Category Development Based on Semantic Principles

Kenneth C. Litkowski
CL Research
20239 Lea Pond Place
Gaithersburg, MD 20879
Telephone: 301-926-5904
Email: ken@clres.com
Web site: http://www.clres.com

Abstract

Category systems extracted from textual material are an important part of the scientific process and qualitative analysis, but are frequently developed on an ad-hoc basis. Their development can be improved by a stronger reliance on linguistic and semantic principles. Tracing and recapitulating the development of semantic principles from the 1950s to the 1990s in linguistics and text analysis shows how the principled use of information from lexical resources can facilitate the development and analysis of category systems. These principles are demonstrated in examining the category systems of Minnesota Contextual Content Analysis, a technique for analyzing textual material from sentences, answers to open-ended questions on questionnaires, expository texts, and verbatim transcripts. These principles are then extended to show how to abstract category assignments for this type of textual material. The discussion identifies computerized systems and data resources used in the application of these principles.

Keywords: Category development, semantic analysis, linguistic analysis, text analysis, content analysis, qualitative analysis

Author's Note: This paper incorporates material from a presentation to the Conference on Computers and the Social Sciences, Minneapolis, 1996. Special thanks to Don McTavish for awakening my interest in content analysis and to David Fan for his patience and comments on earlier drafts of this paper.

1. Introduction

Categories are used in every form of science, and indeed, may constitute the roots of science. A category is an abstract class, group, or set consisting of individual elements of any type. A category is defined by characterizing these elements. In this paper, the elements are individual words of natural language, as well as phrases like county seat that have well defined meanings. Beyond such words and simple phrases, set elements can be more complex phrases, sentences, paragraphs, and even entire texts; at the end of the paper, categories for such larger units of text are discussed.

A vast array of methods, particularly statistical, is used to categorize information and data. But many qualitative approaches are used as well, particularly in social science research. For the most part, such techniques are based on the investigator's intuitions about the meaning of categories, perhaps supported with statistical analysis. While such approaches can and should be continued, some other avenues have opened up with the developments in linguistics and the semantic theories supporting linguistics. This paper presents techniques for category

development based on semantic principles (that is, principles for describing the meaning of words), particularly by weaving in the historical emergence of these principles.

To ground this discussion, the paper (1) characterizes some of the ways in which categories are used in social science, from the simple use of categories like gender in questionnaires, through category development in theory development, to highly intricate category systems involving hierarchical systems, and (2) looks briefly at category development for thesauruses and library cataloguing systems. The paper then describes, in the 1950s (the early days of computers), the beginnings of computerized information retrieval and text analysis, particularly from the perspective of their use of thesauruses and cataloguing systems. A brief historical overview then describes formalizations of linguistic principles in the development of formal grammars (that is, how words can be combined in phrases and sentences) and semantics. This overview is then unfolded in the presentation of principles for category development, based on research in linguistic formalisms continuing with ever richer grammars and semantic formalisms. The progression of these formalisms is described in the examination of the categories used in the Minnesota Contextual Content Analysis (MCCA) approach. Finally, current research toward an integration of semantic principles into content analysis describes abstraction procedures for characterizing the "category" of any text.

2. Category Systems Based on Textual Material

Most scientific endeavors involve defining variables in terms suitable for measurement, developing the measures, and specifying the variable relationships in order to check for misspecification of the variables and for measurement errors (U. S. General Accounting Office, 1993: 26). This requires defining variables "in concrete, specific, unambiguous, and contextual terms that reduce the measure to a single trait or characteristic" (U. S. General Accounting Office, 1993: 29). "Measures must be accurate, precise, valid, reliable, relevant, realistic, meaningful, comprehensive, and in some cases complementary, sensitive, and properly anchored" (U. S. General Accounting Office, 1993: 30). This paper is concerned with the development of variables with conceptual underpinnings (that is, categories) and how to ensure that they are well-defined and meet the requirements for being meaningful, comprehensive, and perhaps properly anchored. Such categories may refer to the ostensibly simple concepts of adult, middle-age, and senior in groupings by age.

A survey researcher engaged in exploratory work may ask open-ended questions whose answers can be analyzed only by examining the texts of the responses. The researcher may have initial, sketchy conceptions of the categories into which the answers will fall. In questionnaire development, the researcher formulates questions where answers should identify a comprehensive set of alternatives (such as list of items, multiple choices, ranking scales, and Likert scales, and range, amount, and frequency intensities) (U. S. General Accounting Office, 1993: 46-78). The set of possible answers should contain all the desired categories, should not overlap, and should have an appropriate level of specificity (U. S. General Accounting Office, 1993: 102-9).

Content analysis of open-ended exploratory questions, verbatim transcripts of speech or interviews, or other free textual material is essentially theory development in which an analyst assigns categories to organize the textual material. This development is very difficult, very subjective, and frequently open to criticism of replicability and interrater reliability. Many investigators eventually create categories featuring particular words such as those expressing

emotion expressing words. The analysis then consists of obtaining the frequencies of such words throughout the textual material. Of course, what constitutes an emotion expressing word is an important issue. Many content analysts have developed dictionaries, assigning words to different categories based on their individual judgments; these analysts may articulate criteria used for the development of their systems, sometimes stating that the words in a category share "semantic components," that is, common elements of meaning. However, the validity of these category systems can frequently be criticized. This paper describes the use of semantic principles for the development of criteria, with the goal of placing category development on a firmer basis.

3. Thesaurus, Library Catalogue, and Information Retrieval Categories

Category development predates the computer era primarily in the classification of (1) human works such as as books, films, and plays into library catalogs and (2) words into thesauruses (Roget's International Thesaurus, 1992). While these systems predate computers, they are still relevant to category development. Library cataloguing systems have attempted to organize the world's knowledge into broad groups, broken down into ever-finer categories, so that each catalogued item is placed within the system. However, such a system is ultimately problematic when an item may logically fall under more than one category. In addition to the general cataloguing systems in public libraries, some disciplines, most notably the medical community, have developed cataloguing systems that reflect the complexity of those disciplines.

A thesaurus, while presenting synonyms and antonyms, is generally organized by grouping words according to ideas. Thus, Roget's International Thesaurus (1992) uses 1,073 categories in 15 classes (with further loose groupings within the classes, down finally to pairings of opposites, such as Assent and Dissent). For example, the subclass Sex, with associated categories Masculinity and Femininity may be used to formulate gender based categories.

With the onset of computers, thesauruses and cataloguing systems gained considerable flexibility in permitting multiple terms or categories for characterizing textual materials. The primary purpose of these categories, of course, is for the retrieval of documents. Thesauruses became important adjuncts to cataloguing systems, since documents could be characterized by key words, the stock in trade of thesauruses.

Thesaurus development expanded dramatically with the advent of computer age in the 1950s. This expansion has contined unabated to the present. The process consists of identifying words and phrases used in documents and then placing them within a thesaurus. Unfortunately, with the rapid expansion of these activities, less attention is placed on the overarching schemata for a thesaurus. Instead, the emphasis has been on "local" placement decisions in which a new entry is related to other entries, primarily by linkages through synonyms, broader terms, narrower terms, and perhaps antonyms. The overall consistency of the thesaurus is seldom examined. Notwithstanding, available thesauruses of this type are valuable resources for category development.

Within the field of information retrieval, classification of documents is a primary endeavor. A considerable amount of research uses the existence of words in a text as the basis for "classifying" the text, often in relation to other texts and documents. This type of research focuses on the frequency of occurrence of words and uses sophisticated statistical techniques for the classification. While many of these techniques may be useful for category development, it is important to distinguish between classification and category development. The difference is largely one of scale, with classification generally focusing on whole texts (books, reports, and

papers), while category development focuses on narrower text segments (individual words, phrases, and sentences). But, as category development attempts to cope with the larger text segments of paragraphs, speeches, and interviews, the boundary with classification begins to blur. The principles described in this paper show how category development may cope with these larger segments and perhaps eventually with classification.

4. Text Analysis

The advent of the computer in the 1950s also saw the expansion of efforts to characterize and critically evaluate writings. The computer enabled rich new areas of research to examine characteristics of textual materials. The computer made it possible to examine a wide variety of statistics about texts, identifying such things as frequencies of words, their average length, sentence complexity, and vocabulary growth.

Beyond benefit to information retrieval and automatic text processing efforts, these statistical analyses also enhanced efforts at more sophisticated analyses of the content of texts. The initial work only looked at patterns for very common words such as articles the and a, pronouns, and prepositions. However, these efforts soon turned to the analysis of more 'content'-ful words. The pace of text analysis has accelerated in literary analysis, authorship attribution, the quantification of qualitative data, as well as the analysis of transcripts from focus groups, psychotherapy, and interviews.

5. Evolution of Grammar and Semantics Research

The 1950s also saw the beginnings of analyses of both the frequencies of words in texts and a theory of syntactic structures describing permissible phrases within sentences (Chomsky, 1956; Chomsky, 1965). The late 1950s and the 1960s saw a greater understanding of language syntax and the accompanying development of modestly efficient parsing routines that provided some capability for representing at least the syntactic structure of text. There was a beginning in the identification and assignment of semantic features to words. An example here would be the assignment of the feature male to the word bachelor. Progress was also made in the assignment of semantic roles to various types of syntactic structures. For instance, the subject of a sentence might be identified as the agent of an action. Or, the object of the preposition with might be identified as being the instrument of an action) (Katz & Fodor, 1963; Fillmore, 1968). In information retrieval, the use of thesauruses (with synonyms and rough “broader than” and “narrower than” hierarchies) led to the organization of concepts into hierarchies useful for grouping text segments into conceptual threads through the text. But, there was not yet an integration of semantics (see (Quillian, 1968; Kucera & Francis, 1967)).

The 1970s saw the emergence in artificial intelligence of techniques for creating knowledge bases and representing various types of relations among logically-stated pieces of knowledge (Winograd, 1972; Schank, 1975). This recognition of these linkages led to initial semantic studies of the lexicon, both the words used in a language and the relationships among them (Jackendoff, 1972; Amsler, 1980; Evens, et al., 1980; Litkowski, 1978). The 1980s saw considerable expansion in the study of semantic relations, leading to further understanding of the importance of the lexicon as the bedrock for understanding the nature of syntactic structures. But, the 1980s also showed the need for the compilation of massive amounts of information for characterizing each piece of the lexicon.

The 1990s has seen the continuation of this accumulation of information, so that today,

the lexicon is populated with information that characterizes the meaning of a word, where that word sits in a hierarchy representing the lexicon, the nature of its relations with other items in the lexicon, what syntactic patterns it may participate in (particularly for verbs and verbal nouns), and what might be its collocations (the company a word keeps). This information is used to identify the specific sense in which a word is used, whether through syntactic analysis or through more statistically-based associations.

With identification of the specific concept associated with each word, it is possible to build a much richer representation of a text passage. It is possible to identify the context, to study the ebb and flow of that context, to place the concept within its proper structure within a sentence, and to organize the sentences (that is, the discourse) into its overall structure, and thus, to identify more precisely the overall organization of a text.

Given this overall process for organizing text, the next task is to bring these techniques into real-world processing. The greatest difficulty lies in what is known as the lexical acquisition and knowledge acquisition bottlenecks. It simply takes a lot of time to put all this information into the lexicon and to build the systems to do the processing. The technology is here, but techniques are needed to put it together efficiently for use in information retrieval and text processing systems. The principles that follow will facilitate this process.

6. Principles of Category Development

6.1 Lexical Resources

Lexical resources include dictionaries, thesauruses, grammars, sets of examples of a word's use, specially constructed databases of information about words, and linguistic analyses of words; they provide information about words; they are used to develop lexicons, systematic representations of characteristics of words suitable for use in computerized text analysis systems.[1] The principles described in this section make use of three distinct lexical resources: (1) a machine-readable dictionary (MRD), a searchable reproduction of a paper dictionary, used to identify parts of speech such as nouns, verbs, and adverbs, inflectional forms such as the past tense or gerundial forms of verbs, and derivational forms such as concept that management is derived from manage; (2) an 1800 page description of grammatical and semantic properties of the English language, used to identify features and characteristics of words (Quirk, et al., 1985); and (3) WordNet, a freely available rigorously developed database of approximately 120,000 words and phrases, with these words and phrases grouped into synonym sets (synsets) and organized into a hierarchical and relational semantic network (Miller, et al., 1990). WordNet can be used to identify common semantic components for words, since its principal relation is the hierarchical ISA relation (a "horse" is a "mammal," establishing that "horse" has the semantic component "mammal").[2]

6.2 Minnesota Contextual Content Analysis

Minnesota Contextual Content Analysis (MCCA) is a technique for characterizing the concepts in textual material, ranging from answers to open-ended questions in surveys through sentences, paragraphs, interview transcripts, and books. MCCA places each English language word into one of 116 categories, counts the words in each category and compares the frequency profile against that for general English usage (McTavish & Pirro, 1990; McTavish, et al., 1997a; McTavish, 1997b).

In the MCCA dictionary of 11,000 words, the average number of words in a category is 95, with a range from 1 to about 300.[3] Each category is given a name, but these names are only heuristic in nature and have no essential meaning. The categories appear internally consistent in that the words in each category have an underlying similarity. However, the characteristics of the categories are not intuitively obvious. Firm principles for category construction can help extend the MCCA dictionary and improve the function of this program (McTavish, et al., 1997a; Litkowski, 1997). These principles are a part of the DIMAP dictionary creation and maintenance software (CL Research, 1997 - in preparation).[4] DIMAP includes MCCA as a module and improves the dictionary and the function of the technique by creating sublexicons for individual categories. These sublexicons are based on WordNet synsets, information from the Merriam-Webster Concise Electronic Dictionary, as well as the other resources described above.

6.3 Initial Stage Based on Part of Speech Analysis

The first stage of category analysis involves looking at the part of speech of the words in the categories. This stage corresponds to the earliest developments in computational text processing in the 1950s, when the focus was on the part of speech of words. Eleven categories in MCCA (such as Have, Prepositions, You, I-Me, He, A-An, The) consist of only a few words in closed classes.[5] The category The contains one word and the category Prepositions contains 18 ~~words~~. About 20 categories (Implication, If, Colors, Object, Being) consist of a relatively small number of words (34, 22, 65, 11, 12, respectively) taken primarily from syntactically or semantically closed-class words such as subordinating conjunctions and relativizers or words which are found at the top levels of WordNet and represent abstract concepts like person, place, and colors. To determine that these categories consist primarily of closed class words, the words in the category were passed through DIMAP to extract just this set from the integrated MRD. Inspection of the part of speech field confirmed the intuitions about the category assignment.

When the parts of speech of words in a category belong to open classes, analysis becomes a little more difficult. When the words are all in one class (that is, all nouns, verbs, adjectives, or adverbs), a unifying principle is sought from the hierarchical relationships among the words. One possible principle is that the words fall into a small number of categories in a thesaurus such as that of Roget. Another possibility is that the words are related by "broader than," "narrower than," or synonymic relations as assigned in keyword indexing thesauruses. Yet another possible principle is one used for dictionary definitions and consists of examining definitions of the words to identify an umbrella genus word with more specific terms underneath. Using WordNet, this step involves identifying the hierarchical groupings of the words in the category.

The remaining 80 or so categories in MCCA consist primarily of just such open-class words (nouns, verbs, adjectives, and adverbs), sprinkled with closed-class words (auxiliaries, subordinating conjunctions). Several categories consist of words from a single part of speech as is the case with Functional roles, Detached roles, and Human roles, which all include only nouns. To examine such unified sets of words, it is valid to examine their definitions for common genus terms. DIMAP implements the more convenient method of using WordNet to examine hierarchical relations as in Table 1, which shows a sample dictionary entry where the field "Isa links" shows that "animal" is of type "creature".

- - - - - -

Table 1 about here

- - - - - -

To see how this field is used, consider the MCCA category Detached roles, which has a total of 66 words, including the words:

ACADEMIC, ARTIST, BIOLOGIST, CREATOR, CRITIC, HISTORIAN, INSTRUCTOR, OBSERVER, PHILOSOPHER, PHYSICIST, PROFESSOR, RESEARCHER, REVIEWER, SCIENTIST, SOCIOLOGIST.

These words fall under the WordNet synsets headed by PERSON (although not including this word), in particular, synsets headed by

CREATOR;
EXPERT: AUTHORITY: PROFESSIONAL;
INTELLECTUAL.

Other synsets under EXPERT and AUTHORITY do not fall into this category (and would thus be included in other MCCA categories). Thus, it is possible to characterize Detached roles as words used to describe persons performing intellectual or thinking activities. This is a concept well captured by its heuristic name, and distinguished from Human roles such as uncle or bride and Functional roles such as janitor or firefighter. Identification of these synsets facilitates extension of the MCCA dictionary for this category to include further hyponyms (that is, types of creators, experts, or intellectuals) of these synsets.

6.4 Semantic Features and Semantic Components

The heuristic name given to the category of Detached roles along with the defining WordNet synsets, suggests the next stage in the process of category development, as well as the next step in linguistic consideration of the lexicon. Table 1 also shows the field "Features", indicating properties of the lexical items, such as "Age" and "Sex". Katz & Fodor (1963) proposed the use of semantic features to characterize entries in a lexicon. In the sample category, there is a feature "Human" with a value "+" and a feature "Role" with a value "Detached." Several more features might be proposed to encode words in this category; hundreds, if not thousands, of other features can be used to characterize the full set of words in the English. For example, Whissell (1996) developed a "Dictionary of Affect", encoding dimensions of emotion-activation and pleasantness.

Laffal (1995) likewise based his dictionary of 43,000 words and 168 concepts on semantic features, coding words in the same category based on the "core meanings of words," that is, having the same semantic component. Nida (1975: 174) characterized a semantic domain as consisting of words sharing semantic components. However, he also suggests (Nida, 1975: 193) that domains represent an arbitrary grouping of the underlying semantic features.

Thus, it is possible to see that the 1960s development of the notion of semantic features has become a very prominent basis for the development of category systems. The subtrees rooted at particular nodes in the WordNet hierarchies provide a readily available basis for category development that reflects implicit assignment of common semantic features and components. Litkowski (1997) proposes making these semantic features and components more explicit, specifically for the purpose of facilitating category development.

6.5 Syntax and Semantic Roles

The 1960s saw the rapid development of formalisms for representing the syntactic structures of phrases, clauses, and sentences, but there was relatively little research toward integrating semantics (that is, meanings) into the representations. Fillmore (1968) began a process of characterizing the semantic roles of noun phrases in a sentence, particularly as related to the main verb. Thus, in addition to identifying the subject and object of a verb and the object of a preposition, it was possible to characterize the role of these syntactic items, by referring to them as, for example, agent, patient, theme, instrument, and location. There are about 30 to 50 semantic roles although there is still no full agreement on what the complete set should be. Table 2 shows a lexical entry for the word eat and illustrates the way in which syntactic and semantic role information is encoded. Important to this example is the requirement that the word eat have associated syntactic items of subject and object. The subject identifying an "agent" who performs the act of eating and a "theme" describing the thing being consumed are both encoded as features of the lexical item.

- - - - - -

Table 2 about here

- - - - - -

Syntactic and semantic role information is normally used for parsing text, but it can be important for category development as well. This can be seen in the analysis for the MCCA category, Sanction, which contains 120 words, including the following words:

> APPLAUD, APPLAUSE, APPROVE, CONGRATULATE, CONGRATULATION, CONVICT, CONVICTION, DISAPPROVAL, DISAPPROVE, HONOR, JUDGE, JUDGMENT, JUDGMENTAL, MERIT, MISTREAT, REJECT, REJECTION, RIDICULE, SANCTION, SCORN, SCORNFUL, SHAME, SHAMEFULLY

While this set of words includes words from several parts of speech (discussed in more detail below), it is rooted primarily in the Levin (1993) verb sets of Characterize (class 29.2), Declare (29.4), Admire (31.2), and Judgment (33). This means that the set has particular syntactic and semantic patterning in addition to the synonymic and hierarchical relations that can be discovered using the techniques described in the previous section. Levin has identified a considerable set of syntactic properties associated with the classes she has developed (and thus a useful resource itself for category development), but has not yet formally characterized the semantic properties. Instead, the definition of this class might, following Davis (1996), inherit a sort notion-rel, which has a "perceiver" and a "perceived" argument (thus capturing syntactic patterning) with perhaps a selectional restriction on the "perceiver" that the type of action is an evaluative one (thus providing semantic patterning). In other words, the underlying conceptualization of the MCCA category indicates that there is an action involved (as indicated by the verb), that this action involves some idea or notion on the part of the actor (the "perceiver"), and that this notion (the "perceived") is inherently an evaluation.

WordNet synsets explicitly contain some syntactic information and implicitly some semantic role information. However, it does not have the depth required for the analysis described above. Other resources, such as Levin (1993), as well as some databases being constructed for on-line access, contain more of this detail. What this means for purposes of characterizing and extending the words in the category Sanction is that not only can the WordNet hierarchy be used, but also it is possible to include words that correspond to conversion of verb

concepts into noun counterparts (for example, the action judge corresponds to the result of a judging action, that is, a judgment).

6.6 Selectional Restrictions, Semantic Relations, and Knowledge Bases

The evolution of artificial intelligence and semantics in the 1970s and the 1980s (Amsler, 1980; Evens, et al., 1980; Winograd, 1972; Schank & Abelson, 1977; Markowitz, et al., 1986) has provided significant amounts of understanding about potential information that can be included in lexical entries that can be used in category development. This discussion illustrates three pieces of information (selectional restrictions, semantic relations, and knowledge base information) that may be included in lexical entries and that can assist in the process of category development. These are discussed for the sake of completeness, but are not described in the present analysis of MCCA categories because of space considerations.

As alluded to in the last section, a restriction was placed on the type of notion involved in the use of a word in the Sanction category, namely, that it had to be evaluative in nature. Table 3 shows a lexical entry for the preposition in with two senses. Basically, this entry says that in is used to begin prepositional phrases (the "pp-adjunct") with noun phrase objects. In the first sense, this says that the phrase may be attached to another noun phrase which may be an "object" or an "event" and that the object of the prepositional phrase is a location in some physical object. The second sense says that the prepositional phrase is attached to a verb which describes an event and that the object of the preposition describes a location which may additionally be characterized as a destination. These specifications are called selectional restrictions and serve to limit the range of words that may appear in the identified syntactic positions.

- - - - - -

Table 3 about here

- - - - - -

Table 4 shows a lexical entry describing an event (of which there may be many types). But, additionally, the entry states that any word describing an event is inherently related, in several possible ways, to other lexical entries. These are known as semantic relations. They are encoded here as features with values preceded by plus (+) signs, which are taken to mean that the following word is actually a selectional restriction on what other lexical entries may appear in the particular relation. The relations shown in Table 4 are quite general and would apply to many lexical entries. However, the number of possible relations is unbounded, similar to the open-class words, and hence, a relation may be of arbitrary depth and specificity. For example, a chemical event relation "hydrogenate" could be defined and specify that its location is a test-tube.

- - - - - -

Table 4 about here

- - - - - -

Table 5 presents a lexical entry for the word or concept "teach". "Teaching" is a communicative event that involves a "teacher" as the agent and "knowledge" as the "thing" that is passed on. The lexical entry specifies that a "teaching" event may consist of three subevents, where a teacher performs a "describing" action, where there may be a "request" subevent (when a student asks for information), and where there may be an "answering" process. The corresponding lexical entries for the "answering" and "describing" subevents show that they

inherit information from the "teaching" event. The three lexical entries, considered as a unit, are construed as part of a script (see Schank & Abelson (1977)).

- - - - - -

Table 5 about here

- - - - - -

Lexical entries containing information on selectional restrictions, semantic relations, and knowledge base data can be used in category development primarily by enabling an analysis of how the embodied concepts fit together, that is, which ones are in more subsidiary positions. The lexical entries described in Tables 3, 4, and 5 illustrate the general linguistic finding that the representation of meaning is focused principally on the verbs and that these verbs may themselves be arranged in hierarchies. Analysis or development of categories should therefore consider this information in identifying the characteristics of the words in the category.

6.7 Lexical Rules, Derivations, and Sense Relations

The final type of information in lexical entries considered here is based on the phenomena by which new lexical entries are derived from existing ones. The most basic of these derivational relations is the one in which inflected forms are generated. These are generally quite simple, and include the formation of plural forms of nouns, the formation of tensed (past, past participle, gerund) forms of verbs, and the formation of comparative and superlative forms of adjectives. The discussion above of the MCCA Detached roles and Sanction categories did not mention the possibility of including these inflected forms, but in fact, these forms are included.

Several more elaborate forms of relations are also possible. For the purpose of illustrating these additional derivational rules, consider another MCCA category, known as Normative. This is a complex category consisting of 76 words, and like the Sanction category, also has words from all parts of speech. This category includes the following (along with various inflectional forms):

> ABSOLUTE, ABSOLUTELY, CONSEQUENT, CONSEQUENCE, CONSEQUENTLY, CORRECT, CORRECTLY, DOGMATISM, HABIT, HABITUAL, HABITUALLY, IDEOLOGICALLY, IDEOLOGY, NECESSARILY, NECESSARY, NORM, OBVIOUSLY, PROMINENCE, PROMINENT, PROMINENTLY, REGULAR, REGULARITY, REGULARLY, UNEQUIVOCALLY, UNUSUAL, UNUSUALLY

The use of the heuristic Normative to label this category clearly reflects the presence in these words of a semantic component oriented around characterizing something in terms of expectations or standards. Of particular interest here are the derivational relations that form adjectives from nouns, nouns from adjectives, and adverbs from adjectives. There were similar kinds of relations in the Sanction category, where most of the concepts seemed to be based on underlying verb forms. In that category, a number of words were clearly noun, adjective, and adverb derivations from the underlying verbs.

These derivational relations can be encoded in lexical entries in the same way as the semantic relations shown in Table 4. The feature name in such relations would describe the relation (such as "nominalization") with a value identifying the derived form, which would also be a lexical entry having the inverse relation ("nominalization_of"), with a value showing the base form of the word. Some of these relations are shown in WordNet, but a more complete source is a dictionary which shows an ordering of derived forms. The MRD included with

DIMAP shows these forms.

The adverb derivations in the Normative category have an additional interesting aspect to them. The heuristic Reasoning has also been used to label this category. Examination of the syntactic and semantic nature of these adverbs shows that they are considered to be content disjuncts (Quirk, et al., 1985: 8.127-33), that is, words indicating that the speaker is making a comment on the content of what the speaker is saying, in this case, compared to some norm or standard. Thus, part of the defining characteristics for this category is a specification for lexical items that have a [content-disjunct +] feature. Analyzing text that contains such words as necessarily, obviously, unequivocally, and consequently would thus indicate the presence of editorial commentary. This shows the value of using non-database sources that describe syntactic and semantic characteristics of the language.

The final type of lexical rule considered here is more subtle and involves the observation that a word may have several senses that are related to one another (usually with one sense as the base from which all the others have been derived). A simple example of such a rule is the word "fish." The base sense of this word refers to an individuated object that is countable; the derived sense is where it refers to the food sense, where the object is not individuated but an undifferentiated mass or substance. Another example of the same process is use of the word "coffee." A lexical rule has been developed to encode this regularity in language and is shown as a lexical entry in Table 6. Note that there is a general rule of "grinding" and then a more detailed entry for "animal-grinding." For the more general rule, a count noun is converted into a mass noun, taking it from an individuated object to a substance. In the more specific rule, the count noun is required to be an animal and then the derived form is a food-substance. Table 7 shows how this might be encoded in a dictionary entry for the word "coffee," where sense 2 of the word is derived from sense 1.

- - - - - -

Table 6 about here

- - - - - -

- - - - - -

Table 7 about here

- - - - - -

These kinds of lexical rules (showing the way different senses are related to one another) are presently a topic of much research, so they are not usually found in any easily accessible databases. However, an awareness of their existence is important for category development.

6.8 Summary of Procedures

In the analysis of MCCA categories, the first step was to extract from the full MCCA dictionary the words in a particular category (performed automatically using DIMAP). This list of words was then passed up against the integrated machine-readable dictionary, automatically creating a sublexicon of entries consisting of just the words on the list. These entries were then visually examined to determine part-of-speech, inflectional, and morphological characteristics. If possible, the words were then grouped in a word processing program so that all words based on a single base word appeared on a single line. Next, the base words were passed through DIMAP to extract and create lexical entries from the WordNet database. Information created automatically in these entries included the relations to other words (in WordNet, but also within the created sublexicon in DIMAP). These relations were visually inspected to determine what hierarchical

relations were present among the words in the category; these relations were then used to rearrange the word lines in the word processing program, so words related hierarchically were indented under their more general words. The words in the group were looked up in Quirk, et al. (1985); if discussed in that text, any properties were identified. The combination of all this information then constituted the definition of the category, permitting a critique of the MCCA categorization and its automatic extension using DIMAP runs based on data from WordNet.

7. Abstraction as Part of Category Development

The preceding section has shown the many ways in which lexical information can be used in category development. While this is important (and all category development can usefully be based on such considerations), categorizations can go beyond the word level. As noted above, the issue of separating categorization from classification comes into play. The techniques of content analysis (including that embodied in the MCCA technique) represent one method of attempting to identify and classify texts that go beyond the single word or phrase. Linguistic techniques are presently emerging that may allow a smoother transition from the word level to the text level.

Burstein, et al. (1996) describe techniques for using lexical semantics to classify responses to test questions. An essential component of this classification process is the identification of sublexicons that cut across parts of speech, along with concept grammars that allow the collapsing of phrases and clauses into a generalized representation that abstracts away from the reliance on individual words. As seen above in the procedures for defining MCCA categories, addition of lexical semantic information in the form of derivational and morphological relations (that is, word formation rules) and semantic components common across part of speech boundaries would justify the development of concept grammars.

Litkowski & Harris (1997) discuss extension of a discourse analysis algorithm incorporating lexical cohesion principles. These principles show how the information in lexical entries, particularly selectional specifications on verbs, maintain cohesion of a discourse. With such information, it is possible to understand how the individual components of a text fit together, and in particular, shows that particular phrases and sentences are elaborations of others (and hence not an essential part of its categorization). As a result, it is possible not only to provide a more coherent discourse analysis of a text segment, but also to summarize the text better and thus provide an overall categorization of a text, rather than just a classification.

8. Conclusions

By following the steps in which the understanding of linguistic processes has evolved since the 1950s, a set of principles has emerged for developing and analyzing category systems. Specifically, these principles require analyzing a lexicon to articulate the specific sets of linguistic and semantic characteristics that define the categories. Many existing and emerging sources of lexical information, including thesauruses, dictionaries, lexical databases, and descriptions of grammatical principles, can be used in category development.[6] Use of these lexical resources and adherence to the category development principles can improve the reliability and validity of category systems used in development of response sets for questionnaire items, analysis of open-ended questions, and analysis of textual material from the sentence to the book level.

Endnotes

1. A lexicon includes phrases as well as individual words. A phrase in a lexicon has the same conceptual status as a word and hence be characterized in the same way as a word. Recognizing phrases in text analysis is very difficult. Since this paper is not concerned with the actual mechanics of text analysis, use of the term phrases is avoided for the sake of simplicity of presentation.
2. Described also on the World Wide Web at http://www.cogsci.princeton.edu/~wn/, from which the database may be downloaded.
3. MCCA incorporates disambiguation procedures for assigning a single category when a word falls into more than one category.
4. A suite of programs for creating and maintaining lexicons for natural language processing, available from CL Research. Elaboration of the procedures used in this paper, applicable to any category analysis using DIMAP, are available at http://www.clres.com. These procedures describe the ordering of the steps, which steps can be performed automatically, how information is merged, and where human intervention is required.
5. Closed classes are syntactic categories, such as prepositions or pronouns, that have relatively few words and are unlikely to have new words. Open classes are nouns, verbs, adjectives, and adverbs; these classes expand as the language evolves.
6. The Special Interest Group on the Lexicon of the Association for Computational Linguistics maintains a set of links to publicly available lexical resources on the World Wide Web at http://www.clres.com/siglex.html.

References

Amsler, R. A. (1980). The structure of the Merriam-Webster pocket dictionary [diss], Austin: University of Texas.

Burstein, J., Kaplan, R., Wolff, S., & Lu, C. (1996). Using lexical semantic information techniques to classify free responses. In E. Viegas & M. Palmer (Eds.), Breadth and Depth of Semantic Lexicons. Workshop Sponsored by the Special Interest Group on the Lexicon. Santa Cruz, CA: Association for Computational Linguistics.

Chomsky, N. (1956). Three models for the description of language. IRE Transactions PGIT, 2, 113-124.

Chomsky, N. (1965). Aspects of the theory of syntax. Cambridge, MA: MIT Press.

CL Research. (1997 - in preparation). DIMAP-3 users manual. Gaithersburg, MD.

Davis, A. R. (1996). Lexical semantics and linking in the hierarchical lexicon [diss], Stanford, CA: Stanford University.

Evens, M., Litowitz, B., Markowitz, J., Smith, R., & Werner, O. (1980). Lexical-semantic relations: A comparative survey. Edmonton, Alberta: Linguistic Research, Inc.

Fillmore, C. J. (1968). The case for case. In E. Bach & R. Harms (Eds.), Universals in linguistic theory (pp. 1-90). New York: Holt, Rinehart, and Winston.

Jackendoff, R. S. (1972). Semantic interpretation in generative grammar. Cambridge, MA: MIT Press.

Katz, J. J., & Fodor, J. A. (1963). The structure of a semantic theory. Language, 39, 170-210.

Kucera, H., & Francis, W. N. (1967). Computerized dictionary of present-day American English. Providence, RI: Brown University Press.

Laffal, J. (1995, October). A concept analysis of Jonathan Swift's A Tale of a Tub and Gulliver's

Travels. Computers and the Humanities, pp. 339-361.
Levin, B. (1993). English verb classes and alternations: A preliminary investigation. Chicago, IL: The University of Chicago Press.
Litkowski, K. C. (1978). Models of the semantic structure of dictionaries. American Journal of Computational Linguistics (Mf.81), 25-74.
Litkowski, K. C. (1997, April). Desiderata for tagging with WordNet synsets and MCCA categories. 4th Meeting of the ACL Special Interest Group on the Lexicon. Washington, DC: Association for Computational Linguistics.
Litkowski, K. C., & Harris, M. D. (1997). Category development using complete semantic networks. Technical Report, vol. 97-01. Gaithersburg, MD: CL Research.
Markowitz, J., Ahlswede, T., & Evens, M. (1986, June 10-13). Semantically Significant Patterns in Dictionary Definitions. 24th Annual Meeting of the Association for Computational Linguistics. New York, NY: Association for Computational Linguistics.
McTavish, D. G. (1997b). Scale validity: A computer content analysis approach. Social Science Computer Review, this issue.
McTavish, D. G., Litkowski, K. C., & Schrader, S. (1997a). A computer content analysis approach to measuring social distance in residential organizations for older people. Social Science Computer Review, 15(2), 170-180.
McTavish, D. G., & Pirro, E. B. (1990). Contextual content analysis. Quality & Quantity, 24, 245-265.
Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., & Miller, K. J. (1990). Introduction to WordNet: An on-line lexical database. International Journal of Lexicography, 3(4), 235-244.
Nida, E. A. (1975). Componential analysis of meaning. The Hague: Mouton.
Quillian, M. R. (1968). Semantic memory. In M. Minsky (Ed.), Semantic information processing. Cambridge, MA: MIT Press.
Quirk, R., Greenbaum, S., Leech, G., & Svartik, J. (1985). A comprehensive grammar of the English language. London: Longman.
Roget's International Thesaurus (R. L. Chapman, Ed.) (5th). (1992). New York: HarperCollins Publishers, Inc.
Schank, R. C. (1975). Conceptual information processing. Amsterdam: North-Holland.
Schank, R. C., & Abelson, R. (1977). Scripts, plans, goals and understanding. Hillsdale, NJ: Lawrence Erlbaum.
U. S. General Accounting Office. (October 1993). Developing and using questionnaires. GAO/PEMD-10.1.7. Washington, D.C.
Whissell, C. (1996). Traditional and emotional stylometric analysis of the songs of Beatles Paul McCartney and John Lennon. Computers and the Humanities, 30(3), 257-265.
Winograd, T. (1972). Understanding natural language. New York: Academic Press.

Software Cited: DIMAP - DIctionary MAintenance Programs, utilities for creating and maintaining lexical knowledge bases, with integrated machine-readable dictionary, WordNet, and MCCA content analysis capability. Available from CL Research, 20239 Lea Pond Place, Gaithersburg, MD 20879 (Telephone: 301-926-5904; email - ken@clres.com; web site - http://www.clres.com).

Biographical Sketch: Kenneth C. Litkowski is the owner of CL Research. He has degrees in mathematics, law, and computer science and has worked extensively in qualitative research and computational lexicology. His interests focus on the design and development of computer-based tools for building lexical knowledge bases. He may be contacted at 20239 Lea Pond Place, Gaithersburg, MD 20879 (Telephone: 301-926-5904; email - ken@clres.com; web site - http://www.clres.com).

Table 1
Lexical entries: Example of semantic features

```
Word: #animal Type=r Code=#00026 No.Defs=1
    Sense: 1 Cat: nil
    Isa links:
        #creature d-0
    Features:
        EDIBLE = +boolean

Word: #creature Type=r Code=#00025 No.Defs=1
    Sense: 1 Cat: nil
    Isa links:
        #ind_obj d-0
    Features:
        AGE = +scalar
        SEX = +gender
```

Table 2
Lexical entries: Example of syntax and semantic roles

```
Word: eat Type=r Code=e00000 No.Defs=1
    Sense: 1 Cat: vrb
    Defin: ingest solid food through mouth and swallow it
    Isa links:
        #ingest d-0
    Features:
        root = $var0
        subj = ((root $var1) (cat n))
        obj = ((root $var2 optional) (cat n))
        AGENT = ^$var1
        THEME = ^$var2
```

Table 3
Lexical entries: Example of selectional restrictions

```
Word: in Type=r Code=i00000 No.Defs=2
    Sense: 1 Cat: prp
    Defin: located within the confines of
    Features:
        root = $var1
```

pp-adjunct = ((root $var0) (obj ((root $var2) (cat n))))
^$var1 = (*OR* +object +event) (location ^$var2 +physobj)

Sense: 2 Cat: prp
Defin: into the destination of
Features:
root = $var1
pp-adjunct = ((root $var0) (obj ((root $var2) (cat n))))
^$var1 = +event (destination ^$var2 +location (relaxable-to +physobj))

Table 4
Lexical entries: Example of semantic relations
Word: #event Type=r Code=#00012 No.Defs=1
Sense: 1 Cat: nil
Isa links:
#all d-0
Features:
SUBEVENTS = +event
SUBEVENT-OF = +event
TIME = > 0 (MEASURING-UNIT +second)
LOCATION = +place
CAUSED-BY = +event
CAUSES = +event
PRECONDITION = +event
EFFECT = +event

Table 5
Lexical entries: Example of knowledge base data
Word: #teach Type=r Code=#00014
Isa links:
#communicative-event d-0
Features:
AGENT = +intentional-agent (default +teacher)
THEME = +knowledge
BENEFICIARY = +intentional-agent (default +student)
PRECONDITION = (default (*AND* #teach-know-1 (NOT #teach-know-2)))
EFFECT = (default #teach-know-2)
SUBEVENTS = (*AND* #teach-describe #teach-request-info #teach-answer)

Word: #teach-answer Type=r Code=#00019
Isa links:
#answer d-0
Features:
AGENT = +teach.agent
THEME = +teach-request-info.theme

BENEFICIARY = +teach.beneficiary

Word: #teach-describe Type=r Code=#00017
Isa links:
#describe d-0
Features:
AGENT = +teach.agent
THEME = +teach.theme
BENEFICIARY = +teach.beneficiary

Table 6
Lexical entries: Example of lexical rules
Word: #grinding Type=r Code=#00032
Sense: 1 Cat: nil
Isa links:
#lexical-rule
Features:
0 = +count-noun (ORTH $var0) (RQS +ind_obj)
1 = +mass-noun (ORTH $var0) (RQS +substance)

Word: #animal-grinding Type=r Code=#00033
Sense: 1 Cat: nil
Isa links:
#grinding
Features:
0 = (RQS +animal)
1 = (RQS +food-substance)

Table 7
Lexical entries: Example of sense relations
Word: coffee Type=r Code=c00000 No.Defs=2
Sense: 1 Cat: nou
Defin: a kind of bean which is roasted and ground to produce coffee-2
Isa links:
#coffee-bean d-0
Features:
count = +
proper = -

Sense: 2 Cat: nou
Defin: a hot drink made from coffee-1
Features:
count = -
proper = -
Role:
#grinding coffee(1)