Using Words

Louise Guthrie, Paul Mc Kevitt, and Yorick Wilks

Computing Research Laboratory
Dept. 3CRL, Box 30001
New Mexico State University
Las Cruces, NM 88003-0001, USA.

{louise, paul, yorick}@nmsu.edu

ABSTRACT

It is well established that the processing of natural language discourse requires pragmatic information about words. Pragmatic information should be added to knowledge bases for natural language processing systems. This information will be useful for both natural language understanding and natural language generation. Pragmatic information can be encoded by hand or, we claim, extracted from a machine-readable dictionary. Machine-readable dictionaries provide a ready-made source of pragmatic information which can be used by natural language processing systems. A number of examples of pragmatic information for words in the Longman's Dictionary of Contemporary English (LDOCE) are discussed.

"I use words to mean what I want them to mean,
neither more nor less."

Humpty Dumpty, in Alice in Wonderland

0. Introduction

While much research in natural language processing has concentrated on the syntax and semantics of words there has been relatively little emphasis on the usage of words. Much of the research has concentrated on lexicons which have syntactic and semantic information about words but little information about the use of the right word in the right context or the odd uses of words in other contexts.

The pragmatics of dictionary entries is concerned with the context of an uttered word and how words are used in natural language discourse. For example, the word "compare" can be used with the words "to" or "with". "With" is used more with long detailed studies as in "a book that compares the human brain with that of the elephant". A natural language generator should note such subtleties if it is to perform at the level of human use of language.

We argue here that natural language processors that understand and generate natural language by means of syntax and word meaning alone cannot possibly represent or generate utterances without pragmatic information. If we take the phrase "bread and butter" then a syntactic-semantic processor may get away with processing this in the utterance "The only food the prisoner had was bread and butter" but would not in "The European Community (EC) is not only concerned with bread and butter issues." In the latter case there needs to be information stored somewhere in the natural language program about the phrasal use of bread and butter and we argue here that such information can be retrieved from machine readable dictionaries.

It is pointed out by Leech and Thomas (1970) that the most serious cross-cultural language misunderstandings occur at the level of speaker-meaning or the pragmatic level. This has obvious implications for natural language processing problems such as machine translation, since a translation which conveyed the wrong pragmatic intent might cause serious problems. It might be better to have no translation at all than to have a pragmatically incorrect one.

1. Background

Recent interest in natural language systems that have vocabularies much bigger than the toy systems of the past has given rise to a great deal of research on how to use Machine-Readable Dictionaries (MRDs) in natural language processing (see Boguraev and Briscoe, 1989). Although most work on MRDs has the goal of extracting information which is useful to natural language processing, early work on MRDs was concerned with the drudgery of manipulating large files on not so large machines (see Amsler, 1989). Later work has examined how to make some of the implicit information in dictionaries, explicit (see Amsler and White 1979, Amsler 1980, Boguraev et al. 1989, Chodorow et al. 1985, Guthrie et al. 1990, Klavans 1990, Markowitz et al. 1986, Nakamura and Nagao 1988, Slator 1988a, 1988b, Slator and Wilks 1990, and Wilks et al. 1988, 1989, 1990). In particular, a great deal of work focuses on obtaining syntactic, and to some extent semantic, information from dictionary entries. We believe that consideration of the pragmatic information found in dictionaries will provide additional useful knowledge to natural language systems.

Several aspects of extracting syntactic and semantic information from MRDs have been studied. We shall mention some distinctions within these and then ask if similar distinctions exist for the pragmatic information found in dictionary entries.

The Longman Dictionary of Contemporary English (LDOCE) (see Procter et al. 1978), is a full-sized dictionary designed for learners of English as a second language that contains 41,122 headword entries, defined in terms of 72,177 word senses, in machine-readable form. The book and tape versions of LDOCE both use a system of grammatical codes of about 110 syntactic categories which vary in generality from, for example, *noun* to *noun/count* to *noun/count/followed-by-infinitive-with-TO*. The machine readable version of LDOCE also contains "semantic category" and "subject" codes that are not found in the book. The semantic category codes use primitives such as *abstract*, *concrete*, and *animate*, organized into a type hierarchy. This hierarchy of primitive types conforms to the classical notion of the IS-A relation as describing proper subsets. These primitives are used to assign type restrictions on nouns and adjectives, and place type restrictions on the arguments of verbs. The subject codes are another set of terms organized into a hierarchy. This hierarchy consists of main headings such as *engineering* with subheadings like *electrical*. These terms are used to classify words by subject. For example, one sense of current is classified as *geology-and-geography* while another sense is marked *engineering/electrical*.

2. Syntax and semantics in dictionaries

We distinguish three levels of syntactic and semantic information that can be identified in MRDs, and which provide important information for natural language processing.

2.1. Syntax

[1] Associated with each word sense defined in a dictionary is a syntactic category to which it belongs. Parts of speech are augmented with additional syntactic information such as count noun, transitive verb, etc. Mott et al. (1986) used the broad syntactic category (e.g. noun, verb, adj) of an entry to tag pieces of text by part of speech, for information retrieval purposes.

[2] Other work identifies the syntax used in the defintion text for entries in a particular dictionary Markowitz et al. 1986 suggests that certain inferences about verb classes can be made from the syntax of definitions.

[3] Boguraev et al. (1989) have studied the syntax of an entire entry in LDOCE and Webster's Seventh (W7) (e.g. headword, pronunciation, part of speech, defintion, examples). This has been used to create a database of dictionary sense definitions which provides ellided information, subdivided sense definitions, identified example sentences, and given cross-references.

Overall, we can talk about the syntax of the headword, the syntax of a particular part of the entry, like the definition text or the example sentence and the syntax used for the entire entry.

2.2. Semantics

[1] LDOCE provides semantic category information for each noun word sense which appears in the dictionary. Examples of these semantic categories are human, movable-solid, liquid, gas. LDOCE uses 34 semantic categories in all. Verb and adjective word senses appear with selection restriction information, i.e. the categories for the arguments of verbs, or the category for the noun which the adjective modifies. Slator et al. (1990) have identified semantic categories for preposition word senses using clustering algorithms. The automatic identification of semantic categories for verbs from machine readable dictionaries will provide important information for natural language processing, but research in this area is in preliminary stages.

[2] Extracting semantic information from the definition texts is a difficult problem, but one which is of interest to many research groups (see Amsler and White 1979, Amsler 1980, Boguraev et al. 1989, Chodorow et al. 1985, Guthrie et al. 1990, Klavans 1990, Markowitz et al. 1986, Nakamura and Nagao 1988, Slator 1988a, 1988b, Slator and Wilks 1990, and Wilks et al. 1988, 1989, 1990). This work includes finding the genus terms for definitions (terms that satisfy an ISA relation with the headword of the defintion) (see Amsler and White 1979, Amsler 1980, Chodorow et al. 85, Nakamura and Nagao 1986, and Slator 1988); and identifying the sense of the genus word (see Guthrie et al. 1990, Klavans 1990); identifying other relations that might exist between words but which are implicit in dictionaries (see Amsler 1980, Guthrie et al. 1990, Nakamura and Nagao 1986); and extracting case relations (see Slator 1988a, 1988b, Slator and Wilks 1990).

[3] As for the semantics of an entire dictionary entry, we know of no work in this area, but we suggest that this is precisely the semantic information that distinguishes one word sense from another. It is the combination of the semantic information in the definition text, the example text, the grammar category, the cross-references and the other fields in a dictionary entry that must be captured in order to map word sense sets across dictionaries as suggested in Byrd (1989).

3. Pragmatics in dictionaries

We suggest that in addition to the research on extracting syntactic and semantic information from MRDs, we should begin examining the pragmatic information that can be found in the dictionaries. We look at LDOCE as a rich source of pragmatic information and discuss why we see this information as having similar distinctions to those made above for syntax and semantics: (1) the pragmatic information (uses) about the word-sense being defined, (2) the pragpragmatic information which is available in the defintion of the word-sense, and (3) the pragmatic information associated with an entire entry.

3.1. Pragmatic information in LDOCE

The compilers of LDOCE took special care in their new edition to mark out pragmatic information in the dictionary. Quirk (1987) says, "Special attention can thus be given to the known needs of advanced students, needs which include the most up-to-date meanings and such pragmatic aspects of usage as courtesy, intention, and speaker-addressee relations." These three types of information for lexical entries could be considered pragmatic.

For example, such aspects may give advice as to the appropriateness of a given word in a particular context, the formality of it, or its connotations. In addition to usage notes, definitions often contain information about collocation and appropriate choice: Words tend to co-occur with other words. Collocations are shown in examples and those collocations which are particularly fixed are shown in heavy type. The "subject" codes described above that are found in the machine readable version of LDOCE provide a context for the word sense being defined, and can be considered to give some pragmatic information about the word. Usage notes also provide advice on which word has appropriate meaning in a particular context.

Usage notes form part of the alphabetic entries for words. and cover four areas: (1) word sets which explain the differences between words of roughly similar meaning (e.g. Usage note at "fat" explains the words "chubby", "stout", and "overweight"; (2) difficult points of grammar and style which explain, for example, whether a plural pronoun can be used with words like anyone and someone; (3) differences in British and American English (e.g. the use of the word "hire" in British and American English); (4) information about pragmatics explain the way some words and phrases can be used in conversation to suggest a meaning or attitude that could not be derived from literal word meanings (e.g. bread and butter).

What is needed initially is a pragmatics processor which takes entries in an MRD and initially detects the keyword USAGE. The processor would then search for important keywords like VERY COMMON. The next step would be to build knowledge representations of pragmatic information for words from the algorithms which process over usage sections. The database derived would then be used in natural language programs.

3.2. Examples

This section gives some examples of the usage sections for words in LDOCE. We have chosen the words (1) bread, (2) compare, (3) between, (4) attendant, (5) biweekly, (6) continual, (7) admittance, (8) blind, (9) burned and (10) hire.

The usage note for bread is shown below pointing out that in the case of bread and butter meaning pieces of bread with butter on them then it is treated as a singular verb. A processor would possibly look for occurrences of WHEN X MEANS Y do Z and encode this pragmatic information about the word.

BREAD

USAGE When bread and butter means pieces of bread with butter spread on them, it takes a singular verb. *This bread and butter is too thick*. Compare: *I bought bread and butter at the shop, and they cost 50p.*

Much of the pragmatic information found in LDOCE is given in the USAGE sections of the definition, as in the example above. It should be noted that, in some cases, information about pragmatic usage is also available in the definition itself. The phrasal definition for bread-and-butter shown below, is such an example. More complicated algorithms for parses of word definitions would need to be used here to derive pragmatic knowledge.

BREAD-AND-BUTTER

DEF 1: concerned with the things that are necessary for life : *low wages, bad houses to live in, and other bread-and-butter political questions*

DEF 2: that can be depended upon : *"Hamlet and Othello" are the bread-and-butter plays of our theatre group*

DEF 3: sent or given as thanks for being treated well by one's host or hostess : esp. in the phr. *a bread-and-butter letter*

The usage note for compare has information about what can follow the word. Keywords such as FOLLOWED BY, MORE OFTEN are significant here.

COMPARE

USAGE Compare can be followed by to or with : *He compared London to / with Paris. London is large, compared to / with Paris.* With is more often used if we are speaking of a long detailed study : *a book that compares the human brain with that of the elephant.* In comparison / by comparison is followed by with, not to : *Paris is small in comparison with London.*

The usage note for between discusses the use of among and between and points out that between should be followed by 2 things. Again, FOLLOWED BY is an important keyword for the retrieval of pragmatic information. Other useful keywords are COMMON, NONSTANDARD, WHEN WE SPEAK OF, ALWAYS USE.

BETWEEN

USAGE Compare among and between: 1) Between must be followed by 2 things. It is right to say *between the 2 houses* or *between each house and the next.* It is common, but nonstandard, to say *between each house.* 2) Some books say that between should be followed by 2 things only, and among by 3 or more : *Divide it between the 2 / among the 3 children.* But when we speak of clear and exact position we always use between : *Ecuador lies between Colombia, Peru, and the Pacific Ocean.*

The next two examples anticipate possible errors or difficulties that a speaker might encounter. This information would not be available in word sense definitions.

ATTENDANT

USAGE An attendant is not someone who attends a play, concert, or church service. Someone who works in a shop is a shop assistant.

BI (prefix)

USAGE Expressions like biweekly are confusing, because they may mean "twice in one week / month / year" or "once in 2 weeks / months / years".

The following four definitions indicate the connotations or mood of a particular word.

CONTINUAL

USAGE Continual is often used of bad things : *continual hammering / these continual interruptions.* Continuous is used of things or events that are connected without a break, but may have a beginning and end : *3 days' continuous flight / 2 rivers connected to form one continuous waterway.*

ADMITTANCE

USAGE In the meaning "permission to go in" admittance is more formal than admission, which is a the more ordinary word. The entrance price is the admission, not the admittance. Admittance could not be used in an expression like "*his admittance of guilt*".

BLIND

USAGE Blinded and deafened are only used when there is a clear cause : *He was blinded by dust / blinded in the war./ The music was so loud I was nearly deafened.* Otherwise use the adjectives blind, deaf : *He became blind / deaf ./ a deaf / blind child*

The following examples illustrate cultural distinctions in the use of words.

BURNED - BURNT

USAGE The British use **burned** as the past tense and participle of burn, only when it is 1) INTRANSITIVE : *The fire burned brightly.* 2) (fig.) *The desire for freedom burned in their hearts.* Otherwise the British past tense and participle is **burnt** : *I ('ve) burnt the dinner!* Americans can use **burned** all the time, but may also use **burnt**, esp. as an adjective : *burnt bread.*

HIRE

In British English one **hires** clothes or a boat for a short time, and their owner hires them out ; *one rents or hires a car; one rents a house, paying regular rent , and the owner lets it. The owner lets out a room or part of a building.* In American English one **rents** all these, and their owner **rents** them out.

4. Conclusions

Pragmatic information on word usage is useful and necessary for natural language processing. Much current work on building lexical information for natural language programs concentrates on the syntax and semantics of words. Pragmatic information is just as important and we argue here that such information can be automatically extracted for Machine Readable Dictionaries (MRD's).

Our initial analyses indicate that much pragmatic information is encoded in the USAGE section of the LDOCE dictionary entries. We argue that such information can be extracted by the recognition of keywords and then further parsing of the information given in the usage section. Examples of usage sections for a number of words are shown and important keywords indicated.

The pragmatic information described in Section 3.2 can be used by a natural language processing system for generation of discourse with a particular style. Style can be of a certain type and then appropriate words are selected which conform to that style type. Also, a machine translation system could have translations done culture-specific for given countries. As we have shown, such cultural information is available in the dictionary.

Further work would involve determining a good set of keywords to be recognized in usage sections. Then a parser could be used in processing the usage information and a knowledge representation of pragmatic information derived for words. The usefulness of this pragmatics knowledge representation could then be tested for a natural language understander/generator. We have identified three levels of syntactic and semantic information that has been studied for machine readable dictionaries. We believe that similar distinctions are useful for organizing the pragmatic information in LDOCE.

The "subject codes" of the machine readable version of LDOCE provide a pragmatic category for each word. The usage notes provide pragmatic information found in one segment of the entry. As is the case with the global semantic information, it is an open question as to what the pragmatic information found in an entire entry will be good for. Our strong belief is that the pragmatic information found in the usage notes, in the definition text, and the collocation information identified in the example text can be formalized into a "pragmatic information" representation, corresponding to a dictionary entry, before sense definitions can be characterized in a way which allows us to achieve significant results on the difficult problems associated with mapping across senses in differnent dictionaries.

5. References

Amsler, Robert A. (1989). Third Generation Computational Lexicology, *Proceedings of the First International Lexical Acquisition Workshop*, Detroit, Michigan.

Amsler, Robert A. and John S. White (1979). Development of a Computational Methodology for Deriving Natural Language Semantic Structures via Analysis of Machine-readable Dictionaries. NSF Technical Report. (MCS77-01315).

Amsler, Robert A. (1980). The Structure of the Merriam-Webster Pocket Dictionary. Technical Report. (TR-164). University of Texas at Austin. Ph.D. Thesis.

Boguraev, Branimir, Roy Byrd, Judith Klavans and Mary Neff. (1989). From Machine Readable Dictionaries to a Lexical Knowledge Base. *Proceedings of the First International Lexical Acquisition Workshop*, Detroit, Michigan.

Boguraev, B. and E. Briscoe. Introduction. (1989). *Computational Lexicography for Natural Language Processing*, Edited by Bran Boguraev and Ted Briscoe, Longman Group UK Limited.

Byrd, Roy J. (1989). Discovering Relationships Among Word Senses. *Proceedings of the 5th Annual Conference of the UW Centre for the New Oxford English Dictionary*, pp. 67-80. St. Catherine's College, Oxford, England.

Chodorow, Martin S., Roy J. Byrd, and George E. Heidorn (1985). Extracting Semantic Hierarchies from a Large On-Line Dictionary. Proceedings of the 23rd Annual Meeting of the ACL, pp. 299-304. Chicago, IL.

Guthrie, Louise, Brian Slator, Yorick Wilks and Rebecca Bruce (1990). Is there content in Empty Heads?, To appear in *Proceedings of Coling 90*, Helsinki, Finland.

Klavans, Judith, Roy Byrd, Nina Wacholder, and Martin Chodorow (1990). Taxonomy and Polysemy, Pre-print presented at the IBM Internal Technical Meeting, Paris, France.

Leech, Geoffrey and Jenny Thomas (1987) *Pragmatics and the dictionary*. In Longman Dictionary Of Contemporary English (LDOCE), New Edition, Essex: Longman, pp. F12-F13.

Markowitz, Judith, Thomas Ahlswede, and Martha Evens (1986). Semantically Significant Patterns in Dictionary Definitions. Proceedings of the 24th Annual Meeting of the Association for Computational Linguistics, pp. 112-119. New York.

Mott, Paul M., David L. Waltz, Howard L. Resnikoff, and George G. Robertson (1986). Automatic Indexing of Text. Technical Report No. 86-1, Thinking Machines Corporation, Cambridge, Ma.

Nakamura, Jun-ichi, and Makoto Nagao (1988). Extraction of Semantic Information from an Ordinary English Dictionary and its Evaluation. *Proceedings of COLING-88*, Budapest, Hungary. pp. 459-464.

Procter, Paul et al. (1978). *Longman Dictionary of Contemporary English (LDOCE)*. Harlow, Essex, UK: Longman Group Ltd.

Quirk, Randolph (1987) *Preface*.ll In Longman Dictionary Of Contemporary English (LDOCE), New Edition, Essex: Longman, pp. F7.

Slator, Brian M. (1988a). Constructing Contextually Organized Lexical Semantic Knowledge-bases. *Proceedings of the Third Annual Rocky Mountain Conference on Artificial Intelligence (RMCAI-88)*. Denver, CO, June 13-15, pp. 142-148.

Slator, Brian M. (1988b). Lexical Semantics and a Preference Semantics Analysis. Memoranda in Computer and Cognitive Science. *(MCCS-88-143)*. Las Cruces, NM: Computing Research Laboratory, New Mexico State University. (Doctoral Dissertation).

Slator, Brian M. and Yorick A. Wilks. (1987). Towards Semantic Structures from Dictionary Entries. *Proceedings of the Second Annual Rocky Mountain Conference on Artificial Intelligence (RMCAI-87)* Boulder, CO, June 17-19, pp. 85-96.

Slator, Brian M., Shahrzad Amirsoleymani, Sandra Andersen, Kent Braaten, John Davis, Rhonda Ficek, Hossein Hakimzadeh, Lester McCann, Joseph Rajkumar, Sam Thangiah, Daniel Thureen (1990). *Towards Empirically Derived Semantic Classes.* In Proceedings of the 5th Annual Rocky Mountain Conference on Artificial Intelligence (RMCAI-90). Las Cruces, NM, June 28-30.

Slator, Brian M. and Yorick A. Wilks (Forthcoming - 1989). Towards Semantic Slator, Brian M. and Yorick A. Wilks (Forthcoming - 1990). Towards Semantic Structures from Dictionary Entries. *Linguistic Approaches to Artificial Intelligence.* Edited by Andreas Kunz and Ulrich Schmitz. Frankfurt: Peter Lang Publishing House. (Revision of *RMCAI-87*).

W7 (1967). *Webster's Seventh New Collegiate Dictionary.* C. & C. Merriam Company, Springfield, MA.

Wilks, Yorick A., Dan C. Fass, Cheng-Ming Guo, James E. McDonald, Tony Plate, and Brian M. Slator (1988). Machine Tractable Dictionaries as Tools and Resources for Natural Language Processing. *Proceedings of the 12th International Conference on Computational Linguistics (COLING-88)*, pp. 750-755. Budapest, Hungary. Aug. 22-27

Wilks, Yorick A., Dan C. Fass, Cheng-Ming Guo, James E. McDonald, Tony Plate, and Brian M. Slator (1989). A Tractable Machine Dictionary as a Resource for Computational Semantics. *Computational Lexicography for Natural Language Processing.* Edited by Bran Boguraev and Ted Briscoe. Harlow, Essex, UK: Longman and New York: Wiley and Sons. pp. 193-228.

Wilks, Yorick A., Dan C. Fass, Cheng-Ming Guo, James E. McDonald, Tony Plate, and Brian M. Slator (Forthcoming - 1990). Providing Machine Tractable Dictionary Tools.*Computers and Translation.* Also to appear in *Theoretical and Computational Issues in Lexical Semantics (TCILS).* Edited by James Pustejovsky. Cambridge, MA: MIT Press.