SEMANTICALLY SIGNIFICANT PATTERNS IN DICTIONARY DEFINITIONS *

Judith Markowitz
Computer Science Department
De Paul University, Chicago, IL 60604
Thomas Ahlswede
Martha Evens
Computer Science Department
Illinois Institute of Technology, Chicago, Il 60616

ABSTRACT

Natural language processing systems need large lexicons containing explicit information about lexical-semantic relationships, selection restrictions, and verb categories. Because the labor involved in constructing such lexicons by hand is overwhelming, we have been trying to construct lexical entries automatically from information available in the machine-readable version of Webster's Seventh Collegiate Dictionary. This work is rich in implicit information; the problem is to make it explicit. This paper describes methods for finding taxonomy and set-membership relationships, recognizing nouns that ordinarily represent human beings, and identifying active and stative verbs and adjectives.

INTRODUCTION

Large natural language processing systems need lexicons much larger than those available today with explicit information about lexical-semantic relationships, about usage, about forms, about morphology, about case frames and selection restrictions and other kinds of collocational information. Apresyan, Mel'cuk, and Zholkovsky studied the kind of explicit lexical information needed by non-native speakers of a language. Their Explanatory-Combinatory Dictionary (1970) explains how each word is used and how it combines with others in phrases and sentences. Their dream has now been realized in a full-scale dictionary of Russian (Mel'cuk and Zholkovsky, 1985) and in example entries for French (Mel'cuk et al., 1984). Computer programs need still more explicit and detailed information. We have discussed elsewhere the kind of lexical information needed in a question answering system (Evens and Smith, 1978) and by a system to generate medical case reports (Li et al., 1985).

This research was supported by the National Science Foundation under IST-85-10069.

A number of experiments have shown that relational thesauri can significantly improve the effectiveness of an information retrieval system (Fox, 1980; Evens et al., 1985; Wang et al., 1985). A relational thesaurus is used to add further terms to the query, terms that are related to the original by lexical relations like synonymy, taxonomy, set-membership, or the part-whole relation, among others. The addition of these related terms enables the system to identify more relevant documents. The development of such relational thesauri would be comparatively simple if we had a large lexicon containing relational information. (A comparative study of lexical relations can be found in Evens et al., 1980).

The work involved in developing a lexicon for a large subset of English is so overwhelming, that it seems appropriate to try to build a lexicon automatically by analyzing information in a machine-readable dictionary. A collegiate level dictionary contains an enormous amount of information about thousands of words in the natural language it describes. This information is presented in a form intended to be easily understood and used by a human being with at least some command of the language. Unfortunately, even when the dictionary has been transcribed into machine-readable form, the knowledge which a human user can acquire from the dictionary is not readily available to the computer.

There have been a number of efforts to extract information from machine-readable dictionaries. Amsler (1980, 1981, 1982) and Amsler and John White (1979) mapped out the taxonomic hierarchies of nouns and verbs in the Merriam-Webster Pocket Dictionary. Michiels (1981, 1983) analyzed the Longman Dictionary of Contemporary English (LDOCE), taking advantage of the fact that that dictionary was designed to some extent to facilitate computer manipulation. Smith (1981) studied the

"defining formulae" - significant recurring phrases - in a selection of adjective definitions from Webster's Seventh Collegiate Dictionary (W7). Carolyn White (1983) has developed a program to create entries for Sager's Linguistic String Parser (1981) from W7. Chodorow and Byrd (1985) have extracted taxonomic hierarchies, associated with feature information, from LDOCE and W7.

We have parsed W7 adjective definitions (Ahlswede, 1985b) using Sager's Linguistic String Parser (Sager, 1981) in order to automatically identify lexical-semantic relations associated with defining formulae. We have also (Ahlswede and Evens, 1983) identified defining formulae in noun, verb and adverb definitions from W7. At present we are working on three interrelated projects: identification and analysis of lexical-semantic relations in or out of W7; generation of computed definitions for words which are used or referred to but not defined in W7; and parsing of the entire dictionary (or as much of it as possible) to generate from it a large general lexical knowledge base.

This paper represents a continuation of our work on defining formulae in dictionary definitions, in particular definitions from W7. The patterns we deal with are limited to recurring phrases, such as "any of a" or "a quality or state of" (common in noun definitions) and "of or relating to" (common in adjective definitions). From such phrases, we gain information not only about the words being defined but also about the words used in the definitions and other words in the lexicon. Specifically, we can extract selectional information, co-occurrence relations, and lexical-semantic relations. These methods of extracting information from W7 were designed for use in the lexicon builder described earlier by Ahlswede (1985a).

The computational steps involved in this study were relatively simple. First W7 definitions were divided by part of speech into separate files for nouns, verbs, adjectives, and others. Then a separate Keyword In Context (KWIC) Index was made for each part of speech. Hypotheses were tried out initially on a subset of the dictionary containing only those words which appeared eight or more times in the Kucera and Francis corpus (1968) of a million words of running English text. Those that proved valid for this subset were then tested on the full dictionary. This work would have been impossible without the kind permission of the G. & C. Merriam Company to use the machine-readable version of W7 (Olney et al. 1967).

NOUN TAXONOMY

Noun definitions which begin with "Any" signal a taxonomic relationship between the noun being defined and a taxonomic superordinate which follows the word "Any." One subset of the formulae beginning with "Any" has the form: "Any"-NP, where the NP can be a noun, noun phrase, or a co-ordinated noun or adjective structure.

1a. **alkyl** any univalent aliphatic, aromatic-aliphatic, or alicyclic hydrocarbon radical.

b. **ammunition** any material used in attack or defense.

c. **streptococcus** any coccus in chains

d. **nectar** any delicious drink

e. **discord** any harsh or unpleasant sound

f. **milkwort** any herb of a genus (Polygala) of the family Polygalaceae, the milkwort family

In these definitions the taxonomic superordinate of the noun being defined is the head noun of the NP immediately following "Any". The superordinate of "alkyl" is "radical," which is the head of the co-ordinated structure following "Any" whereas the superordinate of "ammunition" is the unmodified noun "material." Of the 97 examples of "Any"-NP only two failed to contain an overt taxonomic superordinate following "Any."

2a. **week** any seven consecutive days

b. **couple** any two persons paired together

In each of these cases there is an implicit taxonomic superordinate "set."

The second frequently occurring subset of noun definitions containing "Any" begins with the following pattern: "Any of"-NP. This pattern has two principal realizations depending upon what immediately follows "Any of." In one sub-pattern a quantifier, numeric expression, or "the" follows the initial "Any of" and begins an NP which contains the superordinate of the noun being defined. This pattern is similar to that described above for the "Any"-NP formula.

3a. **doctor** any of several brightly colored artificial flies

b. **allomorph** any of two or more distinct crystalline forms of the same substance.

c. **elder** any of various church officers

The other sub-pattern expresses a biological taxonomic relationship and has the following definition structure:

"Any of a/an"
<optional> modifier
taxonomic level
"("scientific name")"
"of" taxonomic superordinate
either attributes or taxonomic subordinate

The modifier is optional and modifies the taxonomic level of the noun being defined; the capitalized scientific name of the level follows in parenthesis; the taxonomic superordinate can be a noun or a complex NP and is the object of the second "of" in the formula; and the information following the superordinate is generally a co-ordinated structure, frequently co-ordinated NPs. Of the 901 instances of the definition-initial "Any of a/an" sequence 853, or 95 per cent, were biological definitions.

4a. **ant** any of a family (Formicidae) of colonial hymenopterous insects with complex social organization and various castes performing special duties.

b. **grass** any of a large family (Gramineae) of monocotyledonous mostly herbaceous plants with jointed stems, slender sheathing leaves, and flowers borne in spikelets of bracts.

c. **acarid** any of an order (Acarina) of arachnids including mites and ticks.

d. **cercis** any of a small genus (Cercis) of leguminous shrubs or low trees.

e. **nematode** any of a class or phylum (Nematoda) of elongated cylindrical worms parasitic in animals or plants or free-living in soil or water.

f. **archaeornis** any of a genus (Archaeornis) of upper Jurassic toothed birds.

The only sequences which break from the pattern described above are non-biological definitions, which do not have parenthetical information following the head noun of the NP following "Any of a/an" and biological definitions where that head noun is "breed."

5a. **globulin** any of a class of simple proteins (as myosin) insoluble in pure water but soluble in dilute salt solutions that occur widely in plant and animal tissues.

b. **rottweiler** any of a breed of tall vigorous black short-haired cattle dogs.

c. **poland china** any of an American breed of large white-marked black swine of the lard type.

The definition for "globulin" illustrates that even when a non-biological definition has a parenthesis, that parenthetical information does not immediately follow the NP following "Any of a/an." The other definitions in (5) are instances of "breed" following "Any of a/an." In general, when a definition begins with "Any of a/an" it is almost certainly a biological definition and that certainty is increased if the "Any of a/an noun" is immediately followed by parenthesis unless the noun of the pattern is "breed."

THE MEMBER-SET RELATION

Another defining formula with an interesting resemblance to taxonomy also occurs in noun definitions. The pattern "A member of"-NP is similar to the basic organization of the "Any" definitions in that the immediate superordinate of the noun being defined is the object of the preposition "of" except in this pattern the relationship is, of course, member-set.

6a. **hand** a member of a ship`s crew.

b. **earl** a member of the third grade of the British peerage ranking below a marquess and above a viscount.

c. **Frank** a member of a West Germanic people entering the Roman provinces in A.D. 253, occupying the Netherlands and most of Gaul, and establishing themselves along the Rhine.

d. **republican** a member of a political

party advocating republicanism

e. **Fox** a member of an Indian people formerly living in Wisconsin.

f. **Episcopalian** a member of an episcopal church (as the Protestant Episcopal Church).

g. **friar** a member of a mendicant order

What we have here is a generic term for any member of the specified set. It is perhaps best thought of as similar to the part-whole relation -- a hand is part of a crew, a Frank is part of a tribe, an earl is (somewhat inelegantly) part of a peerage.

In our data the nouns being defined with this formula are invariably human. Of the 581 definitions which begin with "A member of" only nine define non-human nouns and two of those are anthropomorphic.

7a. **Jotunn** a member of a race of giants in Norse mythology

b. **Houyhnhnm** a member of a race of horses endowed with reason in Swift's Gulliver's Travels.

Why is it important to mark nouns in a lexicon as explicitly human? Many verbs can take only human subjects or objects. Also, the choice between the relative pronouns who and which depends on whether the referent is human or not.

The member-set relation needs to be distinguished from another relation that classifies a specific individual as in

8a. **Circe** sorceress who changed Odysseus' men into swine.

GENERIC AGENTS

Generic agents are the typical fillers of the agent argument sot for a given verb. They are particularly valuable in understanding intersentential references or generating them. One very surprising source of definitions for human nouns is the formula "One that." Of the 1419 examples of this pattern 694, or 49 per cent were verifiably human. That is, it was possible to determine from the definition itself or from associated definitions, such as a related verb, that the noun being defined was +human. This estimate is, therefore, conservative. It was also determined that a large portion of these definitions (30 per cent) were of occupations.

9a. **goldbeater** one that beats gold into gold leaf

b. **pollster** one that conducts a poll or compiles data obtained by a poll.

c. **schoolmaster** one that disciplines or directs.

d. **hatter** one that makes, sells, or cleans and repairs hats.

e. **assassin** one that murders either for hire or for fanatical motives.

f. **taxpayer** one that pays or is liable to pay a tax

g. **teletypist** one that operates a teletypewriter.

WHAT THE PARENTHESES TELL US

The formula "one (..)" offers very different information. (This formula typically occurs somewhere in the middle of a definition, not at the beginning.) If the first word of the parenthetical information is not "as", a definition which begins with this pattern is a biological definition. The parenthetical material is the scientific name of the noun being defined. These definitions are sub-definitions and almost invariably follow "esp: ".

10a. **pimpernel** any of a genus (Anagallis) of herbs of the primrose family; esp: one (A. arvensis) whose scarlet, white, or purplish flowers close at the approach of rainy or cloudy weather.

b. **whelk** any of numerous large marine snails (as of the genus Buccinum); esp: one (B. undatum) much used as food in Europe.

c. **turnip** either of two biennial herbs of the mustard family with thick roots eaten as a vegetable or fed to stock, one (Brassica rapa) with hairy leaves and usu. flattened roots.

d. **capuchin** any of a genus (Cebus) of So. American monkeys; esp one (C. capucinas) with the hair on its crown resembling a monk's cowl.

e. **croton** any of a genus (Croton) of

herbs and shrubs of the spurge famil, one (C. eluteria) of the Bahamas yielding cascarilla bark.

f. **bully tree** any of several tropical American trees of the Sapodillo family; esp one (Manilkara bidentata) that yields balata gum and heavy red timber.

SUFFIX DEFINITIONS

The defining pattern "One...(... specific /such...)" is an interesting sequence which is only used to define suffixes. The words "specific" and "such" signal this while at the same time indicating what semantic information should be taken from the stem to which the suffix is affixed.

11a. **-er** one that is a suitable object of (a specified action).

b. **-ate** one acted upon (in a specified way).

c. **-morph** one having (such) a form.

d. **-path** one suffering from (such) an ailment.

e. **-ant** one that performs (a specified action).

f. **-grapher** one that writes about (specified) material or in a (specified) way.

Examples associated with some of the definitions in (10) are "isomorph," "psychopath," and "violinist." We are in the process of analyzing all instances of parenthetical "specified" and "such" to determine whether the defining formula exemplified by (10) is a general approach to the definition of affixes. Clearly, the use of parentheses is very significant, signalling an important semantic distinction.

WHAT NOUN DEFINITIONS TELL US ABOUT VERBS

Noun defining patterns can provide important information about specific verbs. Not surprisingly, one of these is the pattern "Act of Ving" which is an indicator of action verbs.

Action verbs differ from stative verbs in a number of important ways. Action verbs like bite and persuade can appear in imperative sentences, while stative verbs like own and resemble cannot:

Bite that man!
Persuade him to go!
*Own the house!
*Resemble your father!

Action verbs take the progressive aspect; stative verbs do not:

She is biting the man.
She is persuading him to go.
*She is owning the house.
*She is resembling your father.

Action verbs can appear in a number of embedded sentences where statives cannot be used.

I told her to bite the man.
*I told her to own the house.

In definitions the action verb appears as the gerundive object of the preposition "of" or as the present-tense verb of the subordinate clause.

12a. **plumbing** the act of using a plumb.

b. **forgiveness** the act of forgiving.

c. **soliloquy** the act of talking to oneself.

d. **projection** the act of throwing or shooting forward.

e. **refund** the act of refunding

f. **protrusion** the act of protruding.

g. **investiture** the act of ratifying or establishing in office.

The examples in (11) indicate that the related verb is not always morphologically related. This pattern could, therefore, be used as a means of accessing semantically related verbs and nouns or as a tool for the construction of a semantic network.

"The act of Ving" definitions have a subpattern which consists of "The act of Ving or the state of being <adj>." There are not many examples of this subpattern, but in all but one instance the noun being defined, the verb and the adjective are morphologically related.

13a. **adornment** the act of adorning or the state of being adorned.

b. **popularization** the act of popularizing or the state of being popularized

c. **nourishment** the act of nourishing or the state of being nourished.

d. **intrusion** the act of intruding or the state of being intruded.

e. **embodiment** the act of embodying or the state of being embodied.

In contrast, our data do not support the use of the corresponding formula "The state of being"-past part. for identifying stative verbs. Many instances of this pattern appear to be passives or stative use of normally non-stative verbs. This position is supported by the presence of a fair number of definitions which conjoin the two formulae.

14a. **displacement** the act or process of displacing: the state of being displaced.

b. **examination** the act or process of examining: the state of being examined.

c. **expansion** the act or process of expanding. The quality or state of being expanded.

It is likely that the formula "The quality or state of being"-past part. is a stative verb indicator when it does not co-occur with "Act of" definitions. Support comes from the frequency with which that pattern alternates adjectives, which are normally stative, with the past participle.

SELECTIONAL INFORMATION FOR VERB DEFINITIONS

Although the structure of verb definitions is much more limited than that of noun definitions, elements of verb definitions do provide interesting insights into collocational information. One striking example of this is the use of parenthetical information which flags typical instantiations of case arguments for the verb being defined. The most consistent of these patterns is "To"-V-(<"as">NP) where the NP is the typical object of the verb being defined.

15a. **mount** to put or have (as artillery) in position.

b. **lay** to bring forth and deposit (an egg).

c. **develop** to subject (exposed photographic material) to a usu. chemical treatment...

We are in the process of determining how consistent the parenthetical "as" is in signalling typical case relations.

SELECTIONAL INFORMATION FOR ADJECTIVES

Adjective definitions differ from those of nouns and verbs in that while nouns are virtually always defined in terms of other nouns and verbs in terms of other verbs, only about 10 percent of adjectives are defined in terms of other adjectives -- the rest are related to nouns or sometimes to verbs. Furthermore, the semantic information in an adjective definition refers more to the noun (or type of noun) modified by the adjective than it does to the adjective itself. This is because an adjective, together with the noun it modifies, defines a taxonomic relationship -- or, to put it another way, denotes a feature of the thing defined in the adjective+noun phrase. For instance, we can say either that the phrase "big dog" denotes a particular kind of (the more general term) "dog"; or that it denotes a dog with the additional feature of "bigness".

A useful piece of information we would like to get from adjective definitions is selectional information -- what sort of noun the adjective can meaningfully modify. Selectional restrictions are harder to find and are largely negative -- for instance, the formula "containing" defines adjectives that do not (in the sense so defined) modify animate nouns.

10a. **basic** containing relatively little silica.

b. **normal** containing neither basic hydroxyl nor acid hydrogen.

The same is true of some other moderately common formulae, such as "consisting of", "extending" and "causing". We hope that further analysis will allow us to find more indications of selectional characteristics of adjectives.

RECOGNIZING ACTION VS. STATIVE ADJECTIVES

One property belonging more to adjectives themselves than to their associated nouns is an active-stative distinction similar to that found in verbs. The test for an "active" adjective is that one may use it in a statement of the form "they are being ----" or in the command "be ----!" e.g. "be aggressive!" or "be good!", but not *"be tall!" or *"be ballistic!" As these examples indicate, most adjectives that can be used actively can also be used

statively -- aggressiveness or goodness may be thought of as a state rather than as an action -- but not the other way around.

Contrary to our expectations, the active-stative parameter of adjectives is much easier to identify in definitions than is selectional information. Some of the defining formulae discussed in Smith (1981) and Ahlswede (1985b) seem to be limited to stative adjectives. "Of or relating to", one of the most common, is one of these:

11a. **ballistic** of or relating to ballistics or to a body in motion according to the laws of ballistics.

b. **literary** of or relating to books.

Although many adjectives defined with "of or relating to" can be used actively in other senses, they are strictly stative in the senses where this formula is used:

12a. **civil** of or relating to citizens <~ liberties>.

b. **peaceful** of or relating to a state or time of peace.

The common formula "being ...", on the other hand, defines adjectives which at least lean toward the action end of the spectrum:

13a. **natural** being in accordance with or determined by nature.

b. **cursed** being under or deserving a curse.

Even such a normally stative adjective as "liquid" is relatively active in one of its senses:

14a. **liquid** being musical and free of harshness in sound.

By no means all formulae give indications of the stative-active qualities of an adjective. A large family of formulae ("having", "characterized by", "marked by", etc.) denoting attribution, are completely neutral with respect to this parameter.

SUMMARY

W7 contains a wealth of implicit information. We have presented methods for making some of this information explicit by focussing on specific formulae found in noun, verb, and adjective definitions. Most of these formulae appear at the start of definitions, but we have also demonstrated that important information can be extracted from syntactic and graphemic elements, such as parentheticals. The information we have extracted involves lexical relationships such as taxonomy and set membership, selectional restrictions, and special subcategories of nouns, verbs, and adjectives. This information is used by an automatic lexicon builder to create lexical entries automatically from W7 definitions.

REFERENCES

Ahlswede, Tnomas. 1985a. "A Tool Kit for Lexicon Building," **Proceedings of the 23rd Annual Meeting of the ACL**, Chicago, pp. 268-276.

Ahlswede, Thomas. 1985b. "A Linguistic String Grammar for Adjective Definitions," in S. Williams, ed., **Humans and Machines: The Interface through Language**. Ablex, Norwood, NJ, pp. 101-127.

Ahlswede, Thomas and Martha Evens. 1983. "Generating a Relational Lexicon from a Machine-Readable Dictionary." Forthcoming.

Amsler, Robert. 1980. **The Structure of the Merriam-Webster Pocket Dictionary.** Ph.D. Dissertation, Computer Science, University of Texas, Austin.

Amsler, Robert. 1981. "A Taxonomy for English Nouns and Verbs." **Proceedings of the 19th Annual Meeting of the ACL**, Stanford, pp. 133-138.

Amsler, Robert. 1982. "Computational Lexicology: A Research Program." **Proceedings of the National Computer Conference**, AFIPS, pp. 657-663.

Amsler, Robert and John White, 1979. **Development of a Computational Methodology for Deriving Natural Language Semantic Structures via Analysis of Machine-Readable Dictionaries.** TR MCS77-01315, Linguistics Research Center, University of Texas.

Apresyan, Yuri, Igor Mel'cuk, and Alexander Zholkovsky. 1970. "Semantics and Lexicography: Towards a New Type of Unilingual Dictionary," in F. Kiefer, ed., **Studies in Syntax and Semantics**, Reidel, Dordrecht, Holland, pp. 1-33.

Chodorow, Martin and Roy Byrd, 1985. "Extracting Semantic Hierarchies from a

Large On-Line Dictionary." **Proceedings of the 23rd Annual Meeting of the ACL**, pp. 299-304.

Evens, Martha and Raoul Smith. 1978. "A Lexicon for a Computer Question-Answering System", **American Journal of Computational Linguistics**, No. 4, pp. 1-96.

Evens, Martha, Bonnie Litowitz, Judith Markowitz, Raoul Smith, and Oswald Werner. 1980. **Lexical-Semantic Relations: a Comparative Survey**, Linguistic Research, Inc., Edmonton, Alberta, 1980.

Evens, Martha, James Vandendorpe, and Yih-Chen Wang. 1985. "Lexical-Semantic Relations in Information Retrieval", in S. Williams, ed., **Humans and Machines**. Ablex, Norwood, New Jersey, pp. 73-100.

Fox, Edward. 1980. "Lexical Relations: Enhancing Effectiveness of Information Retrieval Systems," **ACM SIGIR Forum**, 15, 3, pp. 5-36.

Kucera, Henry, and Nelson Francis. 1967. **Computational Analysis of Present-Day American English**, Brown University Press, Providence, Rhode Island.

Li, Ping-Yang, Thomas Ahlswede, Carol Curt, Martha Evens, and Daniel Hier. 1985. "A Text Generation Module for a Decision Support System for Stroke", **Proc. 1985 Conference on Intelligent Systems and Machines**, Rochester, Michigan, April.

Mel'cuk, Igor, and Alexander Zholkovsky. 1985. **Explanatory-Combinatory Dictionary of Russian**, Wiener Slawisticher Almanach, Vienna.

Mel'cuk, Igor, Nadia Arbatchewsky-Jumarie, Leo Elnitzky, Lidia Iordanskaya, and Adele Lessard. 1984. **Dictionnaire Explicatif et Combinatoire du Francais Contemporain**, Presses de l"Universite de Montreal, Montreal.

Michiels, A., 1981. **Exploiting a Large Dictionary Data Base**. Ph.D. Thesis, University of Liege, Belgium.

Michiels, A., 1983. "Automatic Analysis of Texts." **Workshop on Machine Readable Dictionaries**, SRI, Menlo Park, Ca.

Olney, John, Carter Revard, and Paul Zeff. 1967. "Processor for Machine-Readable Version of Webster's Seventh at System Development Corporation." **The Finite String**, 4.3, pp. 1-2.

Sager, Naomi. 1981. **Natural Language Information Processing**. Addison-Wesley, New York.

Smith, Raoul. 1981. "On Defining Adjectives, Part III." In **Dictionaries: Journal of the Dictionary Society of North America**, no. 3, pp. 28-38.

Wang, Yih-Chen, James Vandendorpe, and Martha Evens. 1985. "Relational Thesauri in Information Retrieval", **JASIS**, Vol. 36, No. 1, pp. 15-27.

Webster's Seventh New Collegiate Dictionary, 1963. G.&C. Merriam Company, Springfield, Massachusetts.

White, Carolyn. 1983. "The Linguistic String Project Dictionary for Automatic Text Analysis," **Workshop on Machine-Readable Dictionaries**, SRI, April.