Building a Computational Lexicon Using Machine Readable Dictionaries

Judith L. Klavans

Abstract. Machine-readable dictionaries (MRD's) provide a key resource for building computational lexicons. Computational lexicons are dictionaries used by natural language systems, and they are different in nature from dictionaries built for human users. This paper gives some examples of the type of information needed for building computational lexicons. We then show some of the ways we have used MRD's to extract both explicit and implicit information in order to build a computational lexicon. We discuss some of the difficulties inherent in utilizing MRD's, in particular the problems of polysemy and mapping. We present preliminary results of a study involving extraction of verbs of manner of movement from different dictionaries, starting with WEBSTER'S SEVENTH NEW COLLEGIATE DICTIONARY (henceforth W7). We argue that the most productive way to utilize the semantic information in MRD's is by mapping from MRD's into an independent Lexical Knowledge Base, rather than by mapping between MRD's.

A description of the Computational Lexicon. Entries in a computational lexicon differ in critical ways from standard dictionary entries. For example, Figure 1 gives extracts of the entry for *see* taken from the computational lexicon built by the IBM Lexical Systems Group. An explanation of the features and attributes is given below.

```
(HEADWORD(see))
(POS(VERB))
(MORPH(INFLECTION(IRREG)))
                    (PASTFORM saw)))
                    (PASTPARTFORM seen)))
(PHON(AXNT))
(SYNTACTIC(CONSTRUCTION(MWESTART)))
                          (VADV off)))
                          (VADV through)))
                          (VPREP about)))
           (INHERENT(INF)))
           (SUBCAT(COMPTYPE(THATCOMP))))
                                (WHCOMP))))
                                (TRAN)))
(SEMANTIC(INHERENT(SENS)))
```

Figure 1: The verb *see*

Figure One is a far cry from what most published dictionaries would list for *see* although some of the information is the same. The first line is the HEADWORD. Next, the field POS gives the part of speech. Then comes a set of MORPHological irregularities, namely that the past and past participal forms of *see* are *saw* and *seen*. Notice that PASTFORM and PASTPARTFORM are attributes with specified values (in this case *saw* and *seen*) whereas IRREG is simply a feature. IRREG has a binary value, and it is a characteristic of the word itself. So far, each of these pieces of information could be found explicitly in published dictionaries. However, the next feature, AXNT, a PHONological feature, is not so easy to find in most published dictionaries. AXNT applies to a word which is accented on the final syllable, or, as with the case of *see*, a one-syllable word, on the only syllable. Some examples of other words with this feature are *abate, allude*, and *annoy*. The feature is needed in a dictionary of this type since it determines the doubling of final consonants when adding certain suffixes beginning with vowels. The feature AXNT can be derived from the pronunciation fields of most dictionaries by looking to see which syllable carries the accent mark. However, AXNT is not explicitly stated, unlike part of speech or irregular forms.

The next set of features is SYNTACTIC in nature. The first type, CONSTRUCTION, reflects the fact that *see* can be the START of a Multi-Word Entry (MWESTART). In this example the two constructions are the verb-adverb (VADV) construction *see off* and the verb-preposition (VPREP) construction *see through*. Verb-particle constructions in English are notoriously difficult to collect and categorize, partly because they are so productive and often idiosyncratic in meaning, and partly because their syntactic properties are complex. Most published dictionaries, and especially learner's dictionaries, have some verb-particle collocations identified, although not all of the grammatical distinctions are specified. In fact, usually an example suffices so an explicit explanation may not, in fact, be helpful to the learner. The difference between a VADV and a VPREP construction is that a VADV construction does not require an object noun phrase, as in *We added the money up* and *The money added up*. On the other hand, a VPREP construction requires an object noun phrase after the prepositional particle in declarative sentences. There are other differences concerning separability in interrogative constructions which we will not go into here. Often the same verb collocates with the same particle, but one sense is grammatically VADV and another is VPREP.

The INHERENT SYNTACTIC feature shows that the form *see* is the INFINITIVE form of the verb. The SUBCATegorization feature THATCOMP expresses the fact that *see* can occur with a clause beginning with *that*, as in *I saw that he was sleeping*. *See* can also take a complement clause beginning with a *wh*-word, as in *I saw why he was tired*, so *see* is labelled WHCOMP. With the exception of LONGMAN DICTIONARY OF CONTEMPORARY ENGLISH (Longman 1978, henceforth LDOCE) and the COLLINS COBUILD ENGLISH LANGUAGE DICTIONARY (1987) this information is not usually explicitly stated in monolingual or bilingual dictionaries, but it may be embedded within example sentences. Sometimes the sense of the word with the *that* clause is separated from other senses, so again example sentences might give the **human** reader a clue about the syntax of the verb. However, unless this property is clearly stated, it will be difficult for a **computer program** to figure it out, even though the human user can. In contrast, the final feature TRANsitive is usually explicitly marked in the dictionary so a program could extract it without difficulty.

The word *see* in this example has only one SEMANTIC feature listed, and this is an INHERENT feature of the verb, rather than a contextual feature. The feature

SENSe holds for a small set of sense verbs in English which omit the *to* in an infinitival complement. For example, *I saw him open the door* but not **I saw him to open the door*. This is an example of an inherent semantic property of a verb which has a syntactic consequence. An example of another inherent semantic feature of verbs is CMMNCT, which holds of a verb of communication such as *tell.* For CMMNCT verbs, the noun phrase object must be animate and generally it must be human. The subject also must be animate and generally it must be human unless the impersonal 'it' is used as the subject. Examples of verbs with this features are *amaze, delight, remind,* and *stun*. Other examples from the structured version of this lexicon are given in Appendix One.

Who Uses a Computational Lexicon? The computational lexicon is created to be used by computer programs for natural language applications. The types of computer systems that might use computational lexicons are machine translation systems, language parsers, language generators, knowledge representation systems, expert systems, and others. For example, a language generator might need to know what kind of complement structure a verb can take. The system will need to know that *I believe that I will come* is acceptable but * *I belive to come* is not. In contrast, *I decided that I would come* is acceptable, and so is *I decided to come*. This information might be found in a traditional dictionary in the example sentences, but it might not be mentioned explicitly. To take another example, a verb like *eat* permits deletion of its direct object in English, but the verb *meet* does not, unless conditions of semantic plurality are fulfilled on the subject. Thus, *I ate* implies *I ate something*, but **I met* is simply not acceptable English, although *we met* or *the committee met* is completely grammatical. Finally, to take an example from semantics, a knowledge system will need to know which nouns are semantically female and which are male. Similarly, for machine understanding, it is necessary to know what nouns are synonyms and hyponyms of *house* to know if the sentence *She lived in a X* makes sense.

What Information Belongs in the Computational Lexicon? The sample entry shown in Fig. 1. One is the tip of the iceberg. Most systems require entries that contain a fuller range of semantic and syntactic information. See Ingria (to appear) for a survey of types of syntactic information found in some existing computational lexicons. Some lexicons represent syntactic information separately from the semantic (or concept) lexicon, and then contain a set of rules often called *linking rules* to state generalizations about the syntactic expression of arguments bearing particular semantic roles. For example, the possessor role of one sense of the verb *have* is carried by the subject, as in *I have a book*. In contrast, in the sentence *She had him clean the porch*, the subject is an actor but not a possessor. Other systems may specify valency (Allerton 1982), thematic roles (Jackendoff 1987), transitivity alternations (Katz and Levin 1988 and Levin, to appear) and other general verb features. Some lexicons contain subcategorization only, such as the Brandeis verb lexicon (developed by Jane Grimshaw and Ray Jackendoff for 950 English verbs), or Gross (developed for French for some 10,000 verbs), described in Gross (1975). In addition, each system usually has special requirements imposed by the architecture and function of the system. For example, an entry from a language translation system will typically include information specific to the source and target languages. An entry from a question-answering system might include specific information about the data base itself. However, all systems need to be able to analyse and generate English sentences, so they each need to have access to a certain core of common information about words. The approach we are following is to represent the core overlapping information in our computational lexicon, and to let individual projects add whatever information is tied to their particular applications.

The computational lexicon illustrated in Fig. 1. is comprised of a set of entries consisting of features and attribute-value pairs. The system (called UDICT, the "Ultimate DICTionary) is described in Byrd (1984). The derivational and inflectional morphological analyzer is described in Byrd (1983) and Byrd et al. (1986). The linguistic motivations for the features and attributes are described in Klavans and Wacholder (1988). Other aspects of our computational lexicon are discussed in Klavans 1988. Among the standard electronic dictionaries that were used in building this lexicon were:

- definitions, synonyms, and etymologies from W7,
- taxonomy files created from W7 using techniques reported in Chodorow et al. (1985),
- grammatical information from LDOCE. In the future, we plan to use:
- definitions from LDOCE and W7,
- synonyms from the COLLINS THESAURUS (Collins 1984),
- entries from the COLLINS bilingual dictionaries for English/Italian (Collins 1980), English/French (Collins 1978), English/Spanish (Collins 1971), and English/German (Collins 1980).

In addition to using MRD's, we anticipate incorporating information from large corpora. There are numerous other sources of lexical information which are not available in electronic form, but which we have entered into UDICT. However, our goal is to extract automatically the maximum of information from our machine readable sources.

Each entry in UDICT consists of lists of features[1] and attribute-value pairs. There is one list merged across senses for each part of speech. For example, the word *claim* has two parts of speech in UDICT, here shown in an abbreviated format[2] (different from Fig. 1):

- claim: (NOUN SING AXNT FACTVE TOV STORED)
- claim: (VERB TRAN AXNT PRES INF THATCOMP STORED HUMSJ COLLHUMSJ HUMEXPSJ)

The question is to decide what features to put into the feature bundle. This is not a trivial matter but there are several options. One is to put only those features that apply to all senses of a word, that is, the *intersection* of the set of features for each sense. Another would be to list the *union* of all features for each sense. Of course, there is the best option of representing different senses of a word, with the corresponding set of features, but then this brings along another more fundamental problem: what is a sense?

Problems in Building a Sense-Disambiguated Computational Lexicon. Consider a system such as that reported in Boguraev (1987) and Boguraev (to appear) in which sense distinctions are in fact made. The grammar development system, intended for a Generalized Phrase Structure Grammar (GPSG), utilizes the grammatical codes from LDOCE as the basis for the listing of feature-value sets. However, notice that this system is forced to accept the sense distinctions from LDOCE, for better or for worse. Similarly, the project described in Wilks et al. (1988) uses LDOCE definitions as the basis for lexical semantic structures. Semantic information is to be extracted from dictionary entries in LDOCE to build sense frames. These structures (with some enhancements) are to provide the basis for knowledge-based parsing. Both projects are pursuing important paths in Natural Language (NL) research, and in particular in the use of machine readable dictionaries. However, each is constrained by the sense distinctions dictated by LDOCE. Similarly, we in the Lexical Systems Group at IBM Research have adopted the sense distinctions in W7 for our taxonym dictionary (see Chodorow et al. 1985). Although W7 has more headwords than LDOCE, the problem of sense distinctions still remains. Further, most dictionary writers have been obliged to merge important grammatical distinctions for the sake of space. As human readers, we may be able to decode such abbreviations, but it is doubtful that computers are capable of such interpretation. Take for example, the entry for the verb *button* from LDOCE:

> button (v)
> T1;IO;
> Subject area: clothing;
> Subject: Human;
> Direct Object: Moveable Solid
> to (cause to) close or fasten with
> buttons: to button (up) one's shirt.
> *My shirt doesn't button (up) easily.*

The entry is listed as requiring a human subject, yet the example sentence has the surface subject *shirt*. The problem here is that the underlying **agent** is *human* but not the surface **subject**. Regular alternations like this are characteristic of fasten-type verbs, such as *zip, clip, lock*. The alternation is sometimes captured implicitly in the definition in the form of the parenthesized (*cause to*), coupled with the fact that the same sense is marked both as *transitive (TI)* and *intransitive (IO)*, but this is in no way explicit in the dictionary itself. The human user might know this about button-type verbs, but it is impossible for a computer program to detect information that is not explicit. The more the program has to guess, the more room for error and, hence, the less useful the resource.

To sum, there are various solutions to the problem of how to list features and attributes. When only one entry is available, the solution to list only the intersection of features (the approach in most of UDICT especially for inherent features holding of nouns) or the solution to list the union of features (taken for the contextual features for verbs in UDICT) does not capture the fact that different senses of a word exhibit different syntactic behavior. Important information is obscured and omitted by these approaches. On the other hand, the solution chosen by Wilks *et al.* (1988) or by Boguraev (1987) and Boguraev (to appear) is to take the sense distinctions provided by LDOCE. But this then requires a system to adopt LDOCE senses, even when they are incomplete or incorrect. In order to use more than one MRD, a way to map senses in one dictionary onto senses in another is required, since sense distinctions across dictionaries rarely correspond.

Mapping vs. Lexical Knowledge Bases. There is a new sense-disambiguated computational lexicon, COMPLEX, which we are currently planning. The goal is to extract information from many machine readable sources into a large sense-disambiguated COMPutational LEXicon (COMPLEX). This lexicon will contain information that is common to systems, and could be accessed by these NL systems. Each application can then enhance the base lexicon (or even eliminate unecessary information) as needed for its customized lexicon. COMPLEX can be viewed as a Lexical Knowledge Base, independent of any specific dictionary, as shown in Appendix Three. One of the major problems that we are tackling in building the new broad-coverage computational lexicon is the representation of polysemous lexical items. Until the problem of sense distinctions is tackled, any computational lexicon will be of limited usefulness. The other problem particular to using machine readable dictionaries is the mapping problem, discussed below.

We have examined mapping between LDOCE, W7, NEW COLLINS THESAURUS, and ROGET'S THESAURUS, and have found it extremely difficult (if not impossible) to map a given sense in one dictionary into a single sense in another. Appendix Two shows one of our most successful attempts. The senses of *mangle* group into the sense of physical disfigurement vs. the sense of pressing wet clothes. In cases like this where a word has two distinct senses, and where the word is not very frequent, the problem of mapping appears to be tractable. (The past participle *mangled* occurs only once in the Brown corpus; in a different corpus of just over a million words, the form *mangled* occurs twice, and *mangling* occurs once). But despite this clear case, so far we have been unable to come up with a convincing way to automatically map even these senses onto each other.

Alternatively, one could abandon the task of mapping dictionaries onto each other and adopt a different approach. One could chose to compose a set of ideal data structures, and then hunt in various resources, including dictionaries, for information which completes the required fields. This is the proposal set forth in Atkins (1987)[3], and it is the route we are currently pursuing. It is also the approach of Calzolari and Picchi (1988), who propose moving from Lexical Data Bases to Lexical Knowledge Bases, and of Fox *et al.* (1988) who have used information from two MRD's to organize information into a semantic network for use in information retrieval. Calzolari and Picchi view the MRD as the "primary source of basic general knowledge" (*op cit* p. 87), from which a knowledge base is derived. A sketch of the Mapping Position contrasted with the Lexical Knowledge Base approach is given in Appendix Three.

An example of Extracting Verb Types.[4] Even with a knowledge base, the question still remains: what information do systems need? A related question is whether that information can be extracted from our MR sources. Other researchers, for example, Atkins, Kegl, and Levin (1986 and 1988) have examined the problem of extracting implicit information from dictionaries, with a somewhat negative prognosis. However, the picture is not altogether bleak. If facts to be extracted rely on certain types of semantic information within the dictionary, then results appear to be more promising than attempts to extract syntactic or linking information. The verb type that we tested was manner of movement verbs. The question asked was whether we could automatically find all verb senses belonging to one semantic class. If so, then we might be able to determine, for example, what the function of a prepositional phrase within a definition might be. At the same time, if we find the manner of movement verbs in a given dictionary, then we can check to see if each prepositional phrase is of manner. Finally, we hypothesized that we could generalize this approach to other verb classes and to other dictionaries.

We started by extracting the hyponyms of *move, go, walk, proceed,* and *advance* from our taxonym dictionary. The tyxonym dictionary (Chodorow *et al.* 1985) is a hierarchy derived from genus terms in W7. We then hand-edited this list, and expanded the list using Filtering. Filtering is a way to use the taxonym files to argument a list of words with a given trait with other words hypothesized to have that trait. We came up with several categories of movement verbs: (1) Manner — *crawl, flounce, hobble,* (2) Sound — *brush, clatter, rustle,* (3) Speed — *accelerate, belt, canter,* and (4) Inherent Direction — *ascend, descend, shin.* We hand-edited the list, added some verbs, and then extracted a test list of thirty-one core verbs from the manner of movement category to use in determining what the properties of the definitions were. We first needed to check for internal consistency in definitions within W7, and then eventually in other dictionaries (such as LDOCE, COLLINS ENGLISH-FRENCH, etc.) The approach was to use the genus terms and modifiers to guess and pick out intransitive senses of verb headwords which are manner of movement senses. Fig. 2 shows some of the information from the parses for three of our core verbs. There are other fields which are not included here.

1. LIMP (vi) to walk lamely; esp : to walk favouring one leg

GENUS	walk	
ADVERBIAL	lamely	
QUALIF—GENUS	walk	
QUALIF—ADV	favouring one leg	
SYNONYMS		
EXAMPLES		

2. REEL (vi) to turn or move round and round : WHIRL

GENUS	turn	move
ADVERBIAL	round and round	
QUALIF—GENUS		
QUALIF—ADV		
SYNONYMS	WHIRL	
EXAMPLES		

3. STAGGER (vi) to rock violently : SHAKE ⟨the ship -ed⟩

GENUS	rock	
ADVERBIAL		violently
QUALIF—GENUS		
QUALIF—ADV		
SYNONYMS	SHAKE	
EXAMPLES	the ship -ed	

Figure 2: Parses for Manner of Movement Verbs

The 8299 intransitive verb senses from W7 were parsed in this way.

Senses of intransitive verbs that might qualify as manner of movement were indentified in a three-step procedure. The first step was to identify all senses that might qualify as manner of movement definitions. This was done by looking for a definition whose genus term matched one of the genus terms on our target list. These were the *called* senses. The next step was to determine which of the called senses was the best one for a given headword. This was done by seeing which genus term came first. In later work, we intend to rank senses based on information in the differentia. This second reduced set is the *chosen*. Finally, if the genus of the definition contained none of the targeted genus terms, we looked up the genus term(s) in the taxonym dictionary. If the first level hypernym contained one of the genus terms we were looking for, then that sense went into a separate file. This third step was incorporated to identify those senses which are in the desired semantic field, but may not have matched the exact genus from the target list.

We ran two selected sets of target genus terms. One is a narrow list of five verbs: *go, move, walk, advance*, and *proceed.* The other was a broader list of 211 verbs. The result are given in Fig. 3:

1. Narrow list — 5 genus terms	
called	594
chosen	469
taxonomy	1320
2. Broad list — 211 genus terms	
called	1131
chosen	802
taxonomy	1301

Figure 3: Intransitive Verb Senses in W7
(total number of senses = 8299)

Comments on Preliminary Results. As expected, the narrow list gave a higher percentage of correct choices, but at the same time many senses were missed altogether. The broader list gave more spurious choices, mostly due to the problem of polysemy. The most serious problem, however, resulted from the use of a verb with a preposition which can change the basic verb into a verb of movement. For example, the verb *hobble* is a verb of movement, whether or not one *hobbles in* or *hobbles out*. In contrast, a verb like *rattle* must be used with a particle to be a verb of movement. The definitions for the intransitive verb senses of the relevant homonym of *rattle* from W7 are:

1. RATTLE (vi) to make a rapid succession of short sharp noises
2. RATLLE (vi) to chatter incessantly and aimlessly
3. RATTLE (vi) a. to move with a clatter or rattle
   b. to have room to move about aimlessly

Looking at the genus terms alone, it would seem that the third definition is the relevant movement sense. However, no movement occurs without a particle, as is captured by the presence of the preposition *about* in sense 3a, for *rattle about.* Consider two citations from the corpus referred to earlier:

1. The moment a chaise was heard rattling over the courtyard cobblestones, Franklin rushed out.
2. Machine guns rattled, and there were dull explosions everywhere.

In the first example, the chaise is definitely moving, related to sense 3 in W7, whereas in the second example, the machine guns are making noise but (probably) not moving, which is captured by sense 1 in W7.

We will be running different lists in the future to determine the optimum size and type. We can then try our methods for other verb types, and on other dictionaries. The goal is to extract senses from different dictionaries, extract features of the verb types from the genus, differentia, synonyms, and information in parentheticals. Once identified, the structured feature clusters associated with appropriate verb senses can be encoded in a knowledge representation formalism utilizing property inheritance and automatic classification of concepts. The result will be semantic network explicitly representing the factoring out of distinctive properties of verb sub-classes as represented in MRD's. We suspect that there will be gaps in our representation, but that by using multiple resources, we have the opportunity to test the hypothesis that MRD's are important resources for the automatic extraction of structured semantic knowledge.

Concluding Remarks. There is a growing number of projects attempting to exploit the information in machine-readable resources. Among them are Michiels (1982), Alshawi (1985), Boguraev (1987), Byrd *et al.* (1987), Fox *et al.* (1988), Calzolari (1983) Calzolari and Picchi (1988), and Wilks (1988). Most natural language systems have been hand-building their lexicons, but it is becoming increasingly clear that broad-coverage is an important goal. Thus, the task of extracting information from already existing sources is an important research area. There is a great need in the computational linguistics community for better and more complete machine-readable dictionary resources aimed at both people and programs. There is much valuable information to be exploited in these resources for the task of automatically creating a wide-coverage computational lexicon, which can then provide natural language systems with needed linguistic information.

Notes

[1] From now on, the term *features* is used to apply to both features and attribute-value pairs in UDICT.

[2] The abbreviations are: SING = singular, AXNT = accent on final syllable, FACTIVE = factive, TOV = takes an infinitival complement, STORED = stored, i.e. not derived from morphological analysis, TRAN = transitive, PRES = present, INF = infinitival form, THATCOMP = takes a that complement, HUMSJ = takes a simple human subject, COLLHUMSJ = takes a collective human subject (such a class,army, group), and HUMEXPSJ = takes a human expression subject (such as film, article,book).

[3] We acknowledge the valuable input of Beryl T. (Sue) Atkins, senior editor of the Collins Robert French-English bilingual dictionary, who was visiting the Lexical Systems Group at IBM during April, 1988. We also acknowledge input from Beth Levin.

[4] The work reported on verb types, and in particular on verbs of manner of movement, was initialed jointly by Sue Atkins and this author. It will be reported in full in a later joint publication.

References

Cited Dictionaries

COLLINS COBUILD ENGLISH LANGUAGE DICTIONARY (COBUILD). 1987. John Sinclair et al. (eds.). London and Glasgow: Collins.

COLLINS GERMAN-ENGLISH/ENGLISH-GERMAN DICTIONARY. 1980. Peter Terrell et al. (eds.). London and Glasgow: Collins.

COLLINS ROBERT FRENCH-ENGLISH/ENGLISH-FRENCH DICTIONARY. 1978, 1987^2. Beryl T. Atkins et al. (eds.). London and Glasgow: Collins, Paris: Dictionnaires Le Robert.

COLLINS SANSONI ITALIAN DICTIONARY: ITALIAN-ENGLISH/ENGLISH-ITALIAN. 1980. London and Glasgow: Collins

COLLINS SPANISH DICTIONARY: SPANISH-ENGLISH/ENGLISH-SPANISH. 1988^2. Colin Smith et al. (eds.). London and Glasgow: Collins, Barcelona: Grijallo.

THE NEW COLLINS THESAURUS. 1984. William T. McLeod, managing editor. Glasgow: Collins.

LONGMAN DICTIONARY OF CONTEMPORARY ENGLISH (LDOCE). 1978. Paul Procter et al. (eds.). Harlow and London: Longman.

ROGET THESAURUS: THESAURUS OF ENGLISH WORDS AND PHRASES. 1982. Peter M. Roget (ed.). Revision ed. Susan M. Lloyd. Harlow and London: Longman.

WEBSTER'S SEVENTH NEW COLLEGIATE DICTIONARY (W7). 1963. Springfield, MA: Merriam.

Other Literature

Allerton, D.J. 1982. *Valency and the English Verb.* London: Academic Press.

Alshawi, H. 1985. 'Processing Dictionary Definitions with Phrasal Pattern Hierarchies'. Unpublished paper. Cambridge, England: University of Cambridge Computer Laboratory.

Atkins, Beryl T. 1987. 'Semantic ID tags: Corpus Evidence for Dictionary Senses' in *The Uses of Large Text Databases,* Waterloo, Canada: *Proceedings of the Third Annual Conference of the UW Centre for the New Oxford English Dictionary,* University of Waterloo.

Atkins, Beryl T., Judy Kegl, and Beth Levin. 1986. 'Explicit and Implicit Information in Dictionaries' in *Proceedings of the Second Annual Conference of the University of Waterloo Centre for the New Oxford English Dictionary: Advances in Lexicology.* Waterloo, Canada: University of Waterloo.

Atkins, Beryl T., Judy Kegl, and Beth Levin. 1988. 'Anatomy of a Verb Entry: From Linguistic Theory to Lexicographic Practice' in *International Journal of Lexicography* 1:84—126.

Boguraev, Branimir. 1987. 'Experiences with a Machine-Readable Dictionary' in *Proceedings of the Third Annual Conference of the UW Centre for the New Oxford English Dictionary: The Uses of Large Text Databases.* Waterloo, Canada: University of Waterloo.

Boguraev, Branimir. (to appear) 'Machine-Readable Dictionaries and Research in Computational Linguistics' in D. Walker, A. Zampolli, N. Calzolari (eds.). *Automating the Lexicon — Research and Practice in a Multilingual Environment.* Cambridge, England: Cambridge University Press.

Byrd, R. J. 1983. 'Word formation in natural language processing systems' in *Proceedings of the International Joint Conference on Artificial Intelligence* 8:704—706.

Byrd, R. J. 1984. 'The Ultimate Dictionary Users Guide'. Unpublished paper. Yorktown Heights, New York: IBM.

Byrd, R. J., J. L. Klavans, M. Aronoff, and F. Anshen. 1986. 'Computer Methods for Morphological Analysis' in *Proceedings of the Association for Computational Linguistics.* 120—127.

Byrd, Roy J., Nicoletta Calzolari, Martin S. Chodorow, Judith L. Klavans, Mary S. Neff, and Omneya A. Rizk. 1987. 'Tools and Methods for Computational Lexicology' in *Computational Linguistics* 13 (3—4): 219—240.

Calzolari, N. 1983. 'Semantic Links and the Dictionary' in S.K. Burton, and D. D. Short (eds.)., *Proceedings of the Sixth International Conference on Computers and the Humanities*. Rockville, Maryland: Computer Science Press. 47—50.

Calzolari, N., and E. Picchi. 1988. 'Acquisition of Semantic Information from an On-Line Dictionary' in *Proceedings of the Twelfth International Conference on Computational Linguistics,* Budapest, Hungary. Association for Computational Linguistics: Morristown, New Yersey. 87—92.

Chodorow, M. S., R. J. Byrd, and G. E. Heidorn. 1985. 'Extracting Semantic Hierarchies from a Large On-Line Dictionary' in *Proceedings of the Association for Computational Linguistics*. 299—304.

Fox, E., J. T. Nutter, T. Alswhede, and M. Evans. 1988. 'Building a Large Thesaurus for Information Retrieval' in *Proceedings of the Second Conference on Applied Natural Language Processing*. Austin, Texas.

Gross, Maurice. 1975. *Méthodes en syntaxe: Régimes des Constructions Complétives*. Paris: Hermann Publishers.

Ingria, Robert. (to appear), 'Lexical Information for Parsing Systems: Points of Convergence and Divergence' in D. Walker, A. Zampolli, and N. Calzolari (eds.). *Automating the Lexicon — Research and Practice in a Multilingual Environment*. Cambridge, England: Cambridge University Press.

Jackendoff, Ray. 1987. 'The Status of Thematic Relations in Linguistic Theory' in *Linguistic Inquiry* 18: (3): 369—411.

Katz, Boris and Beth Levin 1988. 'Exploring Lexical Regularities in Designing Natural Language Systems' in *Proceedings of the Twelfth International Conference on Computational Linguistics,* Budapest, Hungary. 316—323.

Klavans, Judith L. 1988. 'COMPLEX: a computational lexicon for Natural Language Systems' in *Proceedings of the Twelfth International Conference on Computational Linguistics,* Budapest, Hungary. 815—823.

Klavans, Judith L. and Nina Wacholder. 1988. *Features and Attributes in the UDICT Lexicon*. IBM Internal Research Report □ 142451.

Levin, Beth. (to appear) 'The Representation of Semantic Information in the Lexicon' in D.Walker, A. Zampolli, and N. Calzolari (eds.). *Automating the Lexicon — Research and Practice in a Multilingual Environment*. Cambridge, England: Cambridge University Press.

Michiels, Archibald. 1982. *Exploiting a Large Dictionary Data Base*. PhD Dissertation. Liège University of Liège.

Wilks, Y., D. Fass, C-M Guo, J. E. McDonald, T. Plate, and B. M. Slater. 1988. 'Machine Tractable Dictionaries as Tools and Resources for NL Processing' in *Proceedings of the Twelfth International Conference on Computational Linguistics*. Budapest, Hungary. 750—755.

Appendix One — Structured entries

```
might
    (POS(VERB)
    (MORPH              (INFLECTION         (INFORM may)))
                                            (LEMMA may)
    (STYLISTIC          HDG))
    (SYNTACTIC          (INHERENT           (AUX (MODAL))))
                                            (IRREG)
                        (TENSE              (PAST))
    (SYSTEM             (STORED))
princess
    (POS(NOUN)
    (SEMANTIC           (INHERENT           (ANIM))
                                            (FEMALE)
                                            (HUM)
    (SYNTACTIC          (NUMBER             (SING)))
    (SYSTEM             (STORED))
red
    (POS(ADJ))
    (PHON(AXNT))
    (SEMANTIC(INHERENT(COLOR)))
    (SYSTEM(STORED))

    (POS(NOUN))
    (PHON(AXNT))
    (SYNTACTIC(NUMBER(SING)))
    (SYSTEM(STORED))
florin
    (POS(NOUN)
    (SEMANTIC(INHERENT(UNIT(CURRENCY))))
    (SYNTACTIC(NUMBER(SING)))
    (SYSTEM(STORED))
```

Notes on Appendix One

These examples illustrate other features and other attribute-value pairs including the STYLISTIC feature HDG, *hedge*, the syntactic structure of the modals and auxiliaries, and the inherent semantic features ANIMate, FEMALE, and HUMan, COLOR, UNIT, and CURRENCY. These features come from many sources. For example, the FEMALE feature for nouns came partly from LDOCE codes, partly from semantic analysis of the definitions in W7, and partly from our morphological analysis of nouns likely to be female (e.g. ending in -ess). See Klavans and Wacholder (1988) for more detail on features, attributes, and hierarchical structure in the current version of the computational lexicon.

Appendix Two — "Mangle" from four sources

- **Webster7**

mangle 1(vt)
mangling
DEFINITIONS:
1 to cut, bruise, or hack with repeated blows or strokes
2 to spoil or injure in making or performing mangler (n)

mangle 3(vt)
mangling
DEFINITIONS:
to press or smooth (as damp linen) with a mangle

- **Longman**

O: T1 often pass.; ; Subj: Moveable solid; DO:
Human
to tear or cut to pieces; crush: After the accident they tried to find out who the people were, but the bodies were too badly mangled to be recognized
mangle {mangle} (v) /"m NgF1/

O: T1; household; Subj: Human; DO: Moveable solid
to put (wet clothes, sheets, etc.) through a MANGLE or WRINGER

- **New Collins Thesaurus**

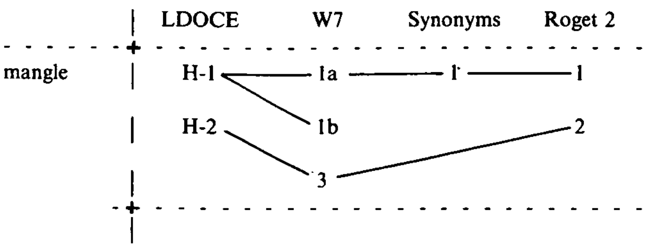
O. butcher, cripple, crush, cut, deform,
destroy, disfigure, distort, hack, lacerate, maim, mar, maul, mutilate, rend, ruin, spoil, tear, wreck<

- **Roget2 (from Gunther)**

M mangle
P v
N 1
D To injure or damage, as by abuse or heavy wear.
S batter/1, knock about/1, knock around/1, maul, rough up
N 2
D To smooth by applying heat and pressure.
S iron, press/1

— The Serse Grid

- **Grid to indicate which senses correspond across sources.**



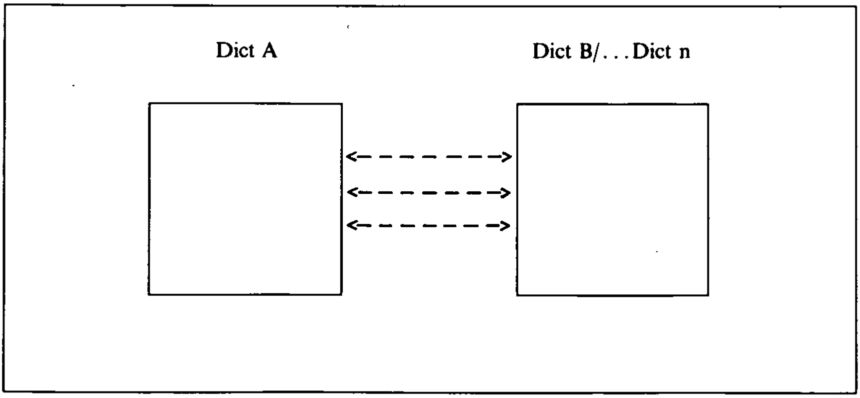
The senses of *mangle* as shown in the preceding definitions map onto each other as above. However, instead of mapping, senses can be analyzed and then represented in a data structure which is not bound to the sense distinctions of any given dictionary. This position is schematized in Appendix Three. An example of a possible data structure for verbs might be:

- **Expanded Sense Grid for Verbs**

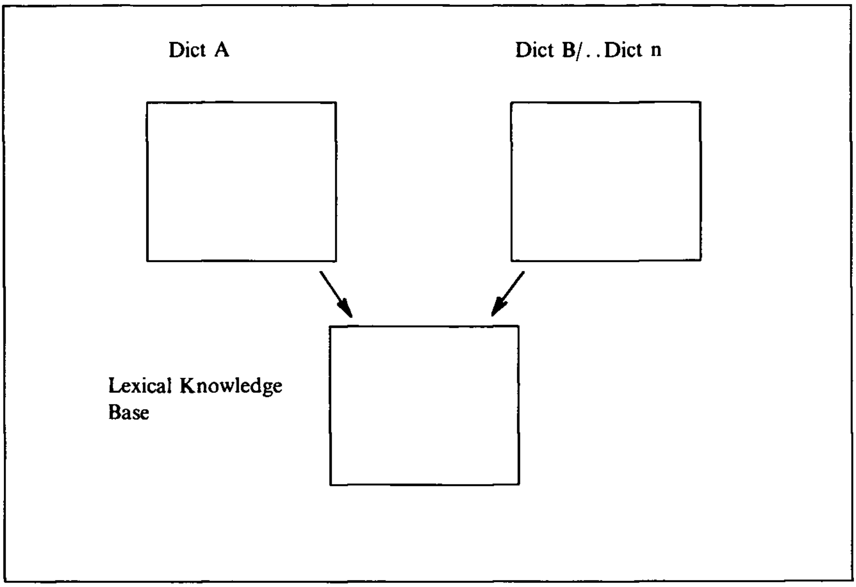
	genus	sbj	obj	mnr	rson	instr	meth	purp	. . .
Dct Sns #									
Dct Sns #									
Dct Sns #									

Appendix Three

Mapping Between Dictionaries:



Mapping From Dictionaries into a Lexical Knowledge Base:



This diagram shows just two dictionaries, Dictionary A and B, but there is no limit on the number of dictionaries that can be either mapped onto each other, or tapped for information for a Lexical Knowledge Base. Furthermore, with the Lexical Knowledge Base, information from any source, structured or unstructured, can be used.