A SEMANTIC CONCORDANCE

George A. Miller, Claudia Leacock, Randee Tengi, Ross T. Bunker

Cognitive Science Laboratory
Princeton University
Princeton, NJ 08542

ABSTRACT

A semantic concordance is a textual corpus and a lexicon so combined that every substantive word in the text is linked to its appropriate sense in the lexicon. Thus it can be viewed either as a corpus in which words have been tagged syntactically and semantically, or as a lexicon in which example sentences can be found for many definitions. A semantic concordance is being constructed to use in studies of sense resolution in context (semantic disambiguation). The Brown Corpus is the text and WordNet is the lexicon. Semantic tags (pointers to WordNet synsets) are inserted in the text manually using an interface, ConText, that was designed to facilitate the task. Another interface supports searches of the tagged text. Some practical uses for semantic concordances are proposed.

1. INTRODUCTION

We wish to propose a new version of an old idea. Lexicographers have traditionally based their work on a corpus of examples taken from approved usage, but considerations of cost usually limit published dictionaries to lexical entries having only a scattering of phrases to illustrate the usages from which definitions were derived. As a consequence of this economic pressure, most dictionaries are relatively weak in providing contextual information: someone learning English as a second language will find in an English dictionary many alternative meanings for a common word, but little or no help in determining the linguistic contexts in which the word can be used to express those different meanings. Today, however, large computer memories are affordable enough that this limitation can be removed; it would now be feasible to publish a dictionary electronically along with all of the citation sentences on which it was based. The resulting combination would be more than a lexicon and more than a corpus; we propose to call it a *semantic concordance*. If the corpus is some specific text, it is a *specific semantic concordance*; if the corpus includes many different texts, it is a *universal semantic concordance*.

We have begun constructing a universal semantic concordance in conjunction with our work on a lexical database. The result can be viewed either as a collection of passages in which words have been tagged syntactically and semantically, or as a lexicon in which illustrative sentences can be found for many definitions. At the present time, the correlation of a lexical meaning with examples in which a word is used to express that meaning must be done by hand. Manual semantic tagging is tedious; it should be done automatically as soon as it is possible to resolve word senses in context automatically. It is hoped that the manual creation of a semantic concordance will provide an appropriate environment for developing and testing those automatic procedures.

2. WORDNET: A LEXICAL DATABASE

The lexical component of the universal semantic concordance that we are constructing is WordNet, an on-line lexical resource inspired by current psycholinguistic theories of human lexical memory [1, 2]. A standard, handheld dictionary is organized alphabetically; it puts together words that are spelled alike and scatters words with related meanings. Although on-line versions of such standard dictionaries can relieve a user of alphabetical searches, it is clearly inefficient to use a computer merely as a rapid page-turner. WordNet is an example of a more efficient combination of traditional lexicography and modern computer science.

The most ambitious feature of WordNet is the attempt to organize lexical information in terms of word meanings, rather than word forms. WordNet is organized by semantic relations (rather than by semantic components) within the open-class categories of noun, verb, adjective, and adverb; closed-class categories of words (pronouns, prepositions, conjunctions, etc.) are not included in WordNet. The semantic relations among open-class words include: synonymy and antonymy (which are semantic relations between words and which are found in all four syntactic categories); hyponymy and hypernymy (which are semantic relations between concepts and which organize nouns into a categorical hierarchy); meronymy and holonymy (which represent part-whole relations among noun concepts); and troponymy (manner relations) and entailment relations between verb concepts. These semantic relations were chosen to be intuitively obvious to nonlinguists and to have broad applicability throughout the lexicon.

The basic elements of WordNet are sets of synonyms (or synsets), which are taken to represent lexicalized concepts. A synset is a group of words that are synonymous, in the sense that there are contexts in which they can be interchanged without changing the meaning of the statement. For example, WordNet distinguishes between the synsets:

{board, plank, (a stout length of sawn timber)}

{board, committee, (a group with supervisory powers)}

In the context, "He nailed a board across the entrance," the word "plank" can be substituted for "board." In the context, "The board announced last quarter's dividend," the word "committee" can be substituted for "board."

WordNet also provides sentence frames for each sense of every verb, indicating the kinds of simple constructions into which the verb can enter.

WordNet contains only uninflected (or base) forms of words, so the interface to WordNet includes `morphy`, a morphological analyzer that is applied to input strings to generate the base forms. For example, given "went" as the input string, `morphy` returns "go"; given "children," it returns "child," etc. `morphy` first checks an exception list; if the input string is not found, it then uses standard rules of detachment.

Words (like "fountain pen") that are composed of two or more simpler words with spaces between them are called collocations. Since collocations are less polysemous than are individual words, their inclusion in WordNet promises to simplify the task of sense resolution. However, the morphology of collocations poses certain problems. Special algorithms are required for inflected forms of some collocations: for example, "standing astride of" will return the phrasal verb, "stand astride of."

As of the time this is written, WordNet contains more than 83,800 entries (unique character strings, words and collocations) and more than 63,300 lexicalized concepts (synsets, plus defining glosses); altogether there are more than 118,600 entry-concept pairs. The semantic relations are represented by more than 87,600 pointers between concepts. Approximately 43% of the entries are collocations. Approximately 63% of the synsets include definitional glosses. And approximately 14% of the nouns and 25% of the verbs are polysemous.

WordNet continues to grow at a rate of almost 1,000 concepts a month. The task of semantic tagging has provided a useful stimulus to improve both coverage and precision.

3. THE BROWN CORPUS

The textual component of our universal semantic concordance is taken from the Brown Corpus [3, 4]. The corpus was assembled at Brown University in 1963-64 under the direction of W. Nelson Francis with the intent of making it broadly representative of American English writing. It contains 500 samples, each approximately 2,000 words long, for a total of approximately 1,014,000 running words of text, where a "word" is defined graphically as a string of contiguous alphanumeric characters with a space at either end. The genres of writing range from newspaper reporting to technical writing, and from fiction to philosophical essays.

The computer-readable form of the Brown Corpus has been used in a wide variety of research studies, and many laboratories have obtained permission to use it. It was initially used for studies of word frequencies, and subsequently was made available with syntactic tags for each word. Since it is well known in a variety of contexts, and widely available, the Brown Corpus seemed a good place to begin.

4. SEMANTIC TAGGING

Two contrasting strategies for connecting a lexicon and a corpus emerge depending on where the process starts. The targeted approach starts with the lexicon: target a polysemous word, extract all sentences from the corpus in which that word occurs, categorize the instances and write definitions for each sense, and create a pointer between each instance of the word and its appropriate sense in the lexicon; then target another word and repeat the process. The targeted approach has the advantage that concentrating on a single word should produce better definitions—it is, after all, the procedure that lexicographers regard as ideal. And it also makes immediately available a classification of sentences that can be used to test alternative methods of automatic sense resolution.

The alternative strategy starts with the corpus and proceeds through it word by word: the sequential approach. This procedure has the advantage of immediately revealing deficiencies in the lexicon: not only missing words (which could be found more directly), but also missing senses and indistinguishable definitions—deficiencies that would not surface so quickly with the targeted approach. Since the promise of improvements in WordNet was a major motive for pursuing this research, we initially adopted the sequential approach for the bulk of our semantic tagging.

A second advantage of the sequential approach emerged as the work proceeded. One objective test of the adequacy of a lexicon is to use it to tag a sample of text, and to record the number of times it fails to have a word, or fails to have the appropriate sense for a word. We have found that such records for WordNet show considerable variability depending on the particular passage that is tagged, but over several months the averaged estimates of its coverage have been slowly improving: coverage it is currently averaging a little better than 96%.

5. CONTEXT: A TAGGING INTERFACE

The task of semantically tagging a text by hand is notoriously tedious, but the tedium can be reduced with an appropriate user interface. ConText is an X-windows interface designed specifically for annotating written texts with WordNet sense tags [5]. Since WordNet contains only open-class words, ConText is used to tag only nouns, verbs, adjectives, and adverbs; that is to say, only about 50% of the running words in the Brown Corpus are semantically tagged.

Manual tagging with ConText requires a user to examine each word of the text in its context of use and to decide which WordNet sense was intended. In order to facilitate this task, ConText displays the word to be tagged in its context, along with the WordNet synsets for all of the senses of that word (in the appropriate part of speech). For example, when the person doing the tagging reaches "horse" in the sentence:

The **horse** and men were saved, but the oxen drowned.

ConText displays WordNet synsets for five meanings of noun "horse":

1. sawhorse, horse, sawbuck, buck (a framework used by carpenters)
2. knight, horse (a chess piece)
3. horse (a gymnastic apparatus)
4. heroin, diacetyl morphine, H, horse, junk, scag, smack (a morphine derivative)
5. horse, Equus caballus (herbivorous quadruped)

The tagger uses the cursor to indicate the appropriate sense (5, in this example), at which point ConText attaches a label, or semantic tag, to that word in the text. ConText then moves on to "men," the next content word, and the process repeats. If the word is missing, or if the appropriate sense is missing, the tagger can insert comments calling for the necessary revisions of WordNet.

5.1. Input to ConText

In the current version of ConText, text to be tagged semantically must be preprocessed to indicate collocations and proper nouns (by concatenating them with underscores) and to provide syntactic tags. Since different corpora come in different formats and so require slightly different preprocessing, we have not tried to incorporate the preprocessor into ConText itself.

A tokenizer searches the input text for collocations that WordNet knows about and when one is found it is made into a unit by connecting its parts with underscores. For example, if a text contains the collocation "took place," the tokenizer will convert it to "took_place." ConText can then display the synset for "take place" rather than successive synsets for "take" and "place."

Syntactic tags indicate the part of speech of each word in the input text. We have used an automatic syntactic tagger developed by Eric Brill [6] which he generously adapted to our needs. For example, "store" can be a noun or a verb; when the syntactic tagger encounters an instance of "store" it tries to decide from the context whether it is being used as a noun or a verb. ConText then uses this syntactic tag to determine which part of speech to display to the user. ConText also uses syntactic tags in order to skip over closed-class words. Since the automatic syntactic tagger sometimes makes mistakes, ConText allows the user to change the part of speech that is being displayed, or to tag words that should not have been skipped.

After the text has been syntactically tagged, all contiguous strings of proper nouns are joined with an underscore. For example, the string "Mr. Charles C. Carpenter" is output as "Mr._Charles_C._Carpenter." Here, too, the user can manually correct any mistaken concatenations.

An example may clarify what is involved in preprocessing. The 109th sentence in passage k13 of the Brown Corpus is:

> He went down the hall to Eugene's bathroom, to turn on the hot-water heater, and on the side of the tub he saw a pair of blue wool swimming trunks.

After preprocessing, this sentence is passed to ConText in the following form:

> br-k13:109: He/PP went_down/VB the/DT hall/NN to/TO Eugene/NP '/POS s/NN bathroom/NN ,/, to/TO turn_on/VB the/DT hot-water/NN heater/NN ,/, and/CC on/IN the/DT side/NN of/IN the/DT tub/NN he/PP saw/VBD a/DT pair/NN of/IN blue/JJ wool/NN swimming_trunks/NN ./.

The version displayed to the tagger, however, looks like the Brown Corpus, except that collocations are indicated by underscores. Note, incidentally, that the processor has made a mistake in this example: "went_down" (as in "the ship went down") is not the sense intended here.

5.2. Output of ConText

The output of ConText is a file containing the original text annotated with WordNet semantic tags; semantic tags are given in square brackets, and denote the particular WordNet synset that is appropriate. For example, when "hall" is tagged with [noun.artifact.1] it means that the word is being used to express the concept defined by the synset containing "hall1" in the noun.artifact file. (Since WordNet is constantly growing and changing, references to the lexicographers' files have been retained; if the lexical component were frozen, some more general identifier could be used instead.) In cases where the appropriate sense of a word is not in WordNet, the user annotates that word with a comment that is later sent to the appropriate lexicographer. After the lexicographer has edited WordNet, the text must be retagged. In the retag mode, ConText skips from one commented word to the next.

In addition to the syntactic and semantic tags, ConText adds SGML markers and reformats the text one word to a line. The SGML markers delimit sentences <s>, sentence numbers <stn>, words in the text <wd>, base forms of text words <mwd>, comments <cmt>, proper nouns <pn>, part-of-speech tags <tag> and semantic tags <sn> or <msn>. The sentence preprocessed above might come out of ConText looking like this:

<stn>109</stn>

```
<wd>He</wd><tag>PP</tag>
<wd>went</wd><mwd>go</mwd><msn>[verb.motion.6]
  </msn><tag>VB</tag>
<wd>down</wd>
<wd>the</wd><tag>DT</tag>
<wd>hall</wd><sn>[noun.artifact.1]</sn><tag>NN</tag>
<wd>to</wd><tag>TO</tag>
<wd>Eugene</wd><pn>person</pn><sn>[noun.Tops.0]
  </sn><tag>NP</tag>
<wd>'</wd><tag>POS</tag>
<wd>s</wd><tag>NN</tag>
<wd>bathroom</wd><sn>[noun.artifact.0]</sn>
  <tag>NN</tag>
<wd>,</wd><tag>,</tag>
<wd>to</wd><tag>TO</tag>
<wd>turn_on</wd><sn>[verb.contact.0]</sn>
  <tag>VB</tag>
<wd>the</wd><tag>DT</tag>
<wd>hot-water_heater</wd><cmt>WORD_MISSING
  </cmt><tag>NN</tag>
<wd>,</wd><tag>,</tag>
<wd>and</wd><tag>CC</tag>
<wd>on</wd><tag>IN</tag>
<wd>the</wd><tag>DT</tag>
<wd>side</wd><sn>[noun.location.0]</sn><tag>NN</tag>
<wd>of</wd><tag>IN</tag>
<wd>the</wd><tag>DT</tag>
<wd>tub</wd><sn>[noun.artifact.1]</sn><tag>NN</tag>
<wd>he</wd><tag>PP</tag>
<wd>saw</wd><mwd>see</mwd><msn>[verb.perception.0]
  </msn><tag>VBD</tag>
<wd>a</wd><tag>DT</tag>
<wd>pair</wd><sn>[noun.quantity.0]</sn><tag>NN</tag>
<wd>of</wd><tag>IN</tag>
<wd>blue</wd><sn>[adj.all.0.col.3]</sn><tag>JJ</tag>
<wd>wool</wd><sn>[noun.artifact.0]</sn><tag>NN</tag>
<wd>swimming_trunks</wd><sn>[noun.artifact.0]</sn>
  <tag>NN</tag>
<wd>.</wd><tag>.</tag>
</s>
```

Note that the tokenizer's mistaken linking of "went_down" has now been corrected by the tagger. Also note "<cmt>WORD_MISSING</cmt>" on line 16 of the output: that comment indicates that the tagger has connected "hot-water" and "heater" to form the collocation "hot-water_heater," which was not in WordNet. This illustrates the kind of comments that are passed on to the lexicographers, who use them to edit or add to WordNet.

The WordNet database is constantly growing and changing. Consequently, previously tagged texts must be updated periodically. In the update mode, ConText searches the tagged files for pointers to WordNet senses that have subsequently been revised. A new semantic tag must then be inserted by the tagger.

5.3 Tracking

As the number of semantically tagged files increased, the difficulty of keeping track of which files had been preprocessed, which had been tagged, which were ready to be retagged, which had been retagged, and which were complete and cleared for use made it necessary to create a master tracking system that would handle the record keeping automatically. Scripts were written that allowed an administrator to preprocess files and add them to the tracking system. Once files are in the tracking system, other scripts keep a log of all the tagging activities pertaining to each file, and insure that taggers will not try to perform operations that are invalid for files with a given status. The administrator can easily generate simple reports on the status of all files in the tracking system.

6. QUERYING THE TAGGED TEXT

A program to query the semantically tagged database has also been written: `prsent` (print sentences) allows a user to retrieve sentences by entering the base form of a word and its semantic tag. It was developed as a simple interface to the semantic concordance, and puts the burden of knowing the word's semantic tag on the user. This program is useful to the lexicographers, who are intimately familiar with WordNet semantic tags and who use it to find sample sentences. A more robust interface is needed, however.

Presently under development is a comprehensive querying tool that will allow a user the flexibility of specifying various retrieval criteria and display options. Envisioned is an X-Windows application with two main windows: one area for entering searching information and another for displaying the retrieved sentences. A primary search key is the only required component. Additional search keys can be specified to find words that co-occur in sentences. This alone is a powerful improvement over `prsent`. Other options will restrict or expand the retrieval, as listed here:

1. Search only given part(s) of speech.
2. Search only for a specific sense.
3. Expand search to include sentences for synonyms of search key.
4. Expand search to include sentences for hyponyms of search key.
5. Use primary key and all secondary keys, or primary key and any secondary key.
6. Search for a secondary key that is within *n* words of the primary key.

As important as specifying searching criteria is how the retrieved information is displayed. An option will be provided to display retrieved sentences in a concordance format (all the target words vertically aligned and surrounded by context to the window's borders) or left justified. Search keys will be highlighted in the retrieved sentences.

Implementation of this program requires the creation of a "master list" of semantically tagged words. Each line in the alphabetized list contains the target word, its semantic tag, and for each sentence containing the word, a list of all the co-occurring nouns, verbs, adjectives, and adverbs with numbers indicating their position in the sentence. For example, the sentence already dissected provides a context for "hall" that might look like this:

hall/5 [noun.artifact.1]:

{bathroom/10 [noun.artifact.0]; hot-water_heater/15 [noun.artifact.0]; side/19 [noun.location.0]; tub/22 [noun.artifact.1]; pair/25 [noun.quantity.0]; wool/28 [noun.artifact.0]; swimming_trunks/29 [noun.artifact.0]}

{go/2 [verb.motion.6]; turn_on/13 [verb.contact.0]; see/23 [verb.perception.0]}

{blue/27 [adj.all.col.3]}

{ }

Collecting entries for this sense of "hall" provides valuable information about the contexts in which it can occur.

7. APPLICATIONS

Our reasons for building this universal semantic concordance were to test and improve the coverage of WordNet and to develop resources for developing and testing procedures for the automatic sense resolution in context. It should be pointed out, however, that semantic concordances can have other uses.

7.1. Instruction

Dictionaries are said to have evolved from the interlinear notations that medieval scholars added for difficult Latin words [7]. Such notations were found to be useful in teaching students; as the number of such notations grew, collections of them were extracted and arranged in lists. When the lists took on a life of their own their educational origins were largely forgotten. A semantic concordance brings this story back to its origins: lexical "footnotes" indicating the meaning that is appropriate to the context are immediately available electronically.

One obvious educational use of a semantic concordance would be for people trying to learn English as a second language. By providing them with the appropriate sense of an unfamiliar word, they are spared the task of selecting a sense from the several alternatives listed in a standard dictionary. Moreover, they can retrieve other sentences that illustrate the same usage of the word, and from such sentences they can acquire both local and topical information about the use of a word: (1) local information about the grammatical constructions in which that word can express the given concept, and (2) topical information about other words that are likely to be used when that concept is discussed.

A use for specific semantic concordances would be in science education: much of the new learning demanded of beginning students in any field of science is terminological.

7.2. Sense Frequencies

Much attention has been paid to word frequencies, but relatively little to the frequencies of occurrence of different meanings. Some lexicographers have atempted to order the senses of polysemous words from the most to the least frequent, but the more general question has not been asked because the data for answering it have not been available. We have enough tagged text now, however, to get an idea what such data would look like. For example, here are preliminary data for the 10 most frequent concepts expressed by nouns, based on some 80 selections from the Brown Corpus:

172 {year, (time_period)}
144 {person, individual, someone, man, mortal, human, soul, (a human being)}
139 {man, adult_male, (a grown man)}
105 {consequence, effect, outcome, result, upshot, (a phenomenon that follows and is caused by some previous phenomenon)}
104 {night, night_time, dark, (time after sunset and before sunrise while it is dark outside)}
102 {kind, sort, type, form, ("sculpture is a form of art" or "what kind of man is this?")}
94 {eye, eyeball, oculus, optic, peeper, (organ of sight)}
89 {day, daytime, daylight, (time after sunrise and before sunset while it is light outside)}
88 {set, class, category, type, family, (a collection of things sharing a common attribute)}
87 {number, count, complement, (a definite quantity)}

Our limited experience suggests, however, that such statistics depend critically on the subject matter of the corpus that is used.

7.4. Sense Co-occurrences

One shortcoming of WordNet that several users have pointed out to us is its lack of topical organization. Peter Mark Roget's original conception of his thesaurus relied heavily on his list of topics, which enabled him to pull together in one place all of the words used to talk about a given topic. This tradition of topical organization has survived in many modern thesauri, even though it requires a double look-up by the reader. For example, under "baseball" a topically organized thesaurus would pull together words like "batter," "team," "lineup," "diamond," "homer," "hit," and so on. Topical organization obviously facilitates sense resolution: if the topic is baseball, the meaning of "ball" will differ from its meaning when the topic is, say, dancing. In WordNet, those same words are scattered about: a baseball is an artifact, batters are people, a team is a group, a lineup is a list, a diamond is a location, a homer is

an act, to hit is a verb, and so on. By itself, WordNet does not provide topical groupings of words that can be used for sense resolution.

One solution would be to draw up a list of topics and index all of the WordNet synsets to the topics in which they are likely to occur. Chapman [8], for example, uses 1,073 such classes and categories. But such lists are necessarily arbitrary. A universal semantic concordance should be able to accomplish the same result in a more natural way. That is to say, a passage discussing baseball would use words together in their baseball senses; a passage discussing the drug trade would use words together with senses appropriate to that topic, and so on. Instead of a long list of topics, the corpus should include a large variety of passages.

In order to take advantage of this aspect of universal semantic concordances, it is necessary to be able to query the textual component for associated concepts. Data on sense co-occurrences build up slowly, of course, but they will be a valuable by-product of this line of work.

7.4. Testing

We are developing a version of the ConText interface that can be used for psychometric testing. The tagger's task in using ConText resembles an extended multiple-choice examination, and we believe that that feature can be adapted to test reading comprehension. Given a text that has already been tagged, readers' comprehension can be tested by seeing whether they are able to choose correct senses on the basis of the contexts of use.

No doubt there are other, even better uses for semantic concordances. As the variety of potential applications grows, however, the need to automate the process of semantic tagging will become ever more pressing. But we must begin with what we have. We are now finishing a first installment of semantically tagged text consisting of 100 passages from the Brown Corpus; as soon as that much has been completed and satisfactorily cleaned up, we plan to make it, and the corresponding WordNet database, available to other laboratories that also have permission to use the Brown Corpus. We expect that such distribution will stimulate further uses for semantic concordances, uses that we have not yet imagined.

8. CONCLUSION

The fact that we have control of the lexical component of our semantic concordance enables us to shape the lexicon to fit the corpus. It would be possible, of course, to create a specific semantic concordance with a lexicon limited strictly to the words occurring in the accompanying corpus. That constraint would have certain size advantages, but would miss the opportunity to build a single general lexicon onto which a wide variety of corpora could be mapped.

The universal semantic concordance described here has enabled us to improve WordNet and has given us a tool for our studies of sense resolution in context. In the course of this exercise, however, it has become apparent to us that cross-referencing a lexicon and a textual corpus produces a hybrid resource that will be useful in a variety of practical and scientific applications. It has occurred to us that semantic concordances might be even more useful if a richer syntactic component could be incorporated, but how best to accomplish that is presently a question for the future.

ACKNOWLEDGMENTS

This work has been supported in part by Grant No. N00014-91-J-1634 from the Defense Advanced Research Projects Agency, Information and Technology Office, and the Office of Naval Research, and in part by grants from the James S. McDonnell Foundation and from the Pew Charitable Trusts. We are indebted to Henry Kučera and W. Nelson Francis for permission to use the Brown Corpus in our research. And we are indebted for assistance and advice to Anthony Adler, Christiane Fellbaum, Kathy Garuba, Dawn Golding, Brian Gustafson, Benjamin Johnson-Laird, Philip N. Johnson-Laird, Shari Landes, Elyse Michaels, Katherine Miller, Jeff Tokazewski, and Pamela Wakefield. The designation, "semantic concordance," was suggested to us by Susan Chipman.

REFERENCES

1. Miller, G. A. (ed.), WordNet: An on-line lexical database. *International Journal of Lexicography* (special issue), 3(4):235-312, 1990.
2. Miller, G. A. and Fellbaum, C. Semantic networks of English. *Cognition* (special issue), 41(1-3):197-229, 1991.
3. Kučera, H. and Francis, W. N. *Computational analysis of present-day American English.* Providence, RI: Brown University Press, 1967.
4. Francis, W. N. and Kučera, H. *Frequency analysis of English Usage: Lexicon and Grammar.* Boston, MA: Houghton Mifflin, 1982.
5. Leacock, C. ConText: A tool for semantic tagging of text: User's guide. Cognitive Science Laboratory, Princeton University: CSL Report No. 54, February 1993.
6. Brill, E. A simple rule-based part of speech tagger. In *Proceedings of Speech and Natural Language Workshop*, 112-116, February 1992. San Mateo, CA: Morgan Kaufman.
7. Landauer, S. I. *Dictionaries: The art and craft of lexicography.* New York: Scribner's, 1984.
8. Chapman, R. L. (ed.) *Roget's International Thesaurus*, (5th edition). New York: HarperCollins, 1992.