# COMPLEX: A Computational Lexicon for Natural Language Systems

**Judith KLAVANS**
**IBM Thomas J. Watson Research Center**
**P. O. Box 704**
**Yorktown Heights, NY 10598**
**USA**

## Abstract

Although every natural language system needs a computational lexicon, each system puts different amounts and types of information into its lexicon according to its individual needs. However, some of the information needed across systems is shared or "identical" information. This paper presents our experience in planning and building COMPLEX, a computational lexicon designed to be a repository of shared lexical information for use by Natural Language Processing (NLP) systems. We have drawn primarily on explicit and implicit information from machine-readable dictionaries (MRD's) to create a broad coverage lexicon.

## 1. The Computational Meta-Lexicon

There is growing awareness among computational linguists that much of the information needed for lexical entries across systems is basically shared or "identical" information /Ingria 1986, Zaenen 1986/. An example for verbs is subcategorization information (transitive, intransitive, takes a that-complement), and selectional features (takes a human object, selects for inanimate subject); an example for nouns is gender (female, male). It should be possible for much of this shared information to be collected into a large "polytheoretical" data base for use by individual systems. This lexicon (sometimes called a "meta-lexicon") would consist of the overlapping set of the various attributes, features, characteristics, etc., that are needed by all or most NLP systems. Each system could then consult the repository of information stored in the central lexicon and extract the information it might need. The extracted information could be enhanced by theory-specific and application-specific information. Thus, instead of each system duplicating efforts, the computational "meta-lexicon" gathers together lexical information for use by programs, in the same way that traditional dictionaries contain information for use by people.

One of the goals of the Lexical Systems project at IBM is to design and build such a lexicon. We have called the system COMPLEX (for COMPutational LEXicon). Although this is an ambitious goal, we believe that careful lexicographic, linguistic, and computational research will permit us to represent whatever information is common to most NLP systems in a neutral representation and in a uniform data structure so as to be compatible with a range of requirements of natural language systems.

Corollary to the goal of designing and building a data structure containing information for different NLP systems is the goal of broad coverage. Indeed, until recently, the lexicon was not the primary focus of most natural language processing (NLP) projects. The result (with a few exceptions) has been a proliferation of descriptively rich syntactic and semantic analyzers with impoverished lexical coverage. Many NLP systems have small hand-built lexicons, hand-tailored to the idiosyncrasies of formatting and processing required by the system. Our aim is to extract information automatically or semi-automatically using machine-readable sources, and in this way to achieve broad coverage. Currently, our primary resources are machine readable dictionaries although we have plans to expand to text corpora in the near future. Initially, we restrict our attention to building English lexicons but there is good evidence that some information may be transferable to computational lexicons for other languages via bilingual dictionaries.

## 2. Applications

The initial impetus for building a computational lexicon arose from the needs of the CRITIQUE text-critiquing system (previously called EPISTLE, Heidorn et al. 1982). Basic syntactic information such as part of speech, subcategorization for verbs (e.g. trans, intrans, complement taking properties), irregular forms, some inherent semantic information (such as male, female for nouns), some graphemic, phonological, and stylistic

features were gathered from a range of (primarily) machine-readable sources. This system (called UDICT, the ultimate dictionary) is described in Byrd 1983 and Byrd et al. 1986. A modified version of the original dictionary is still in use by that project.

Our experience in attempting to build a solid broad-coverage computational lexicon revealed to us the range of projects potentially in need of such a lexical resource. Unfortunately, it also revealed to us a range of problems. First, the projects: we received requests for information from NLP projects such as the experimental English-to-German machine translation system LMT /McCord 1988/, the natural language data base query project TQA /Damerau et al. 1982, Johnson 1984/, the kind-types Knowledge Representation system KT /Dahlgren and McDowell 1986/, and others. In fact, the LMT system uses UDICT for lexicon back-up when the LMT lexicon does not contain or does not analyze an item /McCord and Wolff 1987/. The analyses output from UDICT are compiled into LMT internal format for use by LMT. This is exactly the use we envision for COMPLEX.

In addition to use by NLP systems, some of the information in COMPLEX might be used directly by lexicographers to aid in creating lexicographers' workstations for projects such as dictionary building and machine-assisted translation. It could also be useful to psycholinguists seeking lists of words with particular lexical properties for test materials. /Taft and Forster 1976, Cutler 1983/. Since COMPLEX is machine readable, it is a simple matter to extract lists with selected features.

Some of the problems that arose as a result of our experience in attempting to build and provide a solid broad-coverage computational lexicon for NLP projects are discussed in the next section. Most important is the problem of polysemy. We realized that 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, also discussed below.

## 3. The Polysemy Problem and The Mapping Problem.

Each entry in UDICT consists of lists of features and attribute-value pairs. There is one list for each part of speech. For example, the word "claim" has two parts of speech in UDICT:

1. claim(NOUN SING AXNT FACTVE TOV STORED (STRUCTURE < * > N))

2. claim(VERB PLUR TRAN AXNT PRES INF THATCOMP STORED HUMSJ COLLHUMSJ HUMEXPSJ (STRUCTURE < * > V))

In this case, "claim" is morphologically simple so the STRUCTURE value is the same as the input word.

The polysemy problem arises because of the fact that there is only one list of features[1] permitted for each part of speech. 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 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?

Consider a system such as that reported in Boguraev 1986 and 1987 in which sense distinctions are in fact made. The grammar development system, intended for a GPSG-style parser, utilizes the grammatical codes in the Longman Dictionary of Contemporary English /1978/, henceforth LDOCE, as the basis for 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. 1987 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 NLP research, and in particular in the use of machine readable dictionaries. However, each is constrained by the sense distinctions dictated by LDOCE. LDOCE is a small dictionary, so there are many distinctions omitted. Furthermore, often important grammatical distinctions are merged for the sake of

---

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

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":

```
button (v)

T1;I0; clothing; Subj: Human;
   DO: 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 sometimes captured implicitly in the definition in the form of the parenthesized "(cause to)", but this is in no way explicit in the dictionary resource. A detailed study of the semantic codes for subject from LDOCE is given below.

To sum, there are various solutions to the problem of senses, each of them inadequate in one way or another. The solution to list only the intersection of features (the approach in most of UDICT) or the solution to list the union of features (taken for the 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. 1987 or by Boguraev 1986 and 1987 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. Alternatively, one could 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,[2] and it is the route we are currently pursuing although our results are still too preliminary to be reported.

## 4. COMPLEX - The Lexical Systems Lexicon

### 4.1. COMPLEX Structure

The previous sections of this paper have described the limitations of UDICT. With this in mind, this section gives the information to be contained in COMPLEX. Currently, we draw on the following sources:[3]

1. enhanced UDICT (LexSys)
2. Brandeis Verb Lexicon
3. definitions and grammatical information from LDOCE /Longman 1978/,

We have plans to use information from:

1. definitions, synonyms, and etymologies from Webster's Seventh /Merriam 1963/,
2. taxonomy files created from Webster's Seventh /Merriam 1963/ using techniques reported in Chodorow et al. 1985,
3. synonyms from the Collins Thesaurus /Collins 1984/,
4. Collins bilingual dictionaries for English/Italian, English/French, English/Spanish, and English/German
5. text corpora

We too are using the sense distinctions on LDOCE, although we are aware of its limitations. (See also Michiels 1982). Our system is not hard-wired into LDOCE. Consider the design for one sense of the verb "bring":

**--Lexical Systems Analysis**

```
(MORPH(INFLECTION(PAST brought)))
                  (PASTPART brought)))
(PHON(AXNT))
(SYNTACTIC(CONSTRUCTION (MWESTART)))
          (INHERENT (INF)))
                    (IRREG)))
          (NUMBER (PLUR)))
          (SUBCAT (DITRAN)))
                  (NPING)))
                  (NPTOV)))
                  (TRAN)))
          (TENSE (PRES)))
(SYSTEM(STORED))
```

**--Brandeis Verb Lexicon**

```
DO DO-P10-NP IO-DO DO-TONP
```

---

2 We acknowledge the valuable input of Beryl T. (Sue) Atkins, who was visiting the Lexical Systems Group at IBM during April, 1988. We also acknowledge input from Beth Levin.

3 The Brandeis Verb Lexicon was developed by Jane Grimshaw and Ray Jackendoff's, NSF grant number NSF IST-81-20403 awarded to Brandeis University.

--LDOCE

```
:SENSENUM. 1
:SGRAMCODES. D1 ( to, for );T1
:SUBJCODES. NONE
:SEL_RES_SUBJ. NONE
:SEL_RES_DO. NONE
:SEL_RES_IO. NONE
:DEF. to come with or lead:
```

Note that there are three distinct data sets. Each of these structures will be described in turn.

### 4.2 Lexical Systems.

In the example above, the Lexical Systems data show four feature types: two MORPHological, one PHONological, nine SYNTACTIC and one SYSTEM feature. Other feature types not shown in this analysis are SEMANTIC, STYLISTIC, and GRAPHEMIC. The two morphological features (MORPH) give the irregular inflectional attribute-value pairs for the past and past participial forms of the verb (PAST brought) and (PASTPART brought). The next feature is phonological (PHON); AXNT means that the word is accented on the final syllable. In the case of "bring" the word is monosyllabic, but in a word like "persuade" the AXNT feature distinguishes word initial from word final stress. This phonological feature is needed for some morphological rules in English. The next nine features are syntactic: "bring" can start multi-word constructions such as "bring about" (MWESTART); it is an infinitival form (INF), and it is inherently irregular IRREG; its number is PLUR; it subcategorizes as a di-transitive DITRAN (i.e. it takes two objects), takes an NPING and NPTOV complement, and that it is a transitive verb; its tense is PRES. The SYSTEM feature STORED shows that the word is stored in our database rather than resulting from analysis by our affixation and compounding rules.

The data structure displayed under the Lexical Systems Analysis (LexSys) is based on UDICT. As shown in the example above for "claim", UDICT data is an *unstructured* list of features and attribute-value pairs, This output is then *structured* into a feature hierarchy according to feature type. There are six categories at the top level: SYNTACTIC, PHONological, MORPHological, SEMANTIC, STYLISTIC, and GRAPHEMIC. Features are then listed under part of speech for each category, and there are up to five levels of depth. This has important implications for feature addition, since the system needs to forbid occurrence of certain features under certain nodes. For example, THATCOMP cannot apply to determiners in English or MALE cannot be an inherent property of verbs in English, although a verb could have the contextual property of selecting for MALE arguments. The arrangement of the data in a structure also permits efficient querying. Thus, if an application requires only one type of feature, such as phonological or syntactic, this feature set is easily extracted from the larger data structure.

### 4.3 Brandeis Codes for "bring"

The Brandeis Codes subcategorize "bring" for direct object (DO). Furthermore, if the verb takes a DO with the preposition "to" (P10), then it also takes an NP. If an indirect object is present (IO), then so is a DO. Finally, "bring" will take a DO following by an indirect object introduced by "to"; this code is not intended to apply to other uses of "to".

Observe that, like the features for UDICT, Brandeis Codes represent the intersection of subcategorization properties of verbs. There are about 900 verbs, 28 features, and 19 prepositions or preposition types. The codes characterize some inherent features (such as "Modal"), control properties, and contextual features (such as ACCING "accusative followed by -ing phrase). Cases where combinations of features are required are indicated in the codes.

Note also that there is some overlap of information between the Lexical Systems analysis and the Brandeis analysis, such as SUBCAT(TRAN) and DO. This is a clear example of identical information in different systems. By gathering together different computational lexicons into one general repository, we can both eliminate duplication when two systems overlap, and increase coverage when they differ. Of course, we will also need methods for resolving disagreements when they arise.

### 4.4 LDOCE

The LDOCE data first gives the headword and part of speech; these two values hold for each subsequent sense. Then entries are broken into sense numbers. In this example, sense one has the grammatical codes of "D1" (ditransitive verb used with two or more objects) and "T1" (transitive with one object when used with the prepositions "to" and "for"). There is no subject area, (such as "medicine", "mathematics", "law"), nor are there any selectional restrictions. Next follows the definition and example sentences, which are included for the purpose of helping the human user. They are not relevant to a computational lexicon except as a potential source of implicit information. (See Atkins et al. 1986).

Questions were put to us concerning the accuracy and completeness of the LDOCE codes. We decided to undertake an in-depth study of selectional restrictions for subject to get some concrete data on how precise and thorough the LDOCE codes really are. This study is described in the next section.

## 5. Evaluating the Semantic Codes in LDOCE

### 5.1 Methodology

Selectional restrictions for verbs specify that argument(s) of that verb must have particular semantic properties, as opposed to subcategorization information which simply tells whether the verb can take a certain number of arguments, or can occur in a certain syntactic context. Our position on selectional restrictions is close to that of Jackendoff 1987: "...a selectional restriction ... is part of the verb's meaning and should be fully integrated into the verb's argument structure." (p.285) Although our computational lexicon is far more surface-structure oriented than that required by Jackendoff, the spirit of the claim still applies. We do not yet have a distinct level of Lexical Conceptual Structure /Jackendoff 1983, Levin, to appear/.

Selectional restrictions can be as peculiar and varied as the entire conceptual and semantic system of a language. For this reason, we picked "subject" because all sentences require subjects at some level; we picked "human" because all systems seem to agree on the need for this feature.

The machine-readable form of LDOCE is enhanced with a set of codes called "Box Codes". There are ten fields of information in the Box Codes giving such information as register (e.g. informal), or dialect (e.g. Scottish). For verbs, three of the fields give semantic selectional restrictions on the arguments subject, object, and indirect object.

To illustrate, the following are the two lines of codes from LDOCE for the entry "admire"; there is one line for each sense in the dictionary entry.

```
admire    <....H....Z<v<Wv4;T1>(for)<<
admire    <....H....Z<v<Wv4;T1<<
```

The subcategorization information in these codes, such as "T1" for "verb followed by NP" or "Wv4" meaning "occurs in the gerundive for the adjectival form", is what Boguraev 1986 has used in converting information from LDOCE to more traditional subcategorization formats. In addition to the grammatical codes, there are ten fields for further information. These fields are shown between the first two '<' signs in the previous figure. Each field has letter code, or a '.' for no code. For verbs, field five gives selectional restrictions on subject, field ten on direct object, and field eight on indirect object. In the example above, "H" is "Human", and "Z" is "Unmarked (no semantic restriction)." The box codes are only available in the machine-readable version of the dictionary.

In order to extract a list of verbs from LDOCE that was truly likely to require human subjects in all senses, a constraint was imposed. Only those verbs that are marked with an "H" in position five for *all senses* were considered. This technique yielded a list of 2323 candidate verbs.

Each of the verbs was subjected to six tests reflecting observations about what could count as a human subject, and observations about syntactic variations. Test one was for collective human nouns such as "chorus", "class". Test two was for human actions; this applies to machines such as "robot" or "computer" which are not necessarily humanoid but easily anthropomorphosized. Frame three tested human-expression nouns such as "film", "article", in which case the noun usually refers to the person behind the work. The next test checks to see if a singular human subject is required. The fifth test is to check for cases like "button" where human applies to agent role, but the theme or object can still appear in surface subject position. Finally, we observed that many of the verbs LDOCE claims select for human subject actually take any animate subject. This is particularly applicable to biologically based activities, such as "gag". To sum,

```
a. Collective noun subject
b. Human-action subject
c. "Human expression" subject
d. Obligatorily singular
e. Causative/inchoative alternation
f. Animate subject
```

Simple substitution frame tests were constructed to insert a properly inflected version of the verb to be tested into a set of six representative sentence frames. Judgments on grammaticality were stored in a matrix:

```
                    a b c d e f
admire.............. + - - + - -
caricature.......... + - + + - -
gag................. + - - + - -
```

Features reflecting the judgment patterns were then generated from this matrix. Only the relevant features are shown in the following figure:[4]

1. admire : (VERB +HUMSJ +COLLHUMSJ)
2. caricature: (VERB +HUMSJ +COLLHUMSJ +HUMEXPSJ)
3. gag: (VERB +HUMSJ +COLLHUMSJ +ANIMSJ)

### 5.2 Results

The next figure summarize the results of the judgments on these verbs.[5]

| | LDOCE |
|---|---|
| Broad | 59% |
| Narrow | 36% |
| Animate | 13% |
| Rejects | 28% |
| | n=2323 |

Broad = Human, Human Collective, Human Expression, Human Action
Narrow = Human, Human Collective

We were disappointed that only 59% of these verbs required human subjects in all senses. There may be several reasons for this: LDOCE is small so many senses are omitted; the "button" type verbs were listed as requiring human subjects; and verbs requiring animate subjects were listed as requiring human subjects. We suspect that what may have happened is that the question was asked of these verbs "is a human subject *possible*?" rather than "is a human subject *necessary*?"

This data shows that the Box Codes in LDOCE have to be carefully re-evaluated before they can be used. However, it is not our intention to witch-hunt in LDOCE. The dictionary is immensely useful, particularly since it is based on a detailed and thorough grammatical analysis of English /Quirk and Greenbaum 1972/. Rather, our goal is to utilize what LDOCE has to offer. With this more positive goal in mind, we took the original list of 2323 verbs which select for human subject from LDOCE, and used it as the basis to explore whether the list could be expanded using other tools we have developed. Although our results are limited, it must be remembered that without the LDOCE box codes, we would have had no seed list.

## 6. Using Semantic Codes in LDOCE

### 6.1 Methodology

Our goal in the second study was to use the LDOCE list of 2323 verbs said to select for human subject as the basis to discover other verbs which select for human subject. We compared the results above with two other methodologies:

1. Implicit Information in Concept Hierarchies (Taxonymies)
2. Implicit Information from Morphological Clues

The first technique used clues from the definitions themselves. Chodorow, Byrd, and Heidorn 1985 devised a methodology to construct taxonymic hierarchies based on the genus terms of definitions in Merriam 1963. Two procedures are based on the taxonym files: Sprouting and Filtering. In brief, Sprouting uses concept hierarchies to add words of a semantically related field to a seed list. Filtering is a method to enlarge a list of words in terms of the heads of their definitions.[6] We had used Filtering in the past to augment lists of nouns with a given *inherent* feature, such as HUMAN. However, we had never before tried to filter with a list of verbs with a given selectional feature. If our results were good, we would have some proof of the hypothesis that genus terms reflect certain properties of their corresponding headwords. More specifically, we would have evidence that selectional restrictions may be inherited from hypernyms just as are inherent features. Our results show that this hypothesis is correct.

Using the 2323 verbs from LDOCE we ran Filter on our taxonym files, and extracted 312 candidate human subject verbs. Each of these verbs

---

4 Obviously, more detail would be needed to capture the fact that a verb like "gag" requires an animate subject with a wind-pipe. Can a virus gag?

5 There is some degree of error throughout these judgments. What is needed is a large number of different people giving such judgments. However, I will assume that the errors are equally distributed throughout the data, and thus can be assumed for now to be neutralized. What we have at the very least is a complete and thorough account of at least one person's ideolectical intuitions on human subject verbs.

6 Filtering was used in later parts of the procedure when we started with small seed lists to be used to label nouns which were human_collective, human_expressions, etc. We did not use filtering for verbs.

was subjected to the same six tests, and matrices of properties were constructed. Interestingly, since Merriam 1963 has more headwords than LDOCE, many of the verbs we obtained from Filtering were quite esoteric. These verbs are also less polysemous, probably as a result of their being less common.

For comparison and curiosity, we also tried using a more risky method. Morphology is often a clue to semantic features, both of the base and of the derived word. Under the assumption that the nominalizing *-er* suffix in English sometimes marks agentivity, and in order to test the hypothesis that the verbal bases of these agentive nouns might have a tendency to select for human subjects, we extracted about 4000 nouns ending in -er from a large (100,000+) word list. Then we sent these nouns through our morphological analyzer to extract those with verb bases. Of the over 1000 nouns which had verb bases, 712 were not already on the LDOCE list augmented by Filtering. These verbs were added to the candidate list of possible verbs selecting for human subjects. Although we knew that using the multiply ambiguous *-er* suffix was more speculative, we decided to follow through with our experiment so we could get a measure of how useful the technique is.

### 6.2 Results

The next figure summarize the results of the judgments on these verbs in comparison with the previous results:

| | LDOCE | FILTER | -er |
|---|---|---|---|
| Broad | 59% | 45% | 20% |
| Narrow | 36% | 25% | 8% |
| Animate | 13% | 14% | 21% |
| Rejects | 28% | 41% | 59% |
| | n=2323 | n=316 | n=712 |

Broad = Human, Human Collective, Human Expression, Human Action
Narrow = Human, Human Collective

Not surprisingly, the best source of verbs which select for Human Subject in all senses was LDOCE. However, remember that the candidate verbs were supposed to select for Human Subject in all senses, yet only 59% of these verbs really conformed to that requirement. The next result concerns Filtering. Nearly half of the verbs proposed by Filtering were acceptable. This gives some interesting insights into the internal organization of Merriam 1983. It shows that genus terms reflect certain properties of their corresponding headwords. More specifically, there is some evidence that selectional restrictions may be inherited from hypernyms just as are inherent features. A result like this would be greatly strengthened if sense distinctions were made, rather than requiring that the restriction apply to all senses. Finally, not surprisingly, the morphological method gave the worst results. Only 20% of the candidate verbs fulfilled four of the six tests.

## 7. Future

There are many other ways to tap machine-readable sources that we would like to try. Concerning subjects, we would like to extract data from text corpora to confirm (or refute) our intuitions on the verbs we tested. We would also like to use example sentences to verify hypotheses about lexical features. As shown above "button", example sentences often contradict claims in the Box Codes. Information about verbs, such as "button", which permit an underlying object to appear as subject might be implicit in LDOCE. We are working to develop a mechanism to enable mitosis when sense division is motivated either by semantic or syntactic facts. We are also exploring mechanisms to use several dictionaries to get maximum coverage. We are working on a practical solution to the mapping problem (see Byrd et al. 1987).

The COMPLEX system has been implemented and incorporated into the WordSmith on-line dictionary system, described in Neff and Byrd 1988, which allows flexible access to dictionaries stored as DAM[7] files and lexical data bases. Ultimately, COMPLEX structures will be placed in a Lexical Data Base so they can be queried by the Lexical Query Language /Neff et al. 1988/. We intend to expand our data structures as we incorporate more and different information into our lexical repository. The goal is to create a rich computational lexicon that can be utilized by NLP systems. We are working intensively on a practical solution to both the polysemy problem and to the mapping problem as they apply to the construction of COMPLEX.

[7] DAM ("Dictionary Access Method") is an access method subsystem which gives programs fast and convenient access to large files of information associated with sets of keys.

## Bibliography


Atkins, Beryl T. (1987) "Semantic ID tags: Corpus Evidence for Dictionary Senses", in *The Uses of Large Text Databases*, Proceedings of the Third Annual Conference of the UW Centre for the New Oxford English Dictionary, University of Waterloo: Waterloo, Canada.

Atkins, Beryl S., Judy Kegl, Beth Levin (1986) "Explicit and Implicit Information in Dictionaries", in *Advances in Lexicology*. University of Waterloo: Waterloo, Canada.

Boguraev, Branimir (1986) *Machine-Readable Dictionaries and Research in Computational Linguistics*, Paper presented at the Workshop on Automating the Lexicon, Marina di Grosseto, Italy.

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

Byrd, R. J. (1983) "Word formation in natural language processing systems," Proceedings of IJCAI-VIII, 704-706.

Byrd, R. J., J. L. Klavans, M. Aronoff, and F. Anshen (1986) "Computer methods for morphological analysis," *Proceedings of the Association for Computational Linguistics*, 120-127.

Byrd, Roy J., Nicoletta Calzolari, Martin S. Chodorow, Judith L. Klavans, Mary S. Neff, Omneya A. Rizk (1987) "Tools and Methods for Computational Lexicology," *Computational Linguistics*.

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

Chodorow, Martin, Yael Ravin and Howard E. Sachar (1988) "Investigating the Synonymy Relation in a Sense-Disambiguated Thesaurus" *Proceedings of the 2nd Conference on Applied Natural Language Processing* Association for Computational Linguistics: Morristown, New Jersey.

Collins. 1984 *The New Collins Thesaurus*. Collins Publishers, Glasgow.

Cutler, A. (1983) Lexical Complexity and Sentence Processing, in G. B. Flores d'Arcais and R.J. Jarvella, eds. *The Process of Language Understanding*. Wiley: New York.

Dahlgren, K., and J. McDowell (1986) Kind Types in Knowledge Representation. *Proceedings of COLING86*. Bonn, Germany.

Damerau, F. J., S. R. Petrick, M. Pivovonsky, and W. J. Plath (1982) "Transformational Question-Answering (TQA) System", SIGART Newsletter No. 79:62-64.

Heidorn, G. E., K. Jensen, L. A. Miller, R. J. Byrd, and M. S. Chodorow (1982) "The EPISTLE Text-Critiquing System," *IBM Systems Journal*

Ingria, Robert (1986) "Lexical Information for Parsing Systems: Points of Convergence and Divergence", Paper presented at the Workshop on Automating the Lexicon, Marina di Grosseto, Italy.

Jackendoff, Ray. (1983) *Semantics and Cognition*, MIT Press: Cambridge, Massachusetts.

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

Johnson, David E. (1984) "Design of a Robust, Portable Natural Language Interface Grammar" IBM Research Report Number: RC 10867.

Levin, B., to appear, "Approaches to Lexical Semantics Representation," in D. Walker, A. Zampolli, N. Calzolari, eds., Automating the Lexicon, MIT Press, Cambridge, Massachusetts.

Longman (1978) *Longman Dictionary of Contemporary English*, Longman Group, London.

McCord, Michael C. and Susanne Wolff (1987) "The Lexicon and Morphology for LMT, A Prolog-Based MT System" IBM Research Report, Number: RC 13403

McCord, Michael C. (1988) "Design of LMT: A Prolog-Based Machine Translation System" IBM Report Number: RC 13536

Merriam (1963) *Websters Seventh New Collegiate Dictionary* G. & C. Merriam, Springfield, Massachusetts.

Michiels, Archibal (1982) *Exploiting a Large Dictionary Data Base*. PhD Dissertation. University of Liege, Liege, Holland.

Neff, M. S. and R. J. Byrd (1988) WordSmith Users Guide. IBM Research Report Number: RC 13411, T.J. Watson Research Center, Yorktown Heights, New York.

Neff, Mary S, Roy J. Byrd, and Omneya Rizk (1988) Creating and Querying Lexical Data Bases. *Proceding of the Applied Association of Computational Linguistics* Austin: Texas.

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech, and Jan Svartik. (1972) *A Grammar of Contemporary English.* Longman: Harlow and London, England.

Taft, Marcus and Kenneth I. Forster (1976) Lexical Storage and Retrieval of Polymorphemic and Polysyllabic Words. *Journal of Verbal Learning and Berbal Behavior.*15, pp. 607-620.

Wilks, Y, D. Fass, C-M Guo, J.E. McDonald, T. Plate, and B. M. Slater (1987) A Tractable Machine Dictionary as a Resource for Computational Semantics. Computing Research Laboratory, New Mexico State University, Las Cruces, New Mexico.

Zaenen, Annie (1986) Project Report on the Lexical Project. CSLI Monthly, Volume 1, Number 3.