basic level category. Many attributes are shared due to the relationship of the subordinate categories to the basic level category. More importantly, the subordinate categories contain few additional attributes. This small number of additional attributes, which likely indicate differences in the subordinate categories, can be used by a generation system to describe concisely the differences in the subordinate categories.

Comparing and contrasting basic level categories (e.g. cats and dogs) would be more difficult since there are many attributes at the basic level, some of which are similar and some of which are not. Although the use of a *compare and contrast* schema is possible at this level, the system would have to depend on additional knowledge to determine important differences in two objects. *Compare and contrast* could also be used for an individual member of a natural category that also belongs to a subordinate category by comparing and contrasting it to an individual of another subordinate category. since the individuals inherit attributes from their subordinate categories, the effect would be similar to *compare and contrast* at the subordinate level but less likely to be consistent with the goal of describing an individual item.

A natural taxonomy also provides the opportunity to use the *constituency* schema. This is particularly true in the case of taxonomies pertaining to technological artifacts. A detailed taxonomy would contain a number of attributes describing the major parts of each basic level category, and a description of a basic level category would appropriately use the *constituency* schema to describe the category in terms of its parts. In a description of a subordinate category or a member of the basic category, constituency is a less appropriate discourse goal. For example, when asked to describe the basic level category *automobile*, the system could appropriately indicate the fundamental parts of an automobile: engine, transmission, drive train, body, etc. A description of the subordinate category *Corvette* or the individual item *John's VW Rabbit* is less likely to contain a complete breakdown of the fundamental parts of a car. At these levels, distinguishing attributes associated with the subordinate category (e.g. engine compression) or individual (e.g. the color of John's car) are more appropriate details.

4.2 Enhancing Efficiency of Inheritance

The use of natural categories in a connected text generation system increases the efficiency of inheritance over a uniform taxonomy. By grouping the majority of attributes at the basic level and by relating each object to its basic level category, the attributes of an object can be located quickly without having to search through the entire hierarchy. Values for attributes that are typical are found at the basic level. Before using the typical value, the system must check if the object has an atypical

value by determining if there is a value for the particular attribute associated with the object or one of the subordinate categories to which it belongs. For a complex taxonomy, this check for exceptional attribute values is computationally less expensive than searching the entire taxonomy for attribute values. Although attributes are not positioned to cover the maximal number of categories that contain the attribute, the additional storage requirements are not great in a semantic network such as Peters and Shapiro [1987] used where storage for each particular attribute is unique and where attributes are related to categories by network connections.

Consider a simple natural taxonomy with five levels. The top-level superordinate category is *computer*. At the next level are two superordinate categories, *digital computer* and *analog computer*. The basic level categories under *digital computer* are *microcomputer*, *mainframe*, and *supercomputer*. Each basic level category has a number of subordinate categories identifying individual models, e.g. *IBM-PC* is a subordinate category of the basic level category *microcomputer*. Under the subordinate categories are individual computers, e.g. *John's IBM-PC*. In the natural taxonomy, the relevant attributes for John's PC would appear at the individual level, the subordinate category, or the basic level category. Since typical readers would understand the term *microcomputer*, the additional information at superordinate categories would not typically be needed in a generation system.

Compare this taxonomy to a simple taxonomy with the same five levels but with no distinguished level. In a uniform taxonomy, each attribute would be placed at the highest possible level of the taxonomy. A generation system would have to search at all five levels of the taxonomy to find relevant attributes. For example, in the natural taxonomy each basic level category under *digital computer* would contain the attribute *has an arithmetic-logic unit (ALU)*. While in the traditional taxonomy, this attribute would be associated with the *digital computer* category. In a much more complex taxonomy, this search of a uniform taxonomy would be spread over even more levels.

By localizing relevant attributes at the basic level categories and below, the inference mechanism that performs attribute inheritance can limit its searches to these levels at the expense of the duplication of some attributes across a number of basic level categories. In a uniform taxonomy, speed of inheritance is sacrificed for improved memory usage. We feel the former technique more correctly models human taxonomies.

4.3 Demonstration

We have developed a small natural language generation system in order to demonstrate the use of natural categories in a generation system. In this system, *microcomputer* is a basic level category. There are two subordinate categories, *IBM PC* and *Morrow*, with each subordinate containing an individual microcomputer. The individual Morrow belongs to John, while the IBM PC belongs to Mary. IBM PC's have keyboards and direct video interfaces and do not connect to terminals, while Morrows do not have keyboards or direct video interfaces but do connect to terminals. IBM PC's have an 8088 microprocessor, while Morrows have a Z-80. Both individual units have specific serial numbers, have 8-bit data buses, and optional hard disk drives. There are a number of rules in the system that allow individual items to inherit properties from the categories to which they belong.

The schemata used in our system are

Description Schema

Natural Category	
Subordinate Category	*Compare/Contrast*
Salient Features	Optional Features

Compare/Contrast Schema

Common Salient Features
Different Salient Features

The slots are filled based on information in the knowledge base, and then a surface representation is produced by an ATN. Vertical columns indicate alternatives, and a schema name in a schema means that the named schema is embedded at that point. for example, the *description* schema embeds the *compare/contrast* schema. In the system, the right side of the *description* schema is used for describing individuals while the left side is used to describe subordinate categories. Note that the slot filling mechanism for *compare/contrast* examines pairs of features, one from each of the two categories being compared, where either or both are identified as salient features. When a particular subordinate category is being described, a second subordinate category is selected nondeterministically for use in the *compare/contrast* schema.

Figure 1 is the system output when asked to describe John's Morrow and Mary's IBM PC. This descriptions uses the left side of the *description* schema. Compare this to figure 2, which is the output of the system when asked to describe the subordinate category *IBM PC*. This description is based on the right side of the *description* schema which embeds *compare/contrast*. The subordinate category level of a natural category provides an ideal opportunity for use of the *compare/contrast* schema.

(John owns a microcomputer. It is a Morrow which has a hard disk drive and a Z-80 microprocessor and does not have a direct video interface or a keyboard. It connects to a terminal.)

(Mary owns a microcomputer. It is an IBM PC which has an 8088 microprocessor.)

Figure 1: Description of individual.

(An IBM PC is a type of microcomputer. Although both the IBM PC and the Morrow have 8-bit data buses, the IBM PC has an 8088 microprocessor while the Morrow has a Z-80 microprocessor. Unlike the Morrow, the IBM PC has a keyboard and a direct video interface and does not connect to a terminal. Optional features of the IBM PC are a color graphics adapter, a color monitor, and a hard disk drive.)

Figure 2: Description of subordinate category.

The system also illustrates the improved efficiency of inheritance that can be obtained by using a natural taxonomy. In our system, many attributes are attached to the basic level category. At the subordinate level, only attributes that are specific to the subordinate category are added. Each particular item has some additional attributes added to it. Members of basic level categories inherit attributes from their basic level category and any subordinate level categories to which they belong. The inheritance of attributes is performed using SNePS inference. Since in an natural taxonomy, attributes of superordinate categories are duplicated at the basic level, the inference mechanism need not consider category levels above the basic level when determining the attributes of an individual. This arrangement greatly reduces the amount of inference required to determine inherited attributes.

The duplication of attributes at several category levels increases the amount of storage required to represent the taxonomy. However, this is not a major concern in a system like SNePS-2.1, where attribute nodes are unique. To attach an attribute to another level requires only two network links and one proposition node. We feel that the increase in performance outweighs the additional storage requirements.

5. CONCLUSION

Current research has demonstrated the usefulness of natural category taxonomies in natural language understanding and generation systems. We have demonstrated that two additional benefits are obtained from using natural taxonomies in connected text generation systems. Using natural categories can help identify parts of the taxonomy where the *compare and contrast* schema will be most effective. It also allows taxonomies to be constructed where inheritance is more efficient than in uniform taxonomies.

Our future research will focus on enhancing our natural language generation system by expanding the taxonomy and text production mechanism. Part of this effort will include expanding our current schemata and adding new ones.

REFERENCES

Cline and Nutter 1989
Cline, B.E. and J.T. Nutter. Implications of natural categories for natural language generation, *Proceedings of the First Annual SNePS Workshop*, Buffalo, November 1989, 125-132.

McKeown 1985
McKeown, K.R. *Text Generation*. Cambridge University Press (Cambridge) 1985.

Mervis and Rosch 1981
Mervis, C.B. and E. Rosch. Categorization of natural objects. *In* Rozenzweig, M. R. and Porter, L. W. (eds.), *Annual Review of Psychology* 32, 1981, 89-115.

Nutter 1983
Nutter, J.T. What else is wrong with non-monotonic logics? Representational and informational shortcomings. *Proceedings of the Fifth Annual Conference of the*

Cognitive Science Society, 1983.

Nutter 1985
Nutter, J.T. Deciding what to say: the need for dynamic selection criteria in connected text generation. Tulane University Computer Science Technical Report 85-101, 1985.

Paris 1988
Paris, C.L. Description strategies for naive and expert users. *Proceedings of the 23rd Annual Meeting of the Association for Computational Linguistics*, Chicago, July 1985, 238-245.

Peters and Shapiro 1987
Peters, S.L. and S.C. Shapiro. A representation for natural category systems. *Proceedings of the Tenth International Joint Conference on Artificial Intelligence*, 1987, 140-146.

Peters, Shapiro and Rapaport 1988
Peters, S. L., S. C. Shapiro and W.J. Rapaport. Flexible natural language processing and Roschian category theory. *Proceedings of the Tenth Annual Conference of the Cognitive Science Society*, 1988, 125-131.

Rosch et al. 1976
Rosch, E., C.B. Mervis, W.D. Gray, D.M. Johnson and P. Boyes-Braem. Basic objects in natural categories, *Cognitive Psychology* **8**, 1976, 382-439.

Shapiro and Rapaport 1987
Shapiro, S.C. and W.J. Rapaport. SNePS Considered as a Fully Intentional Propositional Semantic Network. In N. Cercone and G. McCalla, eds., *The Knowledge Frontier*, Springer-Verlag (New York) 1987, 263-315.

Shapiro and SNIG 1989
Shapiro, S.C. and the SNePS Implementation Group. SNePS-2.1 User's Manual. Department of Computer Science, State University of New York at Buffalo, 1989.

Smith and Medlin 1981
Smith, E.E. and D.L. Medin. *Categories and Concepts*. Harvard University Press (Cambridge) 1981.

Tversky and Hemenway 1984
Tversky, B. and K. Hemenway. Objects, parts, and categories. *Journal of Experimental Psychology: General* **113**, 1984, 169-191.