IMPLICATIONS OF NATURAL CATEGORIES FOR NATURAL LANGUAGE GENERATION

BEN E. CLINE AND J. TERRY NUTTER

TR 89-19

Implications of Natural Categories for Natural Language 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. We propose that additional benefits from the use of natural categories can be realized in multi-sentential connected text generation systems. After discussing the psychological research on natural taxonomies that relates to natural language processing systems, the use of natural categorizations in current natural language processing systems is presented. We then describe how natural categories can be used in multiple sentence generation systems to allow the selection of appropriate category names, to provide the mechanism to help determine salience, to aid in the selection of discourse schema, to provide for the shallow modeling of audience expertise, and to increase the efficiency of taxonomy 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 & Shapiro 1987; Peters, Shapiro, & Rapaport 1988]. Benefits include the ability to use appropriate category names and to handle implicit focus. We argue in this paper that the use of natural categories is also important in natural language generation systems that produce multi-sentence texts. In addition to allowing selection of appropriate category names, use of a natural taxonomy provides a mechanism to help determine salience, aids in selection of discourse schema, provides for shallow but potentially useful modeling of audience expertise, and increases the efficiency of 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, the 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 attributes in common. Tversky and Hemenway [1984] argue that basic level objects are distinguished mostly by part attributes, while members of subordinate classes tend to share parts and differ on other attributes.

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.

Researchers have performed a variety of experiments to verify the existence of basic level categories. It was found 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]. It was also found that basic level categories were the most general level at which averaged shapes (produced by overlaying normalized shapes of category members) could be recognized, thus demonstrating that basic level categories are the highest categories for which a concrete mental image of all category members can be formed. For example, subjects could recognize averaged shapes for basic level categories such as dog and chair but not for superordinate categories such as mammal and furniture [Rosch *et al* . 1976]. Tests were also performed to verify that basic level categories are the most inclusive for which highly similar sequences of motor movements are made to objects in the category [Rosch *et al.* 1976].

However, the most important results for this discussion relate basic level categories to language. Without some categorization system, we would need a separate word for each unique item in the world including each blade of grass and each insect. Natural categories provide a way out of this dilemma; as a result they have had a fundamental influence on human language. Regularities in classification across languages have been uncovered [Tversky & 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 & Rosch 1981, Smith & Medin 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 & Medin 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 & 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. It is believed that an expert's knowledge is often confined to specific parts of the taxonomy, thereby creating unevenness in the taxonomy. There also appears to be a level below which basic level categories cannot be formed regardless of the frequency of use or level of expertise due to the lack of attributes to differentiate objects.

3. Applications of Categories in Natural Language Understanding

Peters and Shapiro [1987] have implemented a semantic network system for natural language understanding that models natural category systems. In their representation, a member/class case frame is used to describe the inclusion of an

object in a basic level category. In addition, ISA case frames are used to designate objects as members of subordinate and superordinate categories. The category hierarchy is built from subclass/superclass case frames. 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 using this representation is that this system is able to choose the most appropriate category name for an object in answers to questions. For example, knowledge in the system indicates that *cat* is a basic level category. The system was told that Jane petted a manx, a manx is a cat, a cat is a mammal, and mammals are animals. When asked who petted an animal, it answered that Jane petted a cat. This response is deemed more appropriate than the responses "Jane petted a manx" or "Jane petted an animal." Violations of this rule can produce unintended humor: compare "Jane petted the cat" with "Jane petted the carnivore."

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.

The context-dependent structure associated with concepts is formed by thematic associates (concepts related to events) and by other concepts not related to categorization. Such information is only relevant in particular situations. For example, *mortgage* is a context-dependent concept associated with *house*. *Mortgage* is a useful concept when attempting to understand text concerning the purchase of a house. In understanding the sentences

Jane bought a house
The mortgage was high

the system adds the concept *mortgage* to a potential focus list when it parses the first sentence because the concept *mortgage* is a thematic associate of *house* in this situation. When "mortgage" is read in the second sentence, the system is able to relate this mortgage to the particular house that Jane bought by using the context-dependent knowledge triggered by the first sentence.

4. Natural Categories and Connected Text Generation

The AI system discussed above demonstrates that the use of natural categories enhances natural language understanding systems and single sentence question and answer systems. We propose that the use of natural taxonomies is also beneficial to natural language generation systems that produce multi-sentential output. Whether a generation system is producing descriptions of objects from a knowledge base or

giving extended answers to questions about such objects, knowledge of basic level categories allows the system to produce more natural sounding and more widely understandable text due to the importance of natural taxonomies in psychology and linguistics.

4.1 Selecting Description Level

Since basic level category names are those most frequently used by adults and are widely understood, when describing an object or a subordinate class, a natural language generation system should define the item in terms of its basic level category. Names at the basic level provide the reader with the most information, as typical adults can list a large number of attributes for basic level categories. For example, in describing the subordinate category manx, a generation system should indicate that a manx is a breed of cat. References to superordinate categories (e.g. "The manx is a mammal" or "The manx is an animal") give the reader far less information.

A natural language generation system should also take into account the degree of representativeness of a member or subordinate category of a basic level category when generating qualifying terms [Mervis & Rosch 1981]. Qualifying terms such as "true" or "technically" are typically applicable only to subsets of category exemplars. "True" is applicable to category members that are strongly typical, while "technically" is reserved for atypical members. "A collie is a true dog" is acceptable while "A collie is technically a dog" is not. "A bottle-nosed dolphin is a true whale" is odd at best while "A bottle-nosed dolphin is technically a whale" is a good description. One way of distinguishing atypical individuals or classes is by noting the absence of features typical of the basic kinds to which they belong. That is, bottle-nosed dolphins are atypical whales because they are roughly human-sized, while typical whales are much bigger. Wolves are technically dogs; Persians are true cats.

When describing a class or individual relative to another fixed superclass, selecting the correct modifier depends only on the detection of typicality or atypicality relative to the second class. Hump-backed whales are true whales, but only technically mammals. The usefulness of the distinguished basic level comes in when the system must describe a class or individual without having a fixed superclass supplied.

4.2 Identifying Salient Characteristics

Natural categories which contain a description of typical features can be used for determining salience. In existing generation systems, salience is typically determined by some static measure. For example, the TEXT system [McKeown 1985] is a question and answer system that is used to describe the structure and content of a database that contains descriptions of military hardware, e.g., ships and missiles. In TEXT "distinguishing descriptive attributes" are attribute-value pairs that are used to partition classes of entities into meaningful subclasses. The system also keeps track of attributes that are constant across subclasses. Distinguishing descriptive attributes and constant database attributes indicate important features in describing classes or in comparing and contrasting one class to another. This technique can be viewed as the application of a limited type of salience processing.

The use of distinguishing features to determine salience fails in general for two reasons. First, distinguishing features may not always be salient. Consider for example a comparison of two different computer systems. The serial numbers of the two systems distinguish them, but serial numbers are rarely salient (except in cases

of inventory or theft). Serial numbers would never be included in a discussion of the relative merits of two particular machines. Second, distinguishing features of a category provide little additional information when describing a typical member of the category. The category *bird* contains the distinguishing attribute *can fly*. However, when describing a robin, to indicate that it can fly tells the average reader nothing he doesn't already know. On the other hand, it is an important feature of ostriches that they are birds that do not fly [Nutter 1983, 1985].

As Peters, Shapiro, and Rapaport [1988] used default generalizations to describe basic level category attributes, we believe that these types of default rules are useful for determining salience in a connected text generation system. Because the attributes of basic level categories are familiar to most general readers, it is useless to mention them when describing an object that is strongly typical of the category. Sometimes mentioning the obvious can even imply unintended interpretations because the reader does not expect a speaker or writer to state the obvious unless it is important. On the other hand, when a subordinate category or a member of a basic level category has an attribute that differs from a basic level attribute, this difference is probably salient and is a candidate to be included in the system output. For example, if the particular cat that is being described has only three legs while a default rule of the basic level *cat* category indicates that cats typically have four legs, this difference has potential salience. If the cat has four legs, on the other hand, the system need hardly say so.

4.3 Aiding Schema Selection

A natural language generation system can also 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 85]. 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. 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.

4.4 Shallow Expertise Model

Natural categories provide a way for a natural language generation system to model audience expertise [Peters & Shapiro 1987]. In the domain of their expertise, experts tend to have a different taxonomy structure than nonexperts. The level of abstraction at which the basic level categories of an expert occur differs from that for nonexperts as does the amount of information at both the basic level categories and the subordinate categories. For example, Rosch [Rosch *et al.* 1976] discovered that one of their subjects was a former airplane mechanic. While *airplane* was a basic level category for other students, the former mechanic had basic level categories based on types of airplanes. Furthermore, this student was able to list many more attributes for the categories related to *airplane* than were other students.

Consider an author producing a paper for a meeting of cognitive scientists. The audience will contain a number of experts with varying degrees of expertise in various areas of the field. For the paper to be effective, the author must attempt to write it at a level of expertise that is common to the members of the audience, presenting his arguments in terminology common to the experts and using concepts that they share. If the author writes at a higher level of expertise than is common to the audience, the majority of them will not understand his paper. On the other hand, a paper at a lower level which explains commonly understood concepts will bore them.

In a similar manner, a generation system should tailor its output for a particular audience by modeling audience expertise. Where the audience is a large group, many of whose members are unknown, deep modeling of the audience individual-by-individual is clearly impossible. Furthermore, even if it were possible, it would not be appropriate, since it would be too computationally intensive to be usable. So a shallow mechanism is needed for determining the appropriate level of discourse for a particular audience. An implementation based on natural taxonomies can provide such a shallow model. In order to adjust a generation system for a more expert audience, the basic level categories would be moved to a lower level of abstraction and additional knowledge would be added to the basic and subordinate categories. Since the type of generation system we propose relates objects to their basic level category names, basic categories at a lower level of abstraction would cause the system to use more expert terminology. For example, a system modeling a feline expert would have breed categories at the basic level and would tend to use breed names when describing individual cats instead of the term *cat*. The addition of knowledge to the basic and subordinate categories would allow a generation system to produce text more suitable for experts. Although this technique allows a generation system to produce terminology at the appropriate level of expertise, techniques to select appropriate discourse styles based on the degree of expertise are also needed.

4.5 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.

5. Conclusion

Current research has demonstrated the usefulness of natural category taxonomies in natural language understanding systems and in single sentence question and answer systems. We have argued that connected text generation systems can also benefit from the results of categorization research. Natural categories allow generation systems to produce more understandable text by describing objects and subordinate categories in terms of their basic level category names which are widely understood and have many attributes associated with them. Basic level categories also provide a way for a generation system to determine salient features of a knowledge base by providing typical attributes of basic level classes so that the atypical attributes of a member can be determined. The attributes of an object that differ from these defaults indicate potentially salient information. Natural categories can aid the system in the selection of discourse schema by indicating areas of the knowledge base where differences between objects may be located. Modeling expertise by shifting basic level categories and attributes of subordinate categories and by adding additional knowledge to the categories can be used to have a natural language generation system produce text for different audiences. Finally, the use of a natural taxonomy provides increased efficiency of inheritance over uniform taxonomies by grouping attributes at the basic level and associating each object with its basic level category.

References

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

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

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

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.

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

Peters, S. L., Shapiro, S. C. and Rapaport, W. J. Flexible natural language processing

and Roschian category theory, *Proceedings of the Tenth Annual Conference of the Cognitive Science Society*, 1988, 125-131.

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

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

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