An Investigation of the Semantic Relations in the Roget's Thesaurus: Preliminary Results

Patrick J. Cassidy

As a first step in construction of a lexicon for natural language understanding, we are preparing a hierarchical semantic network using the Roget's thesaurus as a starting database. This work was undertaken because examination of the Roget shows that there are semantic relations considered important for linguistic expression which are not defined in other publicly available semantic networks, such as WordNet. In the process of conversion of the Roget to a semantic network, the first stage has been to reorganize the hierarchy and specify the set of semantic relations necessary to express those conceptual relations which are implied by the relative location of the words within the thesaurus. The explicit marking of semantic relations which are only implied in the original Roget has converted that reference work into a semantic network with a flexible multiple inheritance, which should greatly enhance the utility of the Roget for semantic information processing. We also expect that the resulting set of semantic relations will specify a minimum set required for definitions of words and logical representation of linguistic expressions at a human level of understanding. At the present stage, approximately 170 semantic relations have been defined to express the observed relations. It was found that many semantic relations observed in the Roget could not be expressed with the simple binary predicates often used for semantic relations, and it was found necessary to extend the notation to allow ternary and higher relations, as well as simple frames. At this stage the semantic relations thus defined have not yet been reduced to the first-order logical relations required to allow detailed inferencing, and the resulting semantic network has not yet been evaluated for its utility in practical applications. In future work, the semantic network must be enhanced by defining the semantic relations themselves in logical format, and additional semantic links must be added to distinguish non-synonymous words within paragraphs.

1. Introduction

In order to achieve computerized understanding of human language, the meanings of words and texts must be represented, within the computer lexicon, in a logical format usable by computerized reasoning processes. A critical part of the task of designing a concept-representation system is to specify the relations between concepts; and those ***semantic relations*** form an essential part of the definitions of words and concepts. However, there is no general agreement on the best method to represent the relations between concepts, nor on which set of relations is adequate to represent linguistic and world knowledge. There are numerous suggestions about which sets of semantic relations will be useful for knowledge representation [15, 16], but as soon as one attempts to define words and concepts in a manner that captures the nuances of how people use such words in language, the inadequacy of existing sets of semantic relations quickly becomes obvious. Our ultimate goal is to find a set of logical definitions of words which will allow human-level understanding of natural language while also being sufficiently well-defined so that the same concepts can be used for unambiguous machine-to-machine communication of complex concepts, as in database systems. For that purpose we need first to determine the set of semantic relations required to specify those definitions. This study shows that there are semantic relations used in writing in natural language that have not been discussed in previous literature, and suggests that these relations are likely to be important in capturing the nuances of meaning required for human-level understanding of language.

The optimal form of the lexicon required for machine understanding of language has been the subject of a great deal of research and discussion, but in the years since Quillian first studied the properties of concept networks with semantic links [1], most efforts at knowledge representation for language understanding have used some form of representation in which concept nodes are connected by links representing some type of relationship between the concepts. In many systems the links are defined as ***semantic relations*** and distinguished from the concept nodes, but in other systems (such as SNePS [2]) semantic relations are viewed as another type of node. When a representation formalism also has a defined method for combining (or "unifying") concepts to form larger concepts (as with conceptual graphs [3]), it may also serve as the logical representation for an assertion or an entire discourse. Most semantic networks have been created using semantic relations which are not fully defined logically to allow unambiguous inferencing to be performed, and the FACTOTUM semantic network also lacks full formality in that regard. In contrast, the CYC system places heavy emphasis on reasoning with the semantic relations at an early stage, but CYC has not been demonstrated to be notably useful for language understanding tasks. With language understanding as the main goal, it was nevertheless considered important to determine which semantic relations play a prominent role in the type of thinking that humans perform in linguistic composition, a task for which the Roget's Thesaurus has been much used. At the first stage it is necessary to recognize semantic relations at a level close to the linguistic, and subsequently these relations may be defined in more logical detail to allow the precise inferencing necessary for human-level language understanding.

Construction of a conceptual semantic network presents several major design decisions: (1) defining the structure of the hierarchy; (2) selection of a set of semantic relations to relate the concepts; and (3) specifying the logical operations which operate on the network. For much of the work in knowledge representation, for example in the KL-ONE family of languages [4], the emphasis has been on finding a representation that will allow inferences which are complete and tractable. Thus the third design factor has been of greatest concern. In designing a semantic network for language understanding, other concerns, such as utility in word-sense disambiguation, may be viewed as more urgent, and the emphasis will correspondingly shift to design factors 1 and 2, even though the ultimate purpose of a conceptual semantic network is to support inferencing in the process of language understanding.

A distinction may be drawn between a computer lexicon containing syntactic, morphological, and collocational information, and a semantic network containing only conceptual information. A further distinction is often made between a terminological database, containing only word definitions, and an assertional database, containing facts about situations or events in the "real world". In practice, it is probably impossible to include enough information about the meaning of words to permit human-level understanding of a text, unless a substantial amount of "assertional" world knowledge is include in the database. We therefore do not try to enforce a rigid separation between terminological and world knowledge, though the emphasis is on including only sufficient information to allow understanding of word meanings. We assume here that eventually all information about word usage, grammatical, definitional, and practical uses, will be combined in a single database, and for convenience we will also call that a semantic network.

In all semantic networks, of special concern are the hierarchical links, representing membership in classes. Because of the transitivity of class membership, such links allow the use of inheritance of properties, permitting a more compact and more easily maintainable form for the lexical database. But even with respect to this most fundamental semantic relation, differences among researchers appear, for example, on the degree to which the inheritance may be defeasible, and, importantly, as to which classes (or types) should be defined in the hierarchy. The best set of categories to represent the world has been viewed as very problematic, since different goals for a knowledge representation appear to dictate different methods of splitting concepts into subtypes. One solution, which we adopt here, is to allow multiple inheritance, which allows different users to define special concepts representing aggregates of concepts which may be related in specialized contexts.

However, even if a uniform representation format is used, the resulting categories in different classifications may be so divergent that it may be impossible to transfer concepts between two different systems. There is therefore a strong incentive to try to find areas of agreement to advance the process of developing a standard ontology, at least for the fundamental defining concepts that can be used to create more complex concepts. Since there are aspects of semantic relations that have not been fully treated in the existing literature, we hope that the examination of semantic relations in this study will provide additional data to allow a standardized ontology to be eventually developed and adopted by cooperation among different groups, and that is fully adequate for the task of language understanding.

As a practical matter, most efforts at building practical language understanding systems have concentrated on specific subject matter, where the most important concepts can be represented in sufficient detail to achieve a useful level of understanding [e.g. see 5, 6, 7]. However, the question raised by the specialist approach is whether systems designed in isolation will be able to communicate information between them. The evidence suggests not. Ideally, all knowledge representation systems might use a common general hierarchy and a standardized set of semantic relations, creating new concepts or aggregates of existing concepts as needed for specific applications, but preserving a large area of commonality allowing efficient communication. Some requirements for the development of such a general ontology have been suggested by Gruber [8], and by Bateman [9]. However, skepticism about the possibility of agreeing on such a general ontology is often expressed[1], and even the utility of a general ontology has been questioned [10]. Such skepticism may be due at least in part to a paucity of general ontologies available to be tested in specific applications.

The development of separate representation languages for different applications might not prove a fatal barrier to communication between systems if some method of knowledge conversion is available. One project directed at developing a knowledge-conversion method is the KIF project [11], which is developing specifications for knowledge representation which would allow transfer of knowledge from one formalism to another. The proposed KIF standard has been coordinated with a corresponding conceptual graphs (CG) representation standard [41], and the two standards represent alternate linear and graphical methods for representing knowledge, which are interconvertible with each other. These two representation standards are based predominantly on first-order logic; however the use of these standards for knowledge representation is effectively theory-neutral, and merely provides a method for recording knowledge. The methods for use of the knowledge thus represented need not involve first-order logic, and is unrestricted.

---

[1] An active discussion of such issues can be found via the Conceptual Graphs listserver. Inquiries may be made to cg-request@cs.umn.edu.

Although the KIF and CG standard provide a ***format*** for recording knowledge, there is no corresponding standard for the ***content*** of a common knowledge base. Without some agreement on the type hierarchy and defined semantic relations, the ability to transfer knowledge even in an agreed common format will be severely impeded. Even where hierarchies of two systems look similar, the absence of similarity in the semantic relations used may make it impossible to determine if two concepts are in fact identical, preventing meaningful merger of two knowledge systems. There has been a recent study to determine if a standard ontology can be developed by merger of existing ontologies [39].

Agreement on the basic outlines of such a common semantic network would enhance the utility of a knowledge interchange format such as KIF or CG. Among current efforts to develop large semantic networks, the CYC project [22] and the Japanese EDR project [34] stand out by virtue of the size of the effort expended. The Princeton WordNet system [14] also has a large semantic network, although it uses fewer than fifteen types of semantic links. The WordNet system is being replicated in other languages within the European Computational Linguistics community, which is developing a set of semantic networks in several languages, collectively called Euro WordNet [37].

One concept classification system, Roget's Thesaurus, has been in use for literary composition for over one hundred and forty years. Developed for purposes quite different from computer communication, it nevertheless contains a wealth of explicit and implied information about English words and their underlying concepts, which could be usable in a formally defined semantic network for use by computers. After some examination, it was apparent that the process of extracting implied semantic link information from the Roget into computer-usable form would be very difficult to perform automatically, and would likely be performed more accurately using inspection by a human interpreter to specify the proper location for each concept in a semantic network. Thus the present work was undertaken to convert the information in the Roget's thesaurus [12] into a hierarchical semantic network. The practical utility of the resulting network must be tested in real language-understanding applications, but the tests cannot be performed realistically until all of the words to be found in the test texts are properly classified within the network. The potential of Roget's Thesaurus in its original form as a basis for merging specialist thesauri has previously been discussed by Liddy et al [30].

The completely manual construction of a full semantic network even for only the most common words (e.g., those recognized by a word processor spell-checker) is a very labor-intensive task, requiring probably several hundreds and perhaps over a thousand person-years to enter the most important semantic relations for each word. Thus the present work is not intended to construct a complete semantic network, but to develop the basic outline of such a network to a point where it may be tested for utility in language understanding tasks. It is hoped that subsequent to the initial creation of this bare framework, later stages of enhancement and supplementation will be accelerated by the use of automatic processing of large text corpora and dictionaries.

It remains to be determined how much hand work will be required to get to the point where automatic methods will be able to extract most of the necessary information from dictionaries or free texts. It seems clear that some degree of hand-encoding of word meanings will be needed to bring our automatic systems to the point where they can extract the remaining information, and this work is undertaken with the conviction that a bare outline of a semantic network such as might be developed from the Roget thesaurus is the minimum hand-encoded information that will be required to make the automatic extraction of additional necessary details feasible.

2. Method for Extracting Information from the Roget

We viewed the Roget as a partly-constructed outline of a semantic network, needing some rearrangement in order to make the hierarchy useful for inheritance of conceptual properties. In addition, the semantic links in this network were only implicit, and needed to be marked explicitly. After initial examination it was decided that to mark the semantic relations, we would use only one of the several lists of semantic relations previously proposed, namely that of the UMLS [19].

The Version of Roget's Thesaurus used for this investigation was that published in 1911 [12]. In the course of this work, some additional vocabulary has been added, but no systematic supplementation has yet been undertaken, and this version is deficient in modern technical vocabulary[2]. This version nevertheless represents a complete human language adequate for communication of ideas, definition of new concepts, and learning, using only the natural language itself. The number of headwords, originally 1043 in the 1911 Roget, has been approximately doubled in the present version of the FACTOTUM SemNet, but most of the added headwords were for technical topics. Once a classification of the most general concepts has been completed, supplementation with modern vocabulary should be relatively straightforward.

[2] It was necessary to use this early version due to the inability to obtain an appropriate license from the copyright holder of recent editions.

The Roget's Thesaurus was first published in 1852, and the original classification scheme was retained essentially unchanged through the fourth edition of 1977. A recent fifth edition modified the classification scheme, eliminating the hierarchy. The electronic version of the 1911 edition was prepared by us and used as a simple word-processor file, and modifications were made using a commercial word processor (Microsoft Word).

The 1911 thesaurus was organized in a quasi-hierarchical fashion, with six top categories (Abstract Relations; Space; Matter; Intellect; Volition; and Affections) branching in a shallow tree to about 1050 headwords, all nouns. In many cases the underlying concept of a headword might more appropriately have been categorized as a verb or adjective, but the classification proceeded from the nominalized forms. We have retained the classification by nominalized forms. Wherever possible, the Roget juxtaposed a concept with its antonym. Each headword formed the title for a ***main entry*** containing words conceptually related to the headword, grouped as nouns, verbs, adjectives, and adverbs. However, the conceptual relations within each main entry were quite varied, and were not explicitly marked. Thus subtypes of a headword were not distinguished from parts, causes, or metaphorically related words, although there is a tendency for words with a specific semantic relation to the headword (such as human types who perform certain roles) to be grouped together in a paragraph. It is apparent that there is a large amount of intuitively valid semantic structure in this lexical database, and the task we undertook was to extract that information into a form usable by a computer program.

The main tasks in preparing a semantic network from the Roget were; (1) to modify the hierarchy to allow optimal use of inheritance; and (2) to define and mark the semantic relations between the headword and the related words within that main entry. The work was aided by a custom-designed indexing and viewing program ("Thesview"), which was written for the purpose by Dr. Alexander Gelbukh. This program allows one to rapidly find an index word, distinguish the word senses, and directly view the text of the thesaurus at the point where that word appears. The hierarchy path above the index word can also be viewed.

As discussed above, there is a common perception that there is no unique universally valid hierarchy (ontology) that is likely to be accepted by all computational linguists, and we do not suggest that the hierarchy devised by this procedure is necessarily superior to others created for other purposes. On the other hand, the dense store of implied semantic relations within each main entry of Roget provides an unusual and valuable resource for discovery of semantic relations which may be important in understanding language, but may have been overlooked by research to date.

The semantic network being constructed by this process is named the FACTOTUM(R)[3] SemNet. When the first version of FACTOTUM has been completed, it will still be merely a skeletal outline of what a fully-connected semantic network must be, in order to be truly useful in language understanding. However, it may nevertheless have some utility even at an early state. We believe that Minsky's view of studies of the human mind, "until you've seen some of the rest, you can't make sense of any part."[13], applies equally to understanding the meanings of words, which are intimately connected to others by multiple semantic linkages.

2.1 The Hierarchy

The guiding principle in revising the Roget hierarchy was to allow maximum use of inheritance in defining words. This is convenient both for compactness of representation, and for ease of maintenance of the lexical database. At present we have no automatic procedures to measure compactness, and decisions on hierarchical organization were made solely on the judgment of the author. To optimize utility of the inheritance function, the inheritance of properties is considered in general non-defeasible unless negated by an explicit exception. The semantic network is intended only as a terminological, or type network, and does not purport to contain assertions about objects in the real world (as in the KL-ONE or CYC assertional languages). Nevertheless, the relationships among types mirror those among real-world objects, and in some cases the expression of these relations can require complex frame-like descriptions.

For the purpose of exploiting inheritance, there did not appear to be any benefit in distinguishing ordinary type concepts from roles, since there was no consistent property of the words usually thought of as "roles" (mother, president, student, plumber, carburetor) which could be ascribed to all of them. It appears that the intuitive concept of "role," even if applied only to humans, specifies a heterogeneous collection of concepts which, to a greater or lesser degree, strongly imply that the individual "playing a role" has a necessary link to some other object. But this phenomenon of necessary link is precisely what is intended by all of the semantic relations connecting concepts in the FACTOTUM SemNet, and the distinction between the "role" link and other links is difficult to specify precisely. Among the sets of concepts, however there are some which also carry the attribute "role," namely those which are commonly used in the linguistic expression "x of y", where "x" has the attribute role, and where "x" can also specify an individual (concrete) object, and "y" is also an object. We anticipate that in reasoning with

[3] FACTOTUM(R) is a registered trademark of MICRA, Inc.

this semantic network, the links implied by the "role" attribute will be stronger, in some sense, than the average semantic relation.

For many of the concepts, the hierarchical links as well as the other semantic relations will constitute necessary but not sufficient conditions for definition, thus forming only partial definitions of the concepts in the network. Those concepts which cannot be completely defined solely in terms of other concepts are considered the "primitive" concepts of the system. Although in its mature form, there may be fewer than three thousand true primitive concepts, at the first stages there will be few completely defined concepts.

It was apparent at an early stage that multiple inheritance is the most natural means to represent many of the relations implied in the Roget structure, and this also allows the flexibility to create specialized aggregates of concepts which are useful for specific practical applications. The resulting hierarchy therefore forms a forest, rather than a tree structure. Because of the one-dimensional nature of text, the main hierarchy displayed on paper does not graphically show the numerous secondary hierarchical links, which are logically equally important. After the first version is completed, tests must yet be performed to verify that there are no cycles within the hierarchy.

The hierarchy resulting from the modification of Roget retains a large proportion of the Roget structure, but many segments were shifted to conform to the desired goal of maximizing inherited characteristics. Rather than have separate hierarchies for nouns and verbs as in WordNet [14], we retained the Roget structure of a single hierarchy based on nominalized forms. However, certain subdivisions, such as the sections on actions or interactions of physical objects, might be viewed as principally verb hierarchies. Likewise, certain subdivisions can be viewed as principally adjectival hierarchies, as for properties of physical objects, or human traits.

Starting from Roget's organization, the most natural division appeared to suggest three main top categories. These are (with examples of subtypes):

(1) abstract relations (existence, state, relation, order, sequence, classification, quantity, number, cause, change).
(2) physical world concepts (time, space, shape, location, matter, physical objects, substances, artifacts, physical action, event, motion, physical sciences, life, animals, and medicine).
(3) mental concepts (mind, knowledge acquisition, sensation, ideas, knowledge, thought, emotion, volition, habit, human traits, motive, human goals and behavior, acting, plans, competition, communication, memory, truth, government, law, religion, possession, finance).

The abstract relations include concepts neither principally related to the material world nor the mental realm, but applicable in either category. The mental division is intended to include all concepts which would not exist without minds to invent them. Some subdivisions of the physical division, such as medicine, may have a large component of "mental" character, being largely goal-oriented. But the actions and objects of the actions are predominantly physical in character. Thus this category was included with the physical sciences. Other decisions, especially in the mental realm, may appear arbitrary, and we would expect independently designed hierarchies to have a different structure. In cases where it is not obvious, the reasons for placement of specific categories at specific points in the hierarchy are discussed in comments within the semantic network file itself.

The apparent arbitrariness of the divisions in such a hierarchy is substantially ameliorated by allowing multiple inheritance. Where a certain application may find it beneficial to aggregate a group of concepts under a single category, there is no reason to prevent creation of a new category to accomplish that goal, with all of the desired categories as subtypes. In this way, we may visualize at some point that a general semantic network may be usable by a large number of applications, with specific applications needing only to add details or create a relatively small number of new categories to form aggregates convenient for that application.

The optimal form for such a general ontology will most likely be determined by testing different ontologies in a variety of applications, thereby learning which structures are useful and which are not. An amalgam of ideas from different ontologies may ultimately provide the best hierarchy for diverse applications. This work was undertaken in the anticipation that the Roget Thesaurus, and the derived FACTOTUM hierarchy, may have useful structure to contribute to a general ontology.

2.2 The Semantic Relations

There were several lists of semantic relations available for use in specifying the relations perceived in the Roget structure. A collection of articles discussing semantic nets and their relations can be found in reference [15]. In addition to the hierarchical relations isa/subtype, a number of other relations have been proposed to express the connections between concepts which are used by people in understanding language. Markowitz et al [17] proposed a classification of semantic relations. Chaffin and Hermann [35] present data to show that semantic relations are not simple primitives, but have more complex analyzable structure, and they classify relations according to such structure. Mel'cuk has proposed about 60 relations (called Lexical Functions) required to define words with sufficient precision to represent all of a native speaker's knowledge of a word [20]. Dahlgren [16] lists 54 "feature types" for nouns which appear to be salient characteristics associated by people with various objects. Other than synonymy and antonymy, the most commonly discussed are the part and causal relations. Parts and causes are included, for example, in the WordNet system. Both part and cause have been split by many workers into more specific types; Iris et al. [18], for example, give evidence for four main subdivisions of the part relation. Motschnig-Pitrik and Kaasboll [36] present further discussion of the part-whole relations in the context of object-oriented techniques. Our intention was to discover what relations are present in the Roget's thesaurus, using the thesaurus as a database in which the relations are implied. However, we also wished to keep our relations compatible if possible with the UMLS metathesaurus [19], and for that reason we started with the 30 relations defined within the UMLS, which include part and cause. This initial set of relations was divided into more specific types, or new relations were added as required to express the relations implied by the relative locations of words in each main entry of Roget. For the purpose of discerning relations between words, the author relied upon his own understanding of the concepts involved, consulting commercial dictionaries to resolve uncertainties. As has been discussed by Chaffin and Hermann [35], adult human informants are typically able to recognize relations between words, even when the relations are complex concatenations of simpler relations.

The explicit relations that were marked in the Roget as a result of this work were extracted from a pre-existing text, in contrast to the alternate experimental methodology of querying human informants directly. They nevertheless are likely to have substantial "psychological validity," since the text is of a special form for which the authors had the explicit purpose of juxtaposing words which were, according to their intuition, sufficiently related to have potential utility in composition. However, those authors are no longer available for interrogation concerning their methodology.

As with the UMLS relations, each of the semantic relations suggested here has an inverse relation, although in some cases the relation is symmetrical, *i. e*. the same relation in both directions (e.g. antonyms). The initial set of suggested relations developed from this work should be viewed only as a first approximation, to be refined by more precise definition, and perhaps further subdivision.

The ***format*** of the semantic relations differs from that commonly used, so as to be easily parsed from within the complex word-processing text which was used in effect as the main database. The format also presents some flexibility in expression which was found useful for relations more complex than simple binary relations. A set of double braces was used for binary relations, and triple braces for ternary relations.

In a few cases it appeared desirable to relate two different concepts by more than one semantic relation. For example, in the section dealing with embarrassment, it appears that the feeling of embarrassment has two distinguishable connections to the matter which caused the embarrassment:
{{has_topic(embarrassment)}} {{mcaused_by(embarrassment)}}
source of embarrassment.
These express the notions that a feeling of embarrassment is a mental activity with an external referent (the topic), and also that that topic is the cause of the feeling. Such distinctions in some cases seem necessary to represent the relations between those concepts.

As with WordNet, the individual concept entries, whether individual words or phrases, were organized in synonym groups, and the semantic relations are considered as holding between any word in one synonym group and any word in the related synonym group. It has been remarked that synonymy has varying degrees (see, for example, chapter 12 in Cruse [33]). We use the loosest definition of synonymy, in which the ability to substitute one word for another in at least one context makes two words synonymous. In cases where a general concept is usually expressed in a specific context by only one word, we use a contextual relation (see (5) below) between the base concept and that specific word, rather than including that word in the main synonym group.

The notation was designed for use in the linear-text format of a word processor file. A typical binary relation holds between two categories (and by implication, the members within those categories). Thus the relation:
{{causes(destruction): nonexistence}}
states that destruction causes nonexistence. The patient of the action is in this case implied, and the (as yet undefined) procedural implementation of the relation is responsible for handling such implications. As can be seen from this example, even after recognizing and expressing a semantic relation between two concepts in this way, the practical utility of such a database of relations depends upon a careful and precise definition of the semantic relation. Although Chaffin and Hermann [35] present strong evidence that some semantic relations may be decomposed into more fundamental elements, at present we view the semantic relations as a special subset of the "primitive concepts", each of which must have some unique procedural code which interprets the use of that concept in the context of the surrounding discourse, and creates the logical structure expressing the meaning of the text, under control of a language-understanding system. The "meaning" of each semantic relation will thus be expressed by first-order-logic assertions or at least partly in procedural code, although there is likely to be much structure in common between many semantic relation definitions. We did not expect to find an unambiguously unique set of semantic relations; the question of whether to split a

particular semantic relation, such as part, or cause, into more specific subtypes depends upon how one intends to divide the effort of language understanding between the lexical database and the procedural definition of the semantic relations. The more specific a semantic relation is, the less detail should be needed in its corresponding definition.

In the KL-ONE knowledge representation system, the relations between concepts may take arbitrary forms, and there is no single privileged set of defined semantic relations. We felt it desirable to restrict the number of defined semantic relations to the smallest set which is necessary to express the semantic relations observed.

There were, however, specialized cases where it appeared most natural to define a semantic relation which would be used only one or a very small number of times. In these cases it seemed more appropriate, rather than to proliferate the general list of defined relations, instead to provide a mechanism for the *ad hoc* definition of a semantic relation for such specific cases. In addition to minimizing the general relation list, this allows individuals constructing a specialized application to define semantic relations without fear of such relations being inappropriately used for other cases.

As another consequent of the goal to minimize defined relations, one relation which was found to be valuable was the "property" relation. Rather than define a predicate such as

red(x) or **is_red(x)**

to assert the redness of an object "x", we use the more general predicate "has_property":

{{has_property(x): redness}}.

In fact, for the common properties of color, size, and weight for physical objects, such relations were usually expressed in a slightly different way, which required the modification of the simple binary predicate, as illustrated below for the color of gold. In this way many adjectives can be defined as properties, avoiding an uncontrolled proliferation of semantic relations which are mostly duplicative of adjective definitions.[4]

---

[4] One practical advantage of maintaining a small set of relations is that they can be more easily remembered and used by numerous individuals adding data to the semantic network, thus minimizing the problem of maintaining consistency within a large development effort.

If the semantic relations are themselves maintained within a well-defined hierarchy, it will be possible to allow specializations of each relation, to express that relation in specific contexts. The disadvantage of allowing such proliferation of relations is, as stated, the increased steepness of the learning curve for individuals who wish to add entries to the semantic network. If the relation

It was also considered desirable to the extent possible to define only binary predicate relations, which are most likely to be transferable between different programs. However, in the course of processing the Roget Thesaurus, it became apparent that a proper expression of the relations between words in many main entries was most naturally achieved by allowing some modifications of the standard binary predicate. The most frequently used modifications were:

(1) explicit ternary predicates required to express, for example, "betweenness" and ratios:

> **{{{between(river): left bank + right bank}}}**
> **{{{has_relative_value(dollar + cent): 100}}}**.
> The triple brackets are used for explicit ternary relations, and often the order of the arguments separated by a plus sign is significant.

(2) argument modifiers, used to modify the meaning of one of the arguments, attached to the argument modified. In these cases, the semantic relation usually has its full meaning only when all parts of the relation, including the modifier, are included:

> **{{has_part(bicycle): wheel[num=2]}}**
> meaning: *a bicycle has two wheels*
> **{{has_property(gold): color[val=yellow]}}**
> meaning: *the color of gold is yellow*.

The presence of such modifiers within relations make them in effect of arity greater than two, but they are distinguished from explicit ternary relations in that the modifiers can be viewed as attached to only the second argument of the predicate, rather than to both arguments simultaneously. Considerations concerning the most appropriate formalism for treating such multiple arity predicates have been discussed by Lenat and Guha [22]. For the initial version of this semantic network, we have chosen to keep the actual expression of the relations as close as possible to natural language expressions.

---

hierarchy is properly defined, it would not be an error for a knowledge-enterer to use a more general relation for the same purposes as a subtype specific relation. The use of the more general relation would (1) increase the amount of computation required by the system at run-time, and (2) increase the chance of error in the inferencing process. However, designing the set of semantic relations to allow such substitutions will permit individuals to begin contributing data to the system at an earlier point. This is our goal.

Note that the "property" relation is very flexible when combined with such modifiers. This is the method which allows a relatively small number of defined relations (less than 200) to express a very large number of predicates, which might otherwise require some very specialized predicates. As an example, in their discussion of CYC, Lenat and Guha [22] mention a predicate "surprisingTo", as a predicate modifying a fact in the database, for which predicate the arguments are the people to whom the fact is surprising. In FACTOTUM, the same concept can be expressed, but in a different fashion. The adjective "surprising" would be treated as a property of an assertion, in this manner:

{{has_property(assertion_32): surprising[to=Cassidy]}}

The concept "surprising" is taken from its definition to govern two cases, the fact which is surprising, and the cognitive agent(s) to whom it is surprising. The interpretation of each "has_property" predicate will depend upon the "property" which is predicated, which will be almost any attribute which can modify an object. For example, the procedural code for interpretation of the property "surprising" will look for the two arguments, fact (of which it is a property) and an agent. The code also knows to look for the "to" case marker, to indicate the agent. When called to interpret this semantic relation, the procedure for "surprising" will construct the logical structure with fact and agent in the proper location. If similar predicate names are used in other databases, using an adjectival property with or without a fused case marker, it is likely that such predicates will be translatable automatically to the {{has_property}} format of the FACTOTUM database. In many cases, we may expect that a similar compositional approach will work equally well in languages with other structures, such as inflectional languages with morphological variation indicating the cases, and with, for example, Japanese, with postpositional particles as case markers. Thus allowing such compositional modifiers appears to serve the goal of minimizing unnecessary proliferation of semantic relations without reducing expressiveness of the notation.

The modifiers within the square brackets will, if not simple case markers, probably take on the nature of "primitive" semantic relations themselves, requiring definition. The "[val=x]" modifier, for example, would need to be able to match values of attributes with the range of values which such attributes can take, distinguish between quantitative and qualitative values, and recognize measurement units. It may also be efficient to allow such a predicate modifier to determine whether the value is outside the normal range of values for the object whose attribute is being predicated. This checking process would occur at the time of compiling the semantic network from text format into logical format.

(3) the use of predicates as arguments within predicates,
particularly to express functional relations:
{{has_function(cannon): propel[obj=shell]}}
meaning: *a cannon is designed to propel a shell.*

(4) predicate modifiers, used as a very rough quantification of the frequency or degree of certainty of a relation:

(no modifier)	*by default*
&	*sometimes*
!	*almost always*
!!	*holds by definition*

For example:

{{&caused_by(similarity): imitation}}

meaning: *similarity is sometimes be caused by imitation, but sometimes by other causes*.

The inverse, however, does not have the qualifier:

{{causes(imitation): similarity}}

meaning: *imitation almost always causes similarity*

This illustrates one difficulty in automatically generating inverses from semantic relations. A relationship which may be strong in one direction may be weak in the inverse direction. Most semantic relations in FACTOTUM have thus far been marked only in one direction, and it is likely that determining the level of certainty for the inverse relation will require manual checking.

Two other classes of predicate modifiers are allowed:

Negation. Any predicate with a "not_" modifier attached to it asserts that the relation does not hold.

e.g. **{{has_subtype(graph)}} {{not_part_of(cycle)}} acyclic graph**.

"*An acyclic graph is a graph which does not contain a cycle*"

Disjointness: An "_x" modifier after the predicate asserts that the list of arguments is exhaustive and disjoint.

e.g. **{{has_subtype_x(maximum)}} local maximum; global maximum**.

"*The only subtypes of maximum are local maximum and global maximum*"

Past event: A "_p" modifier after a predicate asserts that the action or property specified occurred in the past.

e.g. **{{property_of_p(married): widow}}.**

"*A widow had the property of being married at some time in the past*"

(5) Contextual relations were also deemed necessary. These are not simple modifications of predicates, as they are not in fact predicates. They express the fact that certain concepts in certain contexts are expressed with specific words. The notation uses a different bracket combination from predicate relations. Thus:

{{has_subtype(render dry)}} {[with_object(corpse)]} mummify.

meaning: *to dry a corpse is expressed as "to mummify"*

Some of these contextual relations are similar in function to the "lexical functions" proposed by Mel'cuk [20].

A number of other deviations from simple binary predicates were found necessary to express the relations between words juxtaposed within each main entry of Roget, but many of these were used for only a few instances. A full list of the relations and their definitions is too long for this article, but is available to interested parties[5].

3. Control of Inference

The purpose of finding the semantic links is to create slots attached to concepts encountered in text, which will allow resolution of references to concepts not explicitly mentioned in the text, as well as inferences such as cause and consequence, in the manner of frame-based and script-based reasoning. The links will also assist in word disambiguation. However, the problem arises as to how far it is necessary to proceed through the chain of linkages in order to achieve the desired level of understanding.

In the CYC project, Lenat reports that their system can follow a chain of links to the sixth level. Depending on the average number of links attached to each concept, proceeding to the sixth level could itself create serious combinatorial problems. For example, if there were an average of six relations attached to each concept, bringing into memory all concepts attached, down to the sixth level, could in the worst case bring an intractable number (6^6= 46,656) of other concepts into active memory, although cycles in the linkage paths would be likely to reduce this number substantially. Clearly, there is a need for some control on how deep the inferencing goes. In CYC, Lenat and Guha [22] describe the use of different functions ("Get0", Get4", etc.) which carry the inferencing to different levels, and

[5] The FACTOTUM Semantic Network, the hierarchy of main entries, and the list of semantic relations are not presently available on the internet, but will be sent by e-mail in response to requests to the author. Paper copies, CD-ROM, or DOS format disks will be supplied to those without e-mail access. The SemNet text and viewing program are copyrighted, but may be used freely for research purposes.

in fact use different sets of inferencing mechanisms. From our examination of the linkage paths in the Roget, it is apparent that different types of links have different "strengths", that is, different links are essential to differing degrees for understanding the concepts to which they are attached. This observations suggests that different types of semantic relations should have different designated strengths attached as attributes of the relations, and these strengths could be used to terminate chains of inference at different levels. In addition, certain individual semantic relations appear essential for the definitions of some concepts, and less important for others. Thus a uniform "strength" for any specific type of semantic relation may not suffice to properly control the inferencing. For this reason, the relations essential to the definition of a concept are prefixed with a "!" symbol to allow differentiation from less essential relations, and to assist control of inferencing. The strengths of relations are not numerically defined in the present version of the SemNet.

The observations made in the course of this study suggest that grasping the basic meaning of different concepts requires traversing the network linkages to differing depths. Unless a system using this network has sufficient speed to explore all concept links to the maximum required depth, the control of such exploration seems likely to be the main challenge in using this network.

4. Related Work

A number of studies have sought to inquire into the variety and meaning of semantic relations for use in coding a lexicon [15, 21], and many semantic relations have been proposed. However none of these studies have systematically explored the Roget's thesaurus as a method of discovering such relations.

The most widely used general ontology available to the research community is the WordNet system [14] developed by the Princeton group under Prof. George Miller. This is a large and mature system, with broad word coverage of modern English, and relatively complete with respect to its basic functions. It has been investigated for various purposes by several research groups, and is being incorporated into other lexical databases. FACTOTUM is still at an early state of development, and is less complete in its word coverage and is both less complete and less consistent in use of the defined semantic relations. After completion of the first version of the FACTOTUM SemNet, we anticipate that there will be much useful data in WordNet which can be adopted to supplement the present work.

Within the hierarchies of WordNet (version 1.4) and FACTOTUM there are many categories which are similar, but numerous differences also are present. A detailed comparison of the two cannot be performed in this space, but the most important general differences are:
(1) FACTOTUM has a single hierarchy organized by nominalized forms, and WordNet has two hierarchies, for the nouns and verbs; and the adjectives are treated separately.
(2) The number of semantic relations (not counting inverses) used in WordNet is presently 7, compared to 150 in FACTOTUM. Even so, WordNet has eight functions, such as entailment, verb case frames, and a meronym hierarchy, which are absent in FACTOTUM, and WordNet also contains definitions.
(3) The word coverage of WordNet is significantly larger than that of FACTOTUM; for individual words (not counting multi-word phrases) FACTOTUM contains 42,000 and WordNet 56,000, with only 26,000 in common. Thus there are 30,000 individual words in WordNet not present in FACTOTUM, however, a meaningful comparison of word stems (as contrasted with the orthographically different words) is made difficult by the variable number of morphological variants present in FACTOTUM.

The PENMAN Upper Model [9, 24], is an ontology developed for the PENMAN language generation system. Accordingly, the categories in the higher levels are motivated in part by an attempt to predict surface syntactic structure, rather than relations between concepts, and are quite language-dependent. The hierarchy diagram therefore has many top categories different from that of FACTOTUM, although the categories several levels lower have some points in common. The two-place relations also have many concepts in common with the semantic relations of FACTOTUM. Although the PENMAN and FACTOTUM hierarchies serve quite different purposes, points of similarity suggest that it may be possible to use both in a single language-understanding system. Where semantic characteristics are reflected in syntactic structure, we can expect some categories to appear similar in PENMAN and FACTOTUM.

The CYC project [22] is a large industrial development effort to build a knowledge base to overcome brittleness in expert systems. This project has been underway for over fifteen years, and in the process has changed its approach in several respects. The current "knowledge base" is quite large; the scale and breadth of the CYC project is many times larger than FACTOTUM. CYC has developed from defining general concepts to the point of defining individual contexts, called "microtheories" which can be accessed for specialized reasoning within the general framework [33]. The most common unit is the ***frame***, and reports indicate that several thousand "slots" (serving the same function as semantic relations) have been used to relate concepts within frames. Because it is an assertional knowledge base (as contrasted with the primarily terminological network of FACTOTUM), the logical relations used can be quite complicated, and are expressed in a notation similar to predicate logic. Some relations are rather specific, such as "ComputersFamiliarWith" for people. Examination of some of those apparently specific relations suggest that they can be represented by combinations of relations as used in FACTOTUM. Thus the differences may not be too great to allow substantial transfer of knowledge across these two notation systems. The hierarchy of CYC is also unusual, with "tangible objects" under the term "process", and "intelligence" as a sister node to "tangible object" under "individual object", disjoint from both the "intangible" and the "represented thing" categories. It would be a challenge to relate knowledge between such disparate ontologies as CYC and FACTOTUM.

The UMLS semantic network [19] is focused on medical topics, and does not treat general concepts in significant detail. There are fewer than 200 categories linked by the semantic relations, and all of the concepts in the UMLS metathesaurus are subsumed by one of those categories. Though we started from the UMLS set of semantic relations, the semantic relational terminology in FACTOTUM has diverged somewhat from UMLS, especially by splitting UMLS relations into more specialized subtypes. Thus the merger of the UMLS thesaurus with the FACTOTUM SemNet is likely to require substantial manual effort.

A group at New Mexico State University has developed a semantic network called Mikrokosmos [38], which is oriented toward machine translation, and has been developed incrementally over a number of years. Another group directed by Eduard Hovy at the Information Sciences Institute of the University of Southern California has explored methods for combining several existing ontologies to make progress toward a standard ontology. This work [39] explored similarities in the SENSUS [40] and CYC [22] ontologies, and prepared an aligned, merged ontology that could serve as the beginning of a standard upper ontology for general use. In neither of these systems is there a major emphasis on developing a sufficient set of semantic relations to permit nuanced definitions of words.

Other semantic networks have been mentioned in the literature, such as the Japanese EDR project [34] and the CODE system of Douglas Skuce [23], but the published information is insufficient to allow a meaningful comparison.

5. Evaluation of and Uses for a Semantic Network

The ultimate utility of any given semantic network must be judged in comparison with alternative general ontologies in specific applications. Preliminary measures of utility may be provided by developing a metric which will allow some judgment utility in limited tasks. For natural language understanding, the subtask which may provide such a metric is word sense disambiguation. One evalutation of this early version of the FACTOTUM SemNet for word sense disambiguation was reported by Gelbukh [42].

A semantic network like FACTOTUM can be useful for both statistical as well as semantic strategies in word sense disambiguation. An example of the statistical use would be to develop a proximity measure based on links in the semantic network to look for associated words in a text and determine which sense is most likely, as explored by Yarowsky [25]. Such a method can be refined by weighting the links according to type, with the link weights determined by training against a corpus with marked word senses. Methodological difficulties in the evaluation of word-sense disambiguation accuracy have been discussed by Ahlswehde and Lorand [26]. Semantically based disambiguation requires testing word combinations (such as adjective/noun or verb/case-filler) to determine whether the semantic attributes of the combining words are of compatible type, as exemplified by the Linguistic String Project's restriction language [27]. In the CYC project [22], this semantic matching function is apparently accomplished by the predicate "MakesSenseFor". Other examples of the statistical use are provided by Morris and Hirst [28], who used a proximity measure based on links in the Roget's Thesaurus to determine topic boundaries in a discourse, and by Voorhees [29] who used WordNet to disambiguate word senses for information retrieval, and found little benefit. To develop an accurate metric of utility for a semantic network from such potential applications will, however, require additional research.

6. Future Work

The semantic relations which have been recognized as used in the Roget are thus far defined only as relatively simple slots. To be useful in inferencing or for natural language understanding, a more detailed definition of each semantic relation in logical format must be developed. Since the final set of semantic relations and modifications was developed gradually in the course of the scanning of the thesaurus, the consistency of the distinctions within each ***main entry*** of those words which are subtypes and those which are otherwise semantically related to the headword needs to be improved. In addition, the referenced words within some semantic relations have not been distinguished as to word sense. The semantic network must also be supplemented with additional words to bring it at least to the level of completeness of the WordNet system. Finding correspondences where possible between the word senses in WordNet and those in the FACTOTUM SemNet is an additional possible future task. Additional automatic supplementation of the FACTOTUM SemNet with new words from unmarked text, might be possible using semantic classification techniques dependent on context, as described by Futrelle and Gauch [31].

7. Conclusions

Examination of the implied semantic relations used to organize the Roget's Thesaurus reveals that the relationships that are considered salient and significant from the point of view of linguistic composition are numerous, nuanced, in some cases complex, and often differ from those used in other systems. Existing ontologies have given much less attention to these nuances within semantic relations than to their hierarchies. Since accurate communication between computer systems will require the use of a standard set of semantic relations to define concepts in a common ontology, considerable additional effort will clearly be needed to develop a consensus on the semantic relations that should be used.

The entire 1911 Roget has been reorganized into a hierarchical semantic network containing approximately 2,000 main entries. This semantic network has more usable structure than the original Roget's Thesaurus, and presents and alternative structure for a semantic network from that of the WordNet. However, considerable additional effort will clearly be needed for the resulting semantic network to reach the level of usability of the WordNet system.

References

[1] M. Ross Quillian, *Word Concepts: A Theory and Simulation of Some Basic Semantic Capabilities*. Behavioral Science 12: 1967, pp. 410-430. Reprinted in R. Brachman and H. Levesque (Eds.) *Readings in Knowledge Representation* (San Mateo, Calif. Morgan Kaufmann, 1985).

[2] S.C. Shapiro and W.J. Rapaport, *The SNePS family*. pp. 243-275 in Fritz Lehmann, Ed. *Semantic Networks in Artificial Intelligence* (New York, Pergamon Press, 1992).

[3] John Sowa, *Conceptual Structures -- Information Processing in Mind and Machine* (Reading, Mass., Addison-Wesley, 1984).

[4] W.A. Woods and J.G. Schmoltze, *The KL-ONE Family*. pp. 133-177 in Fritz Lehmann, Ed. *Semantic Networks in Artificial Intelligence* (New York, Pergamon Press, 1992).

[5] Paul Jacobs and Lisa Rau, *SCISOR: Extracting information from on-line news*. Communications of the Association for Computing Machinery, 33(11): 88-97, November 1990.

[6] Ralph Grishman and Richard Kittredge, Eds. *Analyzing Language in Restricted Domains: sublanguage description and processing*. (Hillsdale, New Jersey, Lawrence Erlbaum Associates, 1986).

[7] Donna Harman. *Overview of the first TREC conference*. pp. 36-47 in *Proceedings of the Sixteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. Robert Korfage, Edie Rasmussen and Peter Willett, Editors. Pittsburgh, PA, July 1993.

These annual evaluations of information retrieval systems focus on some restricted field in order to make the task tractable with present technology.

[8] (a) Thomas R. Gruber. *The Role of Common Ontology in Achieving Sharable, Reusable knowledge Bases*. PP. 601-602 in *Principles of Knowledge Representation and Reasoning: Proceedings of the Second International Conference (KR '91)*. J. Allen, R. Fikes, and E. Sandewall (Eds.) (Morgan Kaufmann, San Francisco and San Mateo, California, 1991).

(b) Thomas R. Gruber, *Toward Principles for the Design of Ontologies Used for Knowledge Sharing*. Technical Report KSL 93-4, January 1993 (Knowledge Systems Laboratory, Computer Science Department, Stanford University).

(c) Thomas R. Gruber, *The Development of Large, Shared Knowledge-Bases: Collaborative Activities at Stanford.* Technical Report KSL-90-62, August 1990 (Knowledge Systems Laboratory, Computer Science Department, Stanford University).

[9] John A. Bateman. *The Theoretical Status of Ontologies in Natural Language Processing*. Technical Report, ISI, March 5, 1992.

[10] John F. Sowa, book review of ref. 16, *in* Artificial Intelligence 61:95-104 (1993).

[11] R. Neches, R. Fikes, T. Finin, T.Gruber, R. Patil, T. Senator, and W. Swartout, Enabling Technology for Knowledge Sharing. AI Magazine 12:36-56 (1991).

[12] C.O. Sylvester Mawson, Litt.D., M.R.A.S., F.S.A., editor, *Roget's Thesaurus of English Words and Phrases Classified and Arranged so as to Facilitate the Expression of Ideas and assist in Literary Composition*, (McDevitt-Wilson's, Inc. New York, 1911). Copyrighted 1911 by Thomas Y. Crowell Company.

The electronic version of this Roget is available by ftp from Project Gutenberg (or on CD-ROM distributed by Walnut Creek CD-ROM).

[13] Marvin Minsky, *The Society of Mind.* (Simon & Schuster, New York, 1985).

[14] George Miller, *WordNet: an on-line lexical database*. International Journal of Lexicography 24: 513-523 (1990). FTP: clarity.princeton.edu.

[15] Representative articles may be found in:

(a) *Relational Models of the Lexicon: Representing knowledge in semantic networks*, Martha W. Evens (Ed.), (Cambridge University Press, 1988); and

(b) Fritz Lehmann, Ed. *Semantic Networks in Artificial Intelligence* (New York, Pergamon Press, 1992).

[16] Kathleen Dahlgren. *Naive Semantics for Natural Language Understanding*. (Boston, Kluwer Academic Publishers, 1988) p. 62.

[17] J.A. Markowitz, J.T. Nutter, and M.W. Evens. *Beyond IS-A and PART-WHOLE: More Semantic Network Links*. pp. 377-390 in Fritz Lehmann, Ed. Semantic Networks in Artificial Intelligence (New York, Pergamon Press, 1992).

[18] Madelyn A. Iris, Bonnie E. Litowitz, and Martha Evens, *Problems of the part-whole relation*. pp.261-188 in *Relational Models of the Lexicon: Representing knowledge in semantic networks*, Martha W. Evens (Ed.), (New York, Cambridge University Press, 1988).

Conferences have already been held on "Formal Mereology" to explore such relations even more deeply.

[19] D.A.B. Lindberg, B.L. Humphries and A.T. McCray, The Unified Medical Language System. Meth. Inform. Med. 32:281-91 (1993).

A complete UMLS bibliography is available by ftp from nlmpubs.nlm.nih.gov in the umls section of the nlmpubs directory. The system is distributed by agreement. (Contact is Betsy L. Humphries at the National Library of Medicine: blh@nlm.nih.gov).

[20] (a) Igor Mel'cuk and Alexander Zholkovsky, *The Explanatory Combinatorial Dictionary*. pp. 41-74 in *Relational Models of the Lexicon: Representing knowledge in semantic networks*, Martha W. Evens (Ed.), (New York, Cambridge University Press, 1988).

(b) James Steele and Ingrid Meyer, *Lexical Functions in an Explanatory Combinatorial Dictionary: Kinds, Descriptions and English Examples*. pp. 41-61 in James Steele, *Meaning-Text Theory* (Ottawa, University of Ottawa Press, 1990).

[21] A list of concept systems (ontologies) from a wide variety of sources appears as an appendix in:

Fritz Lehmann, *CCAT: The Current Status of the Conceptual Catalogue (Ontology) Group, With Proposals*. in *Proceedings, Third PEIRCE Workshop*, held in conjunction with ICCS'94 at the University of Maryland at College Park, August 19, 1994.

[22] D. Lenat and R.V. Guha, *Building Large Knowledge-based Systems -- representation and inference in the CYC project* (Addison-Wesley, Reading, Mass. 1990).

[23] D. Skuce and I. Meyer, *Terminology and knowledge acquisition: a symbiotic relationship*. in: Proceedings of the Sixth Knowledge Acquisition for Knowledge Based Systems Workshop, Banff Alta. (1991).

[24] J.A. Bateman, R.T. Kasper, J.D. Moore, and R.A. Whitney. *A General Organization of Knowledge for Natural Language Processing: The Penman Upper Model*. USC/Information Sciences Institute, Marina Del Rey, CA. Unpublished Research Report, 1990.

[25] David Yarowsky. *Word-sense Disambiguation Using Statistical Models of Roget's Categories Trained on Large Corpora*. pp. 454-460 in: Proceedings of COLING-92 (1992) (available from the Association for Computational Linguistics).

[26] Thomas E. Ahlswehde and David Lorand. *Word Sense Disambiguation by Human Subjects: Computational and Psychological Applications*. pp. 1-9 in: *Acquisition of Lexical Knowledge from Text*. (Proceedings of a Workshop sponsored by ACL/SIGLEX) Branimir Boguraev and James Pustejovsky, Eds. (Association for Computational Linguistics, June 1993).

[27] Naomi Sager, *Natural Language Information Processing*. (Reading, Mass, Addison-Wesley Publishing Co., 1981).

[28] Jane Morris and Graeme Hirst. *Lexical Cohesion Computed by Thesaural Relations as an Indicator of the Structure of Text*. Computational Linguistics 17:21-48 (1991).

[29] E.M. Voorhees. *Using WordNet^{TM} to disambiguate word sense for text retrieval*. Proceedings of ACM SIGIR Conference 16:171-180 (1993).

[30] Elizabeth D. Liddy, Carol A. Hert, and Philip Doty. *Roget's International Thesaurus: Conceptual Issues and Potential Applications*. pp. 93-98 in: *Proceedings of the First ASIS SIG/CR Classification Research Workshop*. Susanne M. Humphrey and Barbara H. Kwasnik, Eds. (Medford, New Jersey, Learned Information, Inc, Nov. 1990).

[31] Robert P. Futrelle and Susan Gauch. *Experiments in Syntactic and Semantic Classification and Disambiguation Using Bootstrapping*. pp. 117-124 in: *Acquisition of Lexical Knowledge from Text*. (Proceedings of a Workshop sponsored by ACL/SIGLEX) Branimir Boguraev and James Pustejovsky, Eds. (Association for Computational Linguistics, June 1993).

[32] Matthew L. Ginsburg. *Knowledge Interchange Format: The KIF of Death*. AI Magazine 12(3): 57-63 (1991).

[33] D.A. Cruse. *Lexical Semantics* (New York., Cambridge University Press, 1986).

[33] R.V. Guha and Douglas B. Lenat. Enabling Agents to Work Together. Comm. of the ACM vol. 37, no. 7, pp. 126-142 (1994).

[34] Toshio Yokoi, *The EDR Electronic Dictionary*. Communications of the ACM vol. 38, no. 11, pp. 42-44 (1995).

[35] Roger Chaffin and Douglas J. Hermann, *The Nature of Semantic Relations: A Comparison of Two Approaches*. pp. 289-334 in ref. 15(a).

[36] Renate Motschnig-Pitrik and Jens Kaasboll, *Part-Whole Relationship Categories and Their Application in Object-Oriented Analysis*, IEEE Trans. on Knowledge and Data Engineering, vol. 11 no. 5, pp. 779-797 (1999).

[37] Piek Vossen and Laura Bloksma, *Categories and Classification in Euro WordNet*, pp. 399-407 in *Proceedings of the First International Conference on Language Resources and Evaluation*, Granada Spain, May 1998.

[38] Kavi Mahesh and Sergei Nirenburg, *A Situated Ontology for Practical NLP*, in proceedings of the *Workshop on Basic Ontological Issues in Knowledge Sharing*, Montreal, Canada, August 1995.

See also http://crl.nmsu.edu/Research/Projects/mikro/index.html

[39] Eduard Hovy, *Combining and Standardizing Large-Scale, Practical Ontologies for Machine Translation and other Uses*. pp. 535-542 in *Proceedings of the First International Conference on Language Resources and Evaluation*, Granada Spain, May 1998.

[40] See: http://www.isi.edu/natural-language/projects/ONTOLOGIES.html

[41] John Sowa, *Conceptual Graphs: draft proposed American National Standard*. pp. 1-65 in William Tepfenhart and Walling Cyre (Eds.) *Conceptual Structures: Standards and Practices* (Springer-Verlag, Berlin and New York, 1999) [This volume is also the Proceedings of the 7th International Conference on Conceptual Structures, held in Blacksburg, Virgina, USA, July 1999. A version is also available: John Sowa, *Conceptual Graphs: draft proposed American National Standard (dpANS) (*NCITS.T2/98-003). at http://concept.cs.uah.edu/CG/cgdpans.html

[42] A.F. Gelbukh. *Using a semantic network for lexical and syntactical disambiguation*. Proc. CIC-97, Nuevas Aplicaciones e Innovaciones Tecnol=gicas en Computacion, Simposium Internacional de Computacion, 12-14, 1997, CIC, IPN, Mexico City, Mexico, pp. 352 - 366. Revised version: A. Gelbukh. *Disambiguation with lexical semantic distance*. In: *Selected Works 1997-1998, Instituto PolitTcnico Nacional, Centro de Investigacion en Computacion*, ISBN 970-18-3427-5, 1999, pp. 150-170. Substantially enlarged version: A. Gelbukh. *Using a semantic network dictionary in some tasks of disambiguation and translation*. Technical report. CIC, IPN, 1998, ISBN 970-18-1894-6.

Patrick Cassidy *is president of MICRA, Inc. at 735 Belvidere Ave., Plainfield, NJ 07062-2054 USA. He can be reached at cassidy@micra.com.*