Building the CODER Lexicon: The Collins English Dictionary and Its Adverb Definitions

Edward A. Fox
Robert C. Wohlwend
Phyllis R. Sheldon
Qi Fan Chen
Robert K. France

Technical Report 86-23

October 1986

Building the CODER Lexicon: The Collins English Dictionary and Its Adverb Definitions

Edward A. Fox
Robert C. Wohlwend
Phyllis R. Sheldon
Qi Fan Chen
Robert K. France

Technical Report 86-23

October 1986

Dept. of Computer Science
Virginia Tech (VPI&SU)
Blacksburg, VA 24061

ABSTRACT

The CODER (COmposite Document Expert/extended/effective Retrieval) project is an investigation of the applicability of artificial intelligence techniques to the information retrieval task of analyzing, storing, and retrieving heterogeneous collections of "composite documents." In order to support some of the processing desired, and to allow experimentation in information retrieval and natural language processing, a lexicon was constructed from the machine readable *Collins Dictionary of the English Language*. After giving background, motivation, and a survey of related work, the Collins lexicon is discussed. Following is a description of the conversion process, the format of the resulting Prolog database, and characteristics of the dictionary and relations. To illustrate what is present and to explain how it relates to the files produced from *Webster's Seventh New Collegiate Dictionary*, a number of comparative charts are given. Finally, a grammar for adverb definitions is presented, together with a description of defining formula that usually indicate the type of the adverb. Ultimately it is hoped that definitions for adverbs and other words will be parsed so that the relational lexicon being constructed will include many additional relationships and other knowledge about words and their usage.

CR Categories and Subject Descriptors: E.2 [**Data Structures**]: Tables; H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing - *dictionaries, linguistic processing, thesauruses*; I.2.1 [**Artificial Intelligence**]: Applications and Expert Systems - *natural language interfaces*; I.2.4 [**Artificial Intelligence**]: Knowledge Representation Formalisms and Methods - *relation systems*; I.2.7 [**Artificial Intelligence**]: Natural Language Processing - *language parsing and understanding, text analysis*

General Terms: Algorithms, Design

Additional Keywords and Phrases: CODER project, defining formula, dictionary parsing, machine readable dictionary, relational lexicon

1. INTRODUCTION

1.1 The CODER Project

The CODER (COmposite Document Expert/extended/effective Retrieval) project is an investigation of the applicability of artificial intelligence (AI) techniques to the information retrieval (IR) task of analyzing, storing, and retrieving heterogeneous collections of "composite documents." CODER has been designed [FOXE 86c, FRAN 86a] in a flexible fashion so it can be adapted to a wide variety of types of documents and users (see Figure 1). Several styles of human-computer interfaces are supported, and there are plans for extending those schemes. As can be seen in Figure 2, an individual with an information need, while searching through the text of possibly relevant documents or document passages [OCON 80], will often directly examine term relationships stored in the lexicon. Indeed, the lexicon plays a central role in the entire system's operations, since it must be accessed by both the analysis and the retrieval subsystems (see Figure 3).

To effectively use a comprehensive lexicon in an information retrieval system requires a number of techniques recently developed by the AI community. As was demonstrated through simulation [BELK 84], a blackboard (see [ERMA 80], [HAYEB 84], [ENSO 85]) is of great value to allow communication among a variety of expert system modules (see, for example [HAYE 83]). I^3R, an information retrieval system with a blackboard and a number of experts coded in LISP [THOM 85], following many of the suggestions of [BELK 84], has been under development, but has not involved lexicon construction or natural language processing. In the CODER project, it is hoped that the lexicon will assist with problems that have plagued others, such as disambiguation [EARL 73] and context recognition [CHAR 82]. The aim is for lexical-semantic relations to be derived from machine readable books with sufficient care and planning that the end result will be readily usable [MILL 85a,b].

CODER is implemented primarily in MU-Prolog [NAIS 85], and while the overall architecture is based upon message passing [FOXE 86b, 87], each of the many specialized experts will typically have its own set of rules [HAYE 85]. Prolog seems especially appropriate for this type of system development [HELM 85]. Logic programming can help not only in general problem

solving [KOWA 79], but also with the natural language component [PERE 83]. In similar fashion to the well-behaved ARGON system [PATE 84], frames (proposed in [MINS 75] and surveyed in [FIKE 85]) have been incorporated to represent various objects and object classes.

Based on previous experience in implementing a modern version of the SMART retrieval system [FOXE 83b], CODER is being developed as an even more versatile test bed [FOXE 85a] for handling composite documents [FOXE 85b]. Knowledge representation schemes are particularly important to investigate [FRAN 86b]. It will incorporate the most common and the most promising retrieval methods [FOXE 83a]. Ultimately, it should combine some of the best features of systems such as TOPIC [HAHN 84], I^3R, SMART, and RUBRIC [TONG 83], and allow careful experimentation to determine which methods do indeed lead to the greatest improvements in performance.

The CODER lexicon has two parts: one with general linguistic knowledge, and one with specialized knowledge about the area(s) covered by the document collection. Since for our initial experimentation a collection of over four years of the ARPANET AIList Digest has been gathered, we have been fortunate to obtain the machine readable text of the *Handbook of Artificial Intelligence* from the SUMEX Computer Project at Stanford University with the permission of William Kaufman. That work has already been analyzed to obtain some of the easily identifiable domain knowledge (e.g., phrases, names, and a taxonomy of topics). Because special knowledge is so important in developing sublanguage grammars [SAGE 75], further work with the *Handbook* may also be needed, especially to aid with document analysis.

General linguistic knowledge makes up the kernel of the lexicon, and can be used with any collection of documents. Indeed, the CODER lexicon may be of value outside CODER, to aid other types of natural language processing research.

1.2 Timeliness of Lexicon Construction Research

New high-capacity storage devices [FUJI 84], especially optical disks like CD-ROMs, can now be integrated with powerful microcomputers and advanced retrieval techniques [FOXE 86a]. Electronic books have been studied [WEYE 82], electronic encyclopedias (see [COOK 84], [WEYE 84]) are now available, and electronic dictionaries [FOXM 80] will soon proliferate. Yet,

while numerous data structures and related algorithms have been thoroughly studied by computer scientists [GONN 84], little attention has been given to access methods for text [FALO 85]. In certain cases, partial matching can be easily handled [PFAL 80]; superimposed coding [RAMA 85] is built-in to the Prolog interpreter being employed in the CODER development effort [NAIS 85] and performs well with large fact bases [SACK 85].

Machine readable dictionaries have been under investigation for a number of years [AMSL 84] and it is hoped that they will become more easily accessible [KAYM 84]. It should soon cost less than $20 each to produce a CD-ROM with the text of a large dictionary accessible by inexpensive personal computers. Even the vast *Oxford English Dictionary* (OED) will be available this decade in a compact machine-readable form for research and ordinary use [LESK 85].

Work in computational linguistics and other areas has brought us to the point of being able to manipulate not only the text but also some of the knowledge in dictionaries. This has long been a goal: Quillian's seminal work on semantic networks [QUIL 68] described a semantic memory including definitions of word senses and a taxonomic hierarchy of word concepts. In recent years, interest in the task has increased rapidly. Ahlswede [AHLS 85] describes a tool kit being developed to build lexicons from machine readable dictionaries. At IBM, semantic hierarchies have been automatically extracted from dictionaries [CHOD 85] and powerful parsers with syntactic and semantic components [MCCO 84] can be used to analyze a large subset of the English language. The newly-emerged discipline of logic programming [GENE 85] provides aid in a variety of problem solving tasks [KOWA 79], especially where (certain limited forms of) inferencing is required. Simmons [SIMM 83] encoded the key information in the first fifty pages of the *Handbook of Artificial Intelligence* (HAI) for subsequent use, and one ambitious project is underway to build an enormous, detailed knowledge base by encoding a large encyclopedia [LENA 86]. Various knowledge representation schemes are becoming widely used, and researchers are identifying what types of reasoning with that knowledge is computationally tractable [LEVE 84].

1.3 Need for Improved Information Retrieval

In spite of the advances just mentioned and others even more directly related, there has been little improvement during the last several decades in the effectiveness of commercially available information retrieval systems. According to one recent study involving a popular retrieval system,

after what lawyers and paralegals thought was thorough searching only about 20% of the relevant cases were identified [BLAI 85]. Yet new storage, processing, communication, and display methods have been applied to make almost 3000 databases [WILL 85b] accessible for a variety of purposes [WILL 85a]. With the advent of front-ends, gateways, and intelligent intermediary systems, there is a new potential to make these databases more convenient and effective [WILL 86]. It is hoped that automatic text retrieval methods will play a role in realizing this potential [SALT 86]. As the number of full-text databases grows [TENO 84], however, automatic methods heretofore tested with bibliographic databases will have to be adapted to allow selection of passages rather than documents [OCON 80]. Especially in that environment, lexical and other natural language processing (NLP) aids will be essential for improving retrieval effectiveness [DOSZ 86].

For information retrieval purposes, it is important to match relevant documents to users' queries even when there are mismatches in language usage. It has become clear in recent years that controlled and uncontrolled vocabularies have different properties in the retrieval process, and that effective use of both yields better results than if only one is allowed [SVEN 86]. A clearer understanding of words, word senses, and their interrelationships is essential to the creation and use of general and specialized thesauri that can lead to improvements in precision as well as recall.

The applicability of linguistics to information science has been studied for a number of years [SPAR 73]. While linguistically-motivated processing such as stemming [LOVI 68] leads to clear gains and can free users from the burden of decision-making about truncation, little benefit resulted from early attempts to apply syntactic analysis [SALT 80]. A recent survey [SALT 83 Chapter 7] indicates that new developments in computational linguistics and knowledge representation may be more promising. Semantic as well as syntactic analysis is being studied in an Intelligent Retrieval Expert System now under development [DEFU 85].

In addition to using linguistic information to improve query formulation, now that natural language front-ends to database management systems are becoming more common and better understood it seems appropriate to provide similar tools for information retrieval systems. Such an interface can not only facilitate a more natural dialog but also build the static and dynamic user models that are necessary for an adaptive mixed-initiative search session [DANI 86b]. For such a system to individualize its behavior [BORG 85], it must understand the user's vocabulary and relate that to the vocabulary of relevant documents [WILLP 85]. While vocabularies in a limited domain can be hand-crafted [NIEH 76], the broader domain of large scale information retrieval

systems requires automatic methods to construct more comprehensive lexicons. Lexical and semantic information is needed if the system is to help searchers map interest statements to documents in one or several collections by more than trivial translations (allowed for example, by Dialindex [DIAL 86]).

1.4 Needs in NLP

Computational linguists have realized the need for procedures to analyze and generate text. Augmented transition networks (ATNs, which are surveyed in [BATE 78]) have been widely used for syntactic and semantic analysis, and Prolog definite clause grammars (DCGs) have popularized language processing as well as enabled construction of powerful syntactic analyzers [PERE 83]. While syntax alone can be effective for sublanguage processing [SAGE 75], and semantic or memory-based parsers work for limited domains or for high level paraphrasing [LEBO 83a, 83b, 84], general applications require morphology, syntax, and semantics [MARC 84]. Integrated processing is needed for robust natural language understanding [SELF 86], so that computers can ultimately be applied to realistic comprehension tasks [RIES 82]. Yet computationally efficient techniques (see [GAZD 85], [POLL 85]) are needed to realize such a system.

Text comprehension requires knowledge as well as processing. Analysis by word experts can be extremely powerful [RIEG 81], and representations of phrasal patterns are also valuable [WILE 84]. Context provides additional clues when it can be recognized [CHAR 82], and methods for basing language processing on situation semantics are evolving [LESP 86]. For greatest flexibility, a declarative knowledge base containing all the objects and procedures involved in all these aspects of language processing must be developed. Simmons has highlighted the power of such an approach for analysis or generation of text and for machine translation [SIMM 84]. Kucera points out the many values of an on-line lexicon [KUCE 85]. Yet most language processing systems know about less than 3000 words, and the few which contain more than 10,000 words (e.g. [WHIT 83]) usually have little semantic information for most entries. Clearly a comprehensive lexicon with declarative representations of morphological, syntactic, phrasal, and semantic information would be a great boon for NLP.

1.5 Available Lexical Resources

Lexical resources of interest presently available fall into the following classes: tagged corpuses of text, machine-readable dictionaries, word lists, small lexicons in specially devised forms, and a few moderately large lexicons containing primarily morphological and syntactic information, supplemented with semantic field and restriction data.

The Brown Corpus [KUCE 67] has been analyzed extensively to determine frequency and grammatical tagging of words in a million-word sample of 1961 American English [FRANC 82]. The LOB Corpus is a similar British English collection, where probabilities have been applied to aid in tagging [BEAL 85]. Such gatherings of text are the raw material for lexicographers and for construction of dictionaries, and can be used to check and tune results of other processing.

Webster's Seventh New Collegiate Dictionary (W7) and *Pocket Dictionary* (WPD) are probably the most studied machine-readable dictionaries. Conversion from the typesetter form into a suitable computer format has been a difficult and time-consuming task— see [MCIL 84] and [MITT 85] regarding the *Oxford Advanced Dictionary of Current English*, and see [WOHL 86] about the *Collins Dictionary of the English Language* (CDEL). There has been a gradual shift from concern with storage efficiency and related matters [SHER 74] to providing a logical and convenient form to work with [PETE 82]. While various forms of W7 and WPD exist, the publisher has been reluctant to permit further release of tape copies. Longman's dictionary has also become a popular resource for linguistic studies, but several years ago the publisher began limiting distribution to the research community.

These dictionaries and other information resources have been collected by researchers such as Amsler and Walker at Bell Communications Resources [WALK 85]. Amsler in particular has worked with nouns and verbs, which make up by far the bulk of the language [AMSL 81]. Amsler and Walker have collected running text from the *New York Times* and noted that more than half of the words in that collection do not appear in W7, and vice versa. As a result, they have launched an ambitious campaign to gather almanacs and other knowledge resources that will supplement the dictionaries and help provide better coverage of the set of words now in active use.

The Waterloo Centre for the New OED has become a focal point for research in dictionary-related processing, owing to its committment to support production of a

machine-readable form of the *Oxford English Dictionary*. That effort is well underway, and should result in creation of the New OED within a few years since the keying of both the OED and the *Supplement* is complete. But copies of the dictionary sections already produced can only be worked with at the Centre, and emphasis to-date has been on practical problems of production and access, as well as on linguistic issues of what should be included, and in what form.

Oxford has supported develoment of another resource of linguistic information in the form of its Text Archive. Dictionaries, books, and a variety of other resources can be purchased for a nominal fee by researchers; many items are in the unrestricted class, which means that any researcher can have access — quite a different situation than is developing with many dictionary producers.

The above list is just a sampling of what is available; each author, editor, or publisher has different policies. Special arrangements can be made with originators of various resources, but for the sake of research in information retrieval or natural language processing it is especially important to obtain assurances from the start that such valuable assets as lexicons will be available to other researchers after completing a long project.

1.6 Lexical Knowledge for IR — Theories

Some type of indexing vocabulary is needed for any retrieval system. Svenomius surveys the evolution of presently used schemes [SVEN 86], and highlights a number of unresolved research questions, including:

- What is the effect of different kinds or degrees of vocabulary control on retrieval effectiveness?
- What is the effect on retrieval performance of automating different forms and kinds of vocabulary control?

Svenomius suggests that it will take more broad-minded research than has previously been carried out to resolve the many issues and interactions among such variables as subject matter of the collection, mix of free text vs. controlled search terms, user need for a few good vs. all relevant items, precision level desired, forms of control of the vocabulary (e.g., orthographic, synonym, hierarchical, related-term), automatic vs. intellectual control, and user need or desire for special control.

While the science of vocabulary control clearly stands in need of an underlying theory, there are useful theories that deal with automatic indexing and retrieval [SALT 83]. Suppose that the words in the text are to be used to produce a set of weighted index terms. Weights then can be based on two statistics: the number of documents in which a term appears, and the number of times a term appears in a document. Unfortunately, while a (medium or low frequency) word that appears frequently in a document is likely to be more important than one that occurs rarely, stylistic conventions confound this effect. In technical articles the general subject matter is often introduced in the title and first few sentences to set a context, and *not* repeated. Also, it is considered bad style to repeat a word many times, so authors tend to vary the terminology and means of expression to make reading more pleasant. Clearly, control of synonyms is important; in some fashion it is necessary to identify "conceptual clusters" of terms so that frequency values are not too low. On the other hand, frequency values can be artificially high because of homonymy and because on average every dictionary headword has two definitions [PETE 82]. While stemming is generally useful, it can further aggravate the homonym problem if searching on precise word senses is desired.

The number of documents in which a term occurs is another problematic statistic. Many studies have assumed that terms co-occurring in more documents than expected bear some relationship, but this is especially difficult to ascertain for low frequency terms, where automatic thesaurus construction is of most use. High frequency terms, according to the theory, should not be very valuable, but they make up the bulk of most queries, and so cannot be ignored, and building phrases from them is expensive.

These indexing theories can be integrated with the vector space retrieval model [SALT 83], where a different dimension is associated with each index term. However, there is a conceptual problem in that dimensions should be independent whereas terms certainly do not occur independently. External knowledge based on pseudoclassification or on a priori relationships among terms can be of particular value in a generalized vector space model where dependencies are accounted for [RAGH 86]. Thus, use of linguistic rather than co-occurrence based term association data may be both practically and theoretically beneficial.

Some recent retrieval systems have espoused theories less dependent on frequency information. Tong *et al.* [TONG 83] hold that (fuzzy) membership of documents relating to the

query rule-set can be determined from a knowledge base that represents a query. Croft and Thompson [THOM 85] suggest that document relevance can be inferred by logic from the query supplied. In both cases it is presumed that having more information than just a word list and using knowledge-oriented processing should provide a sound basis for retrieval.

1.7 Lexical Knowledge for IR — Studies

Lexical information can be used in retrieval in a number of ways. The classic approach has been to take some (i.e. the low frequency) terms in a query and "expand" them by replacing them with a class representative or by adding (possibly down-weighted or OR'ed) terms. Typically this involves using an intellectually or automatically produced thesaurus; the former may be useful for a general subject area while the latter must be created anew for each collection.

In comparing performance of automatic and manual retrieval approaches, Salton observed that a significant benefit resulted from having a good thesaurus [SALT 72]. Sparck Jones [SPAR 71] and others used co-occurrence frequencies and term clustering to automatically develop word classes particularized to a given collection, and found that a number of similar approaches all yielded a small improvement. Rather than develop a global collection thesaurus, Attar and Fraenkel [ATTA 77] produced a set of "searchonyms" from the relevant retrieved documents arising out of a feedback experiment. More widely applicable aids are desired, and linguistic solutions seem a likely source. An early investigation demonstrated that expanding queries by adding in terms that are lexically or semantically related does improve retrieval performance [FOXE 80]. This work was corroborated [EVEN 82], and later extended [FOXE 84]. Further studies by Evens et al. showed that finding related terms worked best with queries that initially led to few relevant documents [WANG 85]. However, online dictionaries with lexical and semantic relations are needed if such an approach is to be incorporated in a practical system.

Early proposals on the structure of these relational lexicons were made in [APRE 69], [MELC 73], and later refined in [EVEN 79,85]. Semi-automatic tools to aid in the construction process have been developed [AHLS 85]. Several large scale lexicons are already under development (e.g., [WHIT 83]), in part to help with parsing [SAGE 81]. Eventually, analysis of documents to identify key topics [HAHN 84] and structure [BABA 85] will become feasible with the aid of such lexicons.

An alternative to thesaurus use is improving the representation of documents and queries. If relationships among words can be determined without reference to the collection, improvements to standard vector processing are possible [RAGH 86]. Many studies have considered identifying and using noun phrases, which often improves precision over using single terms alone (ex. [DEFU 85]). Linguistic analysis of queries can help in this regard [SPAR 84]. A more direct approach is to disambiguate word senses in documents and queries, so that matches are at the sense rather than lexeme or stem level. If the many difficult problems involved could be solved it might then be possible to represent relationships among word senses, thereby combining recall-improving and precision-improving devices. A limited success in both these directions has been achieved by attaching semantic field tags to words to help disambiguate senses and to identify related fields [WALK 85].

Katzer et al. [KATZ 86] studied the effects of resolving anaphora in one experimental test. They replaced anaphoric references with the proper word or phrase and compared retrieval results against those obtained from the original texts. No clear findings were obtained from this trial; it seems that more sophisticated methods are needed to obtain a consistent improvement.

Several studies have suggested providing a natural language dialog interface for users of retrieval systems. [DANI 86b] surveys the early work (in the cognitive science and retrieval areas) and discusses the issues involved. Applying these ideas to the matter of vocabulary suggests the following.

- A user's statements and responses may reflect knowledge and size of active vocabulary in the area of interest, which could be used to predict the degree of expansion that should be applied to a query.
- As a dialog develops, the system should obtain context and background to aid in "understanding" the query.
- During the course of several searches, the system can infer both dynamic (i.e. related to the state of a particular strategy) and static (longer-term) models of the user to aid in selecting search strategies [DANI 86a].
- The system can "explain" (and when appropriate, ask for approval of) its actions.

1.8 Lexical Knowledge for NLP

Since lexical knowledge is so central to NLP, it should be adequate to simply list some of the important applications.

- A complete lexical entry provides information at all linguistic levels regarding a lexeme, so a parser need only possess procedural knowledge, and yet can prepare the fewest number of unambiguous parses.
- Lexical and semantic information can aid in creating knowledge representations.
- Combining a knowledge base and a lexicon should aid in all types of inferencing.
- For generation, having a large lexicon accessible from the semantic representation component should allow a system to vary its presentation content and style, tailor its responses to a user's vocabulary, and prepare text that is both salient and complete.

2. CODER Lexicon

2.1 Long Range Plan

Initially, the CODER lexicon will be made up of information in the *Collins Dictionary of the English Language* (CDEL), and that should be adequate for testing of the prototype system. Later, it is hoped that a more comprehensive lexicon will be developed, including the *Oxford Advanced Learners Dictionary of Current English* (OALDCE), and the *Oxford Dictionary of Contemporary Idiomatic English* (ODCIE, Volumes 1 and 2). Whereas CDEL is a large dictionary with current terms included about scientific and technical areas, OALDCE is designed to help learners; while smaller it has more detail, especially about verbs. ODCIE will help with idioms/phrases and with verb-particle combinations. All in all, this collection of dictionaries should be of enormous aid in building a comprehensive lexicon (of British and U.S. English)!

Long term goals in connection with building a comprehensive lexicon are:

1) to transform the (freely available for research use) machine readable dictionaries of the Oxford Text Archive ([HANK 79], [HORN 74], [COWI 75,83]) from their original typeset form to a Prolog fact base, which can then be requested from that same Archive,

2) to analyze the structure of definitions so they can be automatically parsed,

3) to combine the results of processing several dictionaries to obtain as many lexical and semantic relations as possible, and

4) to apply this lexicon to the processing of documents and queries, by incorporating it in CODER's natural language analysis components.

2.2 Description of the Collins Dictionary

There is a separate entry for each headword in the dictionary. Headwords are words or phrases, usually in lower-case, but they can also be proper names, mixed alphanumerics or numerics (e.g. "A1"), words with embedded/leading/trailing apostrophes, prefixes/suffixes, compound words, foreign terms, or abbreviations/symbols/acronyms. When headwords have the same spelling but have different origins they are assigned homograph numbers. For some headwords there are allowable variant spellings which are given in the entry, sometimes with a country indicator (e.g. "centre or U.S. center"). Headwords and variant spellings both have an associated syllabification. Regular inflections (as defined in the front materials) are not given, but exceptions are shown (e.g. "goose pl. geese"). Morphological variants (different forms of the headword that are not headwords individually) appear at the end of an entry without any definition, but with a part of speech indication.

There are eight standard parts of speech: adjective, adverb, conjunction, interjection, noun, preposition, pronoun, and verb. Verb may be transitive or intransitive. Some other less traditional parts of speech include: determiner, sentence connector (ex. "therefore"), sentence substitute (ex. "aye"), adjective postpositive (i.e. words like "ablaze" that are used predicatively or after the noun), adjective postnominal (i.e. words like "acting" used before the noun), plural nouns (ex. "trousers"), and modifier (ex. "absentee"). Abbreviations and symbols are also indicated in the part of speech location. Approximately 2500 famous people's names appear as headword entries, possibly with first names, pseudonyms, nicknames, titles, and original names.

Associated with word senses are sometimes a label or "category" that restricts the word to subject field, appropriateness, connotation, nationality, etc. The major classes of categories are: temporal, usage, connotative, subject-field, national/regional, or trademarks. Senses appear in rank order, so that the first is most commonly used and the last is least commonly used. Subdefinitions appear when there is a slight variation in sense of a word, or when the meaning is constant while the part of speech varies (ex. "beige" is a noun indicating a color but can also be an adjective when describing items of that color). Categories pertain to a given sense and so apply to all subsenses.

Example sentences and phrases illustrate the use of word senses. Usage notes supplement this information with remarks on how to use a particular word or word sense. Further guidance is

shown through one of several types of cross reference: related adjectives (ex. noun "wall" has related adjective "mural" which is of French origin), comparisons (indicated by "see" or "see also" or "compare"), and alternative names (indicated by "also" or "also called" or "Official name:").

Etymologies and pronunciation appear in entries; occurrences of the latter are scattered throughout entries. Often pronunciation is shown in parenthesis.

2.3 Conversion from Typeset Form

A SUN Microcomputer Inc. workstation with a large disk store was used to process the CDEL tape provided by the Oxford Text Archive. UNIX routines such as lex, yacc, awk, as well as special C programs were developed to convert the typeset form of CDEL into a form that could be loaded into a Prolog fact base. Nine passes were required so that the font changes and special characters present could be analyzed sufficiently to decode the structure of definitions. 95% of the roughly 80,000 headwords were successfully processed in this fashion [WOHL 86].

The remaining 3000 headwords could not be handled by the conversion program and so were manually edited until they could be processed. Included in this set were entries like "take" which had a very large number of senses.

2.4 Characteristics of Constructed Relations

A conceptual structure was imposed upon the dictionary in order to extract useful information. Figure 4 shows the hierarchical structure of the CDEL. For each lexeme there may be several homographs which are each assigned a headword. Below the headword is the syntax level where part of speech is indicated. Next is the sense level, and at the bottom is the sub-sense level. While a lexeme itself determines the first level, a list of numbers indicates a given instance of a sense or sub-sense according to these distinctions.

Table 1 shows the list of relations, along with some description, that have been produced as a result of processing CDEL. Since Prolog can serve as a database query language, use of such normalized relations is especially convenient. Later, organizations that are more space and time efficient may be added or may replace the current collection.

The syntax of the relations is given in Table 2, where "word" and "homnum" and sometimes the "defnum" and "subnum" specify varying levels of specificity as discussed above and shown in Figure 4. As can be seen from the format, c_ABBREV and c_ALSO_CALLED both allow reference to a separate headword. The c_COMPARE relation can associate with a particular word sense where a comparision is illustrative, and the c_RELADJ also relates to another word sense. When a headword has two or more morphological variants (see c_MORPH), the preferred is first and others follow in order. When a person's name appears, all parts that are not the last name are shown in the c_NLAST relation; it is often difficult to separate the parts of a name more thoroughly. The c_PLURAL relation often shows irregular plural forms but may show other irregular inflections or forms, which can be distinguished by knowing the part of speech. The singular form or ending may appear, as is the case for c_SINGULAR. Sometimes semicolons appear in the string of a c_SAMP usage example, to separate compatible definitions at the sense level. Underscores appear in the c_SYLL and c_VAR_SYLL strings to separate syllables. It should be noted that since CDEL is a British dictionary, in most cases the headword in the c_VAR_SPELL relation is the British form and the variant is the U.S. form.

Because there is a length limit on tuples that will be stored in the MU-Prolog external database, long strings appearing in definitions (c_DEF) and usage examples (c_USAGE) were split into multiple numbered tuples. The number of such tuples is given for each entry in the c_DEF_NUM or c_USAGE_NUM relation.

2.5 Comparison with Other Dictionaries

Tables below give frequency information about the CDEL (also referred to as CED) fact database according to the initial processing of almost 80,000 headwords. From Table 3 it can be seen that CDEL is actually larger than *Webster's Seventh New Collegiate Dictionary* (see discussion in [PETE 82]) and, in many cases, has more types of information (i.e. an average of about 5.5 as opposed to 4.5 tuples per headword). This can be of particular value. Indeed, there are more than twice as many morphological variants and categories. W7 has no usage samples while CDEL has a large number. Note that "CED freq" refers to the number of tuples in a given relation, "CED freq/hd" is the average number of times a relation appears for a headword, and "CED % usage" is the relative frequency of a relation compared with all other relations.

The category entries, the most common of which are shown in Table 4, give clues as to semantic field. As is usual with most systems of manually assigned categories, however, a fair amount of cleanup is needed to make the category scheme more systematic and less redundant. Thus, there are a number of broad terms (e.g. "sport") and many more specific terms which could be related to parent broader terms.

Table 5 shows how different dictionary editors approach the use of homograph numbers. In W7 there is a new entry just about every time a part of speech changes. This can be seen since W7 has more frequent use of larger homograph numbers. In CDEL, on the other hand, multiple parts of speech often appear for a given headword rather than appearing as a separate headword.

Table 6 is the Part of Speech list; here again it is seen how CDEL compares with *Webster's Seventh New Collegiate Dictionary* and *Pocket Dictionary* (studied by Amsler). It should be noted that since in [PETE 82] there is a distinction between the primary and secondary parts of speech, it was necessary to combine those results to aid in comparison. Almost two thirds of the entries are nouns. While CDEL is larger than W7, it has (in absolute terms) slightly fewer adjective and adverbs, but this may be influenced by the fact that there are more parts of speech used to classify entries.

W7 and CDEL have very similar distributions of the number of word senses per headword, as can be seen in Table 7. Unhappily, the tail of the distribution is much longer for CDEL, which forced the processing programs to have to deal with very long entries (up to 50 senses!).

On the other hand, the tail of the distribution on number of sub-senses is much longer for W7 than for CDEL, as can be seen in Table 8. Apparently, there are fewer word senses, but more sub-senses per sense (up to 14!), in W7.

2.4 Revisions Needed to Load into MU-Prolog Fact Base

To load the relations into Prolog, it was necessary to make some character substitutions so that the normal read-in processing could be used. Since the single quote delimits atoms, apostrophes were converted to the caret ("^"). Periods are used to end Prolog facts and rules, so all periods were changed to a vertical bar ("|").

In some cases, during the nine passes of processing the routines did not recognize beginnings of certain relations appearing in the midst of definitions. In particular, "compare" and "see also" entries had to be removed from definitions (i.e. the c_DEF relation) and manually entered as new relations (e.g. c_COMPARE).

It was noted that pronunciation information appears in many different places in several of the relations. Usually, pronunciations are found in parentheses immediately following some words in the text field of the c_DEF file. To allow decoding of these appearances, the mapping of character sequences in the Prolog fact base to IPA codes is shown in Table 9.

2.5 Status

All entries in the dictionary are now in a form that can be consulted by the MU-Prolog interpreter. Spot checking has shown that there are still a small number of "compare" and "see also" phrases appearing in the text field of other relations, so a bit more cleanup is needed. After that, specifications must be prepared so that the many relations can be loaded into an external MU-Prolog fact base, to be accessed using the two-level superimposed coding scheme [SACK 83]. A certain amount of tuning of the parameters is expected so that access to the CDEL facts will be rapid for the most common types of processing.

With the CDEL relations available, a pilot study has been undertaken to delve further into construction of lexical and semantic relations. While others have studied nouns and verbs [AMSL 81] and adjectives [AHLS 83], there is little information available on adverb definitions (except the work by Klick discussed in [EVEN 82]). The remainder of this report deals with the content and structure that have been observed in CDEL adverb definitions.

3. CDEL Adverb Definitions

3.1 Adverb Types and Functions

Adverbs are very useful modifiers. Their function allows them to be typed according to the following scheme:

FUNCTION	TYPE
(logical)	transition
when	time
where	location direction
why	purpose
how	manner state degree

Types are often easy to identify by examining the text of the adverb definition.

TRANSITION is for idiomatic expressions such as "on the other hand" which must be identified as phrases, not parsed. Therefore, recognition of this type should be done before further processing.

TIME is indicated by "from, at, after, before, between, during, in, throughout, to." Less often, participles "coming" and "following" indicate TIME. When both "from" and "to" are present, a time span is usually specified.

LOCATION is for adverbs indicating place. Defining formulas (i.e., a pattern present in the definition) usually begin with "at, in, before, behind, beneath, between, on, outside, over, throughout, to, within."

DIRECTION is used for adverbs that indicate movement from one position to another and have such definitions as "from one side to the other" or "on or to the other side." Definitions tend to contain "from" or "to" and sometimes "towards, up, in."

PURPOSE is for adverbs which tell for what purpose or reason something is done. The definitions usually begin with "for" as in "for an express purpose" or "for one reason only."

MANNER type definitions are usually prepositional phrases beginning with "as, by, in, like, with, without" or participial phrases beginning with "using." They describe the manner, way, or fashion in which something exists or is done, and most often follow the pattern "in a <ADJSTRG> manner." Manner can be viewed as having two subcategories: STATE and DEGREE.

STATE refers to state of being. It is the manner in which something exists. As with MANNER, prepositional phrases beginning with "in" often define it. "On" is also common, as in "on fire." Present participles such as "burning" can be used. The "-ly VED" combination, as in "emotionally aroused" or "brightly illuminated" is easy to identify. Less common but still reliable are definitions beginning with "being" or "-ly VING" such as "being behind a competitor" or "aimlessly drifting."

DEGREE describes the intensity with which something is done. It includes, especially, those words defined as "intensifier" which modify adjectives that can be graded. An example is the adjective "interested" which takes adverbs "very, slightly, somewhat, not, not very." Any adjective that refers to speed or force or steadiness or quality is a likely candidate to be modified by an adverb of degree.

3.2 Grammar for Adverb Definitions in CDEL

CDEL definitions often have "usage" information which indicates how the adverb can be used or in what context it can be used. There are special tokens, for example, ":" or ", esp.", and word patterns such as:

used (in | to)
as modifier
a less common (spelling of | word for)
a variant [spelling] of
that | as
an (archaic | informal | obsolete) word for
an (emphatic | intensive) form of
an exclamation used to
another (term | word) for
dialect
short for
the (comparative | superlative) of

Sometimes there are several definitions that appear together; they are usually separated by ";" or "!" as is shown in the following grammar rule:

ADVDEF -> [USAGE] [ONEDEF [';' ONEDEF | '!' ONEDEF]*] ['.'] [USAGE]

A preliminary version of a grammar has been developed as can be seen in Table 10.

Much clearer, however, is Table 11, which gives the defining formulas in chart form, relating each defining formula to a type and indicating the complement type expected. With the aid of Table 11, types following the above descriptions can be easily obtained.

Analyzing definitions can be done by considering the structures (pattern of parts of speech) present or by considering specific wording. Thus, when we see

ADV (; ADV)*

then each separate adverb definition is a synonym; when we see

ADV ADV

this indicates that both entries together form a synonym phrase (like "all around"); when we see

ADV VED

this indicates that ADV is a synonym while the VED is a state; and when we see

ADV VING

this indicates that ADV is a synonym while VING is a motion verb. Finally, when we have

a less common spelling of ADV

then we know we have an ADV that could substitute for the headword.

In some cases of Table 11 there are more than one possible type of definition. There are, though, algorithms that can be used to decide, such as:

after: if the NP complement contains "time" or "delay"
then type = time
else type = location

to: if the complement NP contains "extent" or "degree"
then type = degree
else type = direction or location

3.3 Problems

Further work is needed to actually program a parser for the adverb definitions. However, there are some problems to resolve beforehand. A notational scheme for handling grades of intensity (ex., +++ for "very accelerated") must be developed and tied in with grades in the definitions. Lexical handling of the ";" and "!" entries, and separation of the USAGE forms is needed. Most difficult, however, is the issue of handling the word "not" appearing in the middle of a definition. Scopes must be determined accurately so that antonyms can be properly recognized or other appropriate analysis invoked.

4. Summary and Conclusions

A large and comprehensive lexicon would be of particular value to help with IR and NLP research investigations. Due to the availability of machine readable dictionaries, it is now possible to develop such a lexicon. With optical disk technology rapidly improving, it is clear that a large lexicon could be easily distributed on inexpensive read-only or write once media. As a result of progress in NLP and IR, it is clear that dictionary entries can be at least partially analyzed, to provide even more useful information for such tasks as query expansion.

In connection with the CODER Project, initial processing of the Collins dictionary has been completed, and a Prolog fact base that is syntactically correct has been developed. Additional cleaning will take place, and then the MU-Prolog external database routines will be used to index these facts for fast partial match access according to a two-level superimposed coding scheme. At that time the fact base will be usable to support testing of the CODER prototype now under development.

More information is available in the CDEL than is readily accessible, however, since in the Prolog fact base definitions and usage examples are but simple uninterpreted text strings. Because Ahlswede, Amsler, Evens and others have already studied definitions of nouns, verbs, and adjectives, it was decided that definitions of adverbs in CDEL should be analyzed (though there has been work by Klick that is described in [EVEN 82]). A simple grammar for CDEL adverb definitions has been specified, and related to the various types of adverbs that are present.

In the future, a parser for adverb definitions could be prepared, and, ultimately, parsers for the other parts of speech could be developed. Later, other dictionaries could be analyzed too. As new sets of information become available they can in turn become the subject of further experimental studies aimed at applying lexicology and computational linguistics to help develop more effective information retrieval systems.

BIBLIOGRAPHY

[AHLS 83] Ahlswede, T. E. A Linguistic String Grammar of Adjective Definitions from *Webster's Seventh Collegiate Dictionary*. In *Humans and Machines*, S. Williams (ed.), 101-127, 1983.

[AHLS 85] Ahlswede, T.E. A Tool Kit for Lexicon Building. In *Proc. of the 23rd Annual Meeting of the ACL*, 268-276, July 1985.

[AMSL 80] Amsler, R.A. The Structure of the Merriam-Webster Pocket Dictionary. Dissertation. TR-164, Univ. of Texas at Austin, Dec. 1980.

[AMSL 81] Amsler, R.A. A Taxonomy for English Nouns and Verbs. In *Proc. Annual Mtg. of the ACL*, 133-138, June 29-July 1, 1981.

[AMSL 84] Amsler, R.A. Machine-Readable Dictionaries. *ARIST*, 19:161-209, 1984.

[APRE 69] Apresyan, Y. D., I.A. Mel'cuk, and A.K. Zolkovsky. Semantics and Lexicography: Towards a New Type of Unilingual Dictionary, In *Studies in Syntax and Semantics*, ed. F. Kiefer, 1-33. Dordrecht - Holland, D. Reidel, 1969.

[ATTA 77] Attar, R. and Aviezri S. Fraenkel. Local Feedback in Full-Text Retrieval Systems. *J. ACM*, 24(3): 397-417, July 1977.

[BABA 85] Babatz, R. and M. Bogen. Semantic Relations in Message Handling Systems: Referable Documents. In *Proc. IFIP WG 6.5 Symposium*, Sept. 1985.

[BATE 78] Bates, Madeleine. The Theory and Practice of Augmented Transition Networks. In L. Bolc: *Natural Language Communication via Computers*, Springer-Verlag, Berlin, 1978.

[BEAL 85] Beale, Andrew David. Grammatical Analysis by Computer of the Lancaster-Oslo/ Bergen (LOB) Corpus of British English Texts. In Proc. *1985 ACL Annual Meeting*, 293-298.

[BELK 84] Belkin, N.J., Hennings, R.D., and T. Seeger. Simulation of a Distributed Expert-Based Information Provision Mechanism. In *Inf. Tech.: Res. Dev. Applications*. 3(3): 122-141, 1984.

[BLAI 85] Blair, D.C. and M.E. Maron. An Evaluation of Retrieval Effectiveness for a Full-Text Document-Retrieval System. *Commun. ACM*, 28(3):289-299, March 1985.

[BORG 85] Borgman, Christine L. Designing an Information Retrieval Interface Based on User Characteristics. In *Res. & Dev. in Inf. Ret., Eighth Annual Int. ACM SIGIR Conf.*, Montreal, 139-146, June 1985.

[CHAR 82] Charniak, E. Context Recognition in Language Comprehension. In *Strategies for Natural Language Processing*, ed. by Wendy G. Lehnert and Martin H. Ringle, Lawrence Erlbaum Assoc., Hillsdale NJ, 1982, 435-454.

[CHOD 85] Chodorow, Martin S., Roy J. Byrd, and George E. Heidorn. Extracting Semantic Hierarchies from a Large On–Line Dictionary. In *Proc. of the 23rd Annual Meeting of the ACL*, 299-304, July 1985.

[COOK 84] Cook, P.R. Electronic Encyclopedias. *Byte*, 9(7):151-170, July 1984.

[COWI 75] Cowie, A.P. and R. Mackin. *Oxford Dictionary of Current Idiomatic English. Volume 1: Verbs with Prepositions & Particles*. Oxford Univ. Press, Oxford, 1975.

[COWI 83] Cowie, A.P., R. Mackin, and I.R. McCaig. *Oxford Dictionary of Current Idiomatic English. Volume 2: Phrase, Clause & Sentence Idioms*. Oxford Univ. Press, Oxford, 1983.

[DANI 86a] Daniels, Penny J. The User Modelling Function of an Intelligent Interface for Document Retrieval Systems, In Proc. *IRFIS 6. Intelligent information systems for the information society*, Frascati, Sept. 1985, Amsterdam, North-Holland, 1986.

[DANI 86b] Daniels, P.J. Cognitive Models in Information Retrieval - an Evaluative Report. Final Report to the British Library Research and Development Department on Project Number SI/G/753, May 1986.

[DEFU 85] Defude, B. Different Levels of Expertise for an Expert System in Information Retrieval. In *Res. & Dev. in Inf. Ret., Eighth Annual Int. ACM SIGIR Conf.*, Montreal, 147-153, June 1985.

[DIAL 86] DIALOG Information Services Inc. DIALINDEX. Database 411 from DIALOG Information Retrieval Service, Specification Sheets, Jan. 1986.

[DOSZ 86] Doszkocs, Tamas E. Natural Language Processing in Information Retrieval. *J. Am. Soc. Inf. Sci.*, 37(4), 191-196, 1986.

[EARL 73] Earl, Lois. Use of Word Government in Resolving Syntactic and Semantic Ambiguities. *Inform. Stor. Retr.*, 9:639-664, 1973.

[ENSO 85] Ensor, R.J., and J.D. Gabbe. Transactional Blackboards. *Proc. of IJCAI '85*, 340-344, August 1985.

[ERMA 80] Erman, L.D., Hayes-Roth, F., Lesser, V.R., and D.R. Reddy. The Hearsay-II Speech-Understanding System: Integrating Knowledge to Resolve Uncertainty. *ACM Comp. Surveys*, 12:213-253, 1980.

[EVEN 79] Evens, M.W. and R.N. Smith. A Lexicon for a Computer Question-Answering System. *Am. J. Comp. Ling.*, Microfiche 83: 1-93, 1979.

[EVEN 82] Evens, Martha. Structuring the Lexicon and the Thesaurus with Lexical-Semantic Relations. Final Report to the NSF, 1985.

[EVEN 85] Evens, Martha., J. Vandendorpe, and Yih-Chen Wang, Lexical Semantic Relations in Information Retrieval. In *Humans and Machines: The Interface Through Language*, Ablex, ed. S. Williams, 73-100, 1985.

[FALO 85] Faloutsos, C. Access Methods for Text. *ACM Comp. Surveys*, 17(1):49-74, March 1985.

[FIKE 85] Fikes, Richard and Tom Kehler. The Role of Frame-Based Representation in Reasoning. *Commun. ACM*, 28(9):904-920, Sept. 1985.

[FOXE 80] Fox, E.A. Lexical Relations: Enhancing Effectiveness of Information Retrieval Systems. *ACM SIGIR Forum*, 15(3):5-36, Winter 1980.

[FOXE 83a] Fox, E.A. Extending the Boolean and Vector Space Models of Information Retrieval with P-Norm Queries and Multiple Concept Types. Dissertation, Cornell University, University Microfilms Int., Ann Arbor MI, Aug. 1983.

[FOXE 83b] Fox, E.A. Some Considerations for Implementing the SMART Information Retrieval System under UNIX. TR 83-560, Cornell Univ., Dept. of Comp. Sci., Sept. 1983.

[FOXE 84] Fox, E.A. Improved Retrieval Using a Relational Thesaurus Expansion of Boolean Logic Queries. In Proc. *Workshop on Relational Models of the Lexicon*, Martha Evens (ed.), Stanford, CA, July 1984 (to appear).

[FOXE 85a] Fox, E.A. Composite Document Extended Retrieval: An Overview. In *Res. & Dev. in Inf. Ret., Eighth Annual Int. ACM SIGIR Conf.*, Montreal, 42-53, June 1985.

[FOXE 85b] Fox, E.A. Analysis and Retrieval of Composite Documents. In *ASIS '85, Proc. 48th ASIS Ann.Mtg.*, 54-58, Oct. 1985.

[FOXE 86a] Fox, E.A. Information Retrieval: Research into New Capabilities. In *CD-ROM: The New Papyrus*, Steve Lambert and Suzanne Ropiequet (eds.), Microsoft Press, 1986, 143-174.

[FOXE 86b] Fox, E. A., and R. K. France. Architecture of a Distributed Expert System for Composite Document Entry, Analysis, Representation, and Retrieval. Proceedings *Third Annual USC Comp. Sci. Symp.; Knowledge-Based Systems: Theory and Applications*, March 31-April 1, 1986, Columbia, S.C.

[FOXE 86c] Fox, E. A. A Design for Intelligent Retrieval: The CODER System. *The Second Conference on Computer Interfaces and Intermediaries for Information Retrieval*, 28-31 May 1986, Boston MA.

[FOXE 87] Fox, Edward A. and Robert K. France. Architecture of an Expert System for Composite Document Analysis, Representation and Retrieval. *Int. J. of Approximate Reasoning*, 1(2) to appear.

[FOXM 80] Fox, M.S., D.J. Bebel, and A.C. Parker. The Automated Dictionary. *IEEE Computer,* 35-48, July 1980.
[FRAN 86a] France, Robert K. An Artificial Intelligence Environment for Information Retrieval Research. MS Thesis, VPI&SU Dept. of Comp. Sci., Blacksburg VA, July 1986.
[FRAN 86b] R.K. France, and E.A. Fox. Knowledge Structures for Information Retrieval: Representation in the CODER Project. *Proceedings IEEE Expert Systems in Government Conference*, October 20-24, 1986, McLean VA, (to appear).
[FRANC 82] Francis, W. Nelson and Henry Kucera. *Frequency Analysis of English Usage: Lexicon and Grammar*. Boston: Houghton Mifflin Company, 1982.
[FUJI 84] Fujitani, L. Laser Optical Disk: The Coming Revolution in On-Line Storage. *Commun. ACM*, 27(6):546-554, June 1984.
[GAZD 85] Gazdar, Gerald and Geoffrey K. Pullum. Computationally Relevant Properties of Natural Languages and their Grammars. Report No. CSLI-85-24. Center for the Study of Language and Information, Stanford CA, 1985.
[GENE 85] Genesereth, Michael R. and Matthew L. Ginsberg. Logic Programming. *Commun. ACM,* 28(9):933-941, Sept. 1985.
[GONN 84] Gonnet, G.H. *Handbook of Algorithms and Data Structures*. Addison-Wesley, Reading, MA, 1984.
[HAHN 84] Hahn, U. and Reimer, U. Heuristic Text Parsing in 'Topic': Methodological Issues in a Knowledge-based Text Condensation System. In *Representation and Exchange of Knowledge as a Basis of Information Processes*, ed. by Hans J. Dietschmann, North-Holland, New York, 143-163, 1984.
[HANK 79] Hanks, P. ed. *Collins Dictionary of the English Language,* William Collins Sons & Co., London, 1979.
[HAYE 83] Hayes-Roth, F., Waterman, D.A. and Lenat, D.B., eds. *Building Expert Systems*, Addison-Wesley, Reading, MA, 1983.
[HAYE 85] Hayes-Roth, F. Rule-Based Systems. *Commun. ACM,* 28(9):921-932, Sept. 1985.
[HAYEB 84] Hayes-Roth, B. BB1: An Architecture for Blackboard Systems that Control, Explain, and Learn About Their Own Behavior. Technical Report No. STAN-CS-84-1034, Stanford University Dept. of Comp. Science, December 1984.
[HELM 85] Helm, A.R., Marriott, Kimbal, and Catherine Lassez. Prolog for Expert Systems: An Evaluation. *Proc. of Expert Systems in Government Symp.*, 284-293, October 1985.
[HORN 74] Hornby, A.S. ed. Oxford Advanced Dictionary of Current English, Oxford University Press, Oxford, 1974.
[KATZ 86] Katzer, Jeffrey, Susan Bonzi and Elizabeth Liddy. The Effects of Anaphoric Resolution on Retrieval Performance: Preliminary Findings. *Proc. 49th Ann. Mtg. Amer. Soc. Inf. Sci.*, Sept. 28 - Oct. 2, 1986, Chicago, IL, 118-122.
[KAYM 84] Kay, Martin. The Dictionary Server. In Proc. *COLING 84*, Stanford, CA, July 2-6, 1984.
[KOWA 79] Kowalski, R.A. *Logic for Problem Solving*. Elsevier North-Holland, New York, 1979.
[KUCE 67] Kucera, Henry and W. Nelson Francis. *Computational Analysis of Present-Day American English*. Brown University Press, Providence, 1967.
[KUCE 85] Kucera, Henry. Uses of On-Line Lexicons. Proc. *First Conference of the UW Centre for the New Oxford English Dictionary: Information in Data*. Nov. 6-7, 1985, Waterloo, Canada, 7-10.
[LEBO 83a] Lebowitz, Michael. Memory-Based Parsing. *Artificial Intelligence* 21: 285-326, 1983.
[LEBO 83b] Lebowitz, Michael. RESEARCHER: An Overview. Columbia University Dept. of Computer Science, Tech. Report CUCS-54-83, May 1983.
[LEBO 84] Lebowitz, Michael. Using Memory in Text Understanding. Columbia University Dept. of Computer Science, Tech. Report CUCS-121-84, May 1984.

[LENA 86] Lenat, Doug, Mayank Prakash and Mary Shepherd. CYC: Using Common Sense Knowledge to Overcome Brittleness and Knowledge Acquisition Bottlenecks. *The AI Magazine*, 65-84, Winter 1986.

[LESK 85] Lesk, M. Report on *Information in Data: Using the Oxford English Dictionary on a Computer,* In *IRList Digest,* 1(28), 31 Dec. 1985.

[LESP 86] Lesperance, Yves. Toward a computational interpretation of situation semantics. *Comput. Intell.*, 2:9-27, 1986.

[LEVE 84] Levesque, Hector J. "A Fundamental Tradeoff in Knowledge Representation and Reasoning." *Proceedings of the Fifth CSCSI National Conference (London, ON, May 1984),* 141-152.

[LOVI 68] Lovins, B.J. Development of a Stemming Algorithm. *Mech. Trans. and Comp. Ling.*, 11(1-2):11-31, March-June 1968.

[MARC 84] Marcus, Mitchell P. Some Inadequate Theories of Human Information Processing. In *Talking Minds: The Study of Language in Cognitive Science,* eds. Thomas G. Bever, John M. Carroll, and Lance A. Miller, MIT Press, Cambridge MA, 1984, 253-278.

[MCCO 84] McCord, Michael C. Semantic Interpretation for the Epistle System. In Proc. *ILPC-84,* 65-76.

[MCIL 84] McIlroy, M.D. Personal Communication on February 8, 1984.

[MELC 73] Mel'cuk, I.A. Towards a Linguistic 'Meaning <=> Text' Model. In *Trends in Soviet Theoretical Linguistics,* ed. F. Kiefer, 33-57. Dordrecht - Holland, D. Reidel, 1973.

[MILL 85a] Miller, George A. Dictionaries of the Mind. In *Proc. of the 23rd Annual Meeting of the ACL,* 305-314, July 1985.

[MILL 85b] Miller, George A. Wordnet: A Dictionary Browser. Proc. *First Conference of the UW Centre for the New Oxford English Dictionary: Information in Data.* Nov. 6-7, 1985, Waterloo, Canada, 25-28.

[MINS 75] Minsky, M. A Framework for Representing Knowledge. In *The Psychology of Computer Vision,* ed. by P. Winston, McGraw-Hill, New York, 1975.

[MITT 85] Mitton, Roger. A Description of the File OALD.DAT. Personal Communication, July 18, 1985.

[NAIS 85] Naish, Lee. *MU-Prolog 3.2db Reference Manual.* Melbourne Univ., July 1985.

[NIEH 76] Niehoff, R.T. Development of an Integrated Energy Vocabulary and the Possibilities for On-line Subject Switching. *J. Am. Soc. Inf. Sci.* 27(1): 3-17, Jan.-Feb. 1976.

[OCON 80] O'Connor, J. Answer-Passage Retrieval by Text Searching. *J. Am. Soc. Inf. Sci.*, 31(4):227-239, 1980.

[PATE 84] Patel-Schneider, P.F., R.J. Brachman, and H.J. Levesque. ARGON: Knowledge Representation meets Information Retrieval. Fairchild Technical Report No. 654, FLAIR Technical Report No. 29, Sept. 1984.

[PERE 83] Pereira, F. Logic for Natural Language Analysis. Tech. Note 275, SRI Int., Jan. 1983.

[PETE 82] Peterson, James L. Webster's Seventh New Collegiate Dictionary: A Computer-Readable File Format. TR-196, Univ. of Texas at Austin, Dept. of Comp. Sci., May 1982.

[PFAL 80] Pfaltz, J.L., Berman, W.H., and E.M. Cagley. Partial-match retrieval using indexed descriptor files. *Commun. ACM,* 23(9):522-528, Sept. 1980.

[POLL 85] Pollard, Carl. J. and Lewis G. Creary. A Computational Semantics for Natural Language. In *Proc. 23rd Ann. Mtg. ACL,* 8-12 July 1985, 172-179.

[QUIL 68] Semantic Memory. In *Semantic Information Processing,* Marvin Minsky (ed.), Cambridge, Massachusetts: MIT Press, 1968.

[RAGH 86] Raghavan, Vijay V. and S.K.M. Wong. A Critical Analysis of Vector Space Model for Information Retrieval, *J. Am. Soc. Inf. Sci.*, 37(5):279-287, Sept. 1986.

[RAMA 85] Ramamohanaro, Kotagiri and John Shepherd. A Superimposed Codeword Indexing Scheme for Very Large Prolog Databases. Tech. Report 85/17, Dept. of Comp. Sci., Univ. of Melbourne, 1985.

[RIEG 81] Rieger, C. and S. Small. Toward a Theory of Distributed Word Expert Natural Language Parsing. *IEEE Trans. on Systems, Man, and Cybernetics,* SMC-11(1):43-51, Jan.1981.

[RIES 82] Riesbeck, C.K. Realistic Language Comprehension. In *Strategies for Natural Language Processing*, ed. by Wendy G. Lehnert and Martin H. Ringle, Lawrence Erlbaum Assoc., Hillsdale NJ, 435-454, 1982.

[SACK 85] Sacks-Davis, Ron. Performance of a multi-key access method based on descriptors and superimposed coding techniques. *Inform. Systems,* 10(4), 391-403, 1985.

[SAGE 75] Sager, N. Sublanguage Grammars in Science Information Processing. *J. Am. Soc. Inf. Sci.*, 26(1):10-16, Jan.-Feb. 1975.

[SAGE 81] Sager, N. *Natural Language Information Processing*. Addison-Wesley, New York, 1981.

[SALT 72] Salton, G. A New Comparison Between Conventional Indexing (Medlars) and Text Processing (SMART). *J. Am. Soc. Inf. Sci.*, 23(2):75-84, 1972.

[SALT 80] Salton, G. The SMART System 1961-1976: Experiments in Dynamic Document Processing. In *Encyclopedia of Library and Information Science*, 1-36, 1980.

[SALT 83] Salton, G. and M.J. McGill. *Introduction to Modern Information Retrieval.* McGraw-Hill, New York, 1983.

[SALT 86] Salton, G. Another Look at Automatic Text-Retrieval Systems. *Commun. ACM*, 29(7): 648-656, July 1986.

[SELF 86] Selfridge, Mallory. Integrated Processing Produces Robust Understanding. *Comp. Ling.*, 12(2): 89-106, April-June 1986.

[SHER 74] Sherman, D. A New Computer Format for *Webster's Seventh Collegiate Dictionary.* In *Computers and the Humanities,* 8:21-26, 1974.

[SIMM 83] Simmons, Robert F. A Text Knowledge Base for the AI Handbook. Univ. of Texas at Austin Dept. of Comp. Sci., Technical Report TR-83-24, Dec. 1983.

[SIMM 84] Simmons, R.F. *Computations from the English.* Prentice-Hall, Englewood Cliffs, NJ, 1984.

[SPAR 71] Sparck Jones, K. *Automatic Keyword Classifications*. Butterworths, London, 1971.

[SPAR 73] Sparck Jones, K. and Martin Kay. Linguistics and Information Science, Academic Press, New York, 1973.

[SPAR 84] Sparck Jones, K. and J.I. Tait. Automatic Search Term Variant Generation. *J. Doc.*, 40(1):50-66, March 1984.

[SVEN 86] Svenonius, Elaine. Unanswered Questions in the Design of Controlled Vocabularies, *J. Am. Soc. Inf. Sci.*, 37(5):331-340, Sept. 1986.

[TENO 84] Tenopir, Carol. Full-Text Databases. *ARIST*, 19:215-246, 1984.

[THOM 85] Thompson, R.H. and W.B. Croft. An Expert System for Document Retrieval. *Proc. Expert Systems in Gov. Symp.*, IEEE, 448-456, Oct. 1985.

[TONG 83] Tong, R.M. et al. A Rule-Based Approach to Information Retrieval: Some Results and Comments. *Proc. AAAI-83,* 411-415, 1983.

[WALK 85] Walker, Donald E. Knowledge Resource Tools for Accessing Large Text Files. Proc. *First Conference of the UW Centre for the New Oxford English Dictionary: Information in Data.* Nov. 6-7, 1985, Waterloo, Canada, 11-24.

[WANG 85] Wang, Y.-C., J. Vandendorpe, and M. Evens. Relational Thesauri in Information Retrieval. *J. Am. Soc. Inf. Sci.*, 36(1): 15-27, Jan. 1985.

[WEYE 82] Weyer, S.A. The Design of a Dynamic Book for Information Search. In *International Journal of Man Machine Studies*, 17(1): 87-107, July 1982.

[WEYE 84] Weyer, S.A., and A.H. Borning. A Prototype Electronic Encyclopedia. In *ACM Trans. on Office Information Systems*, 3(1): 63-68, January 1984.

[WHIT 83] White, C. The Linguistic String Project Dictionary for Automatic Text Analysis. In Proc. *Workshop on Machine Readable Dictionaries,* SRI, Menlo Park, CA, May 1983.

[WILE 84] Wilensky, R., Arens, Y., and D. Chin. Talking to UNIX in English: An Overview of UC. *Commun. ACM*, 27(6):574-593, 1984.

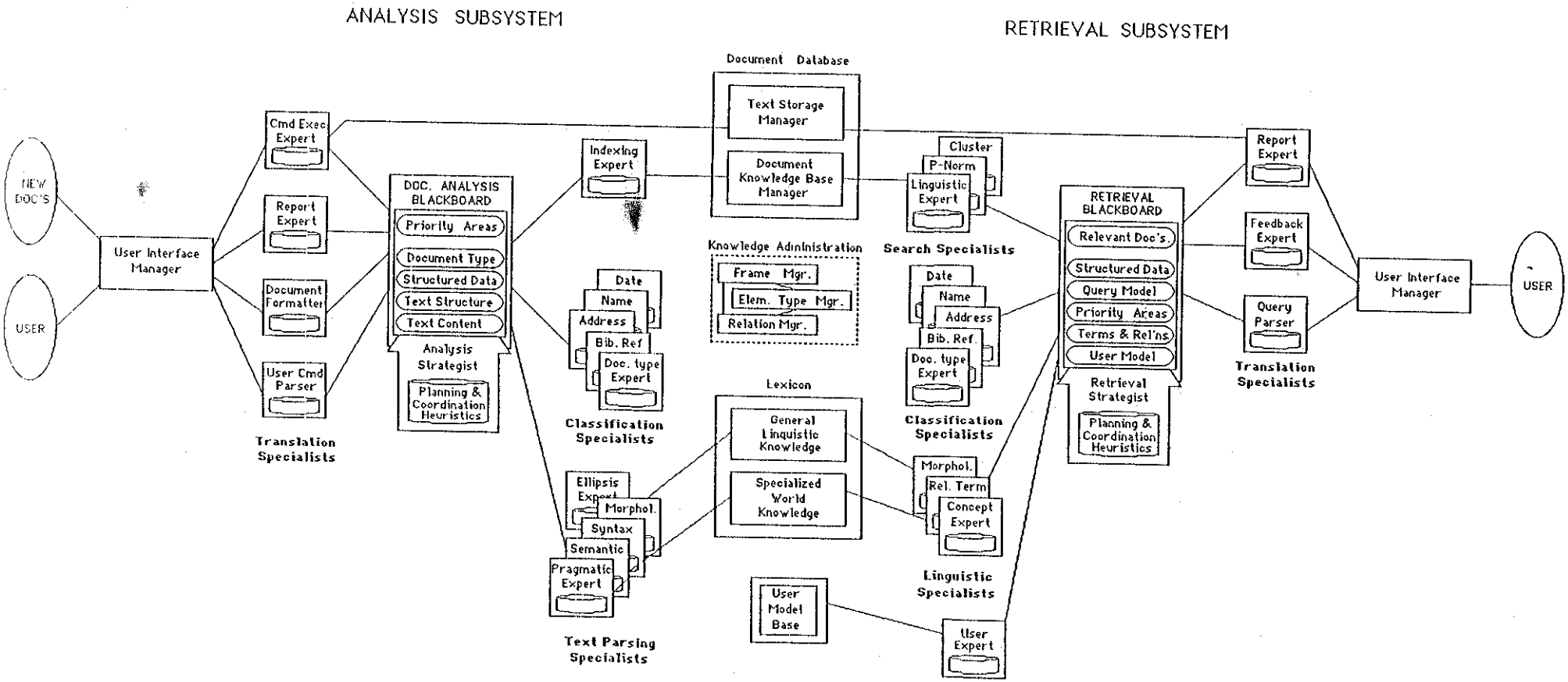
[WILL 85a] Williams, Martha E. Electronic Databases. *Science* 228(4698): 445-456, 26 April 1985.

[WILL 85b] Williams, Martha E., ed. *Computer-Readable Databases.* American Library Assoc., Chicago 1985.

[WILL 86] Williams, Martha E. Transparent Information Systems Through Gateways, Front Ends, Intermediaries, and Interfaces. *J. Am. Soc. Inf. Sci.* 37(4), 204-214 (July 1986).

[WILLP 85] Williams, Phil W. How Do We Help the End User? Proc. *6th National Online Meeting*, April 30-May 2, 1985, 495-505.

[WOHL 86] Wohlwend, Robert C. Creation of a Prolog Fact Base from the Collins English Dictionary. MS Report, VPI&SU Computer Science Dept., Blacksburg, VA, March 1986.



Overview of the CODER System
Figure 3

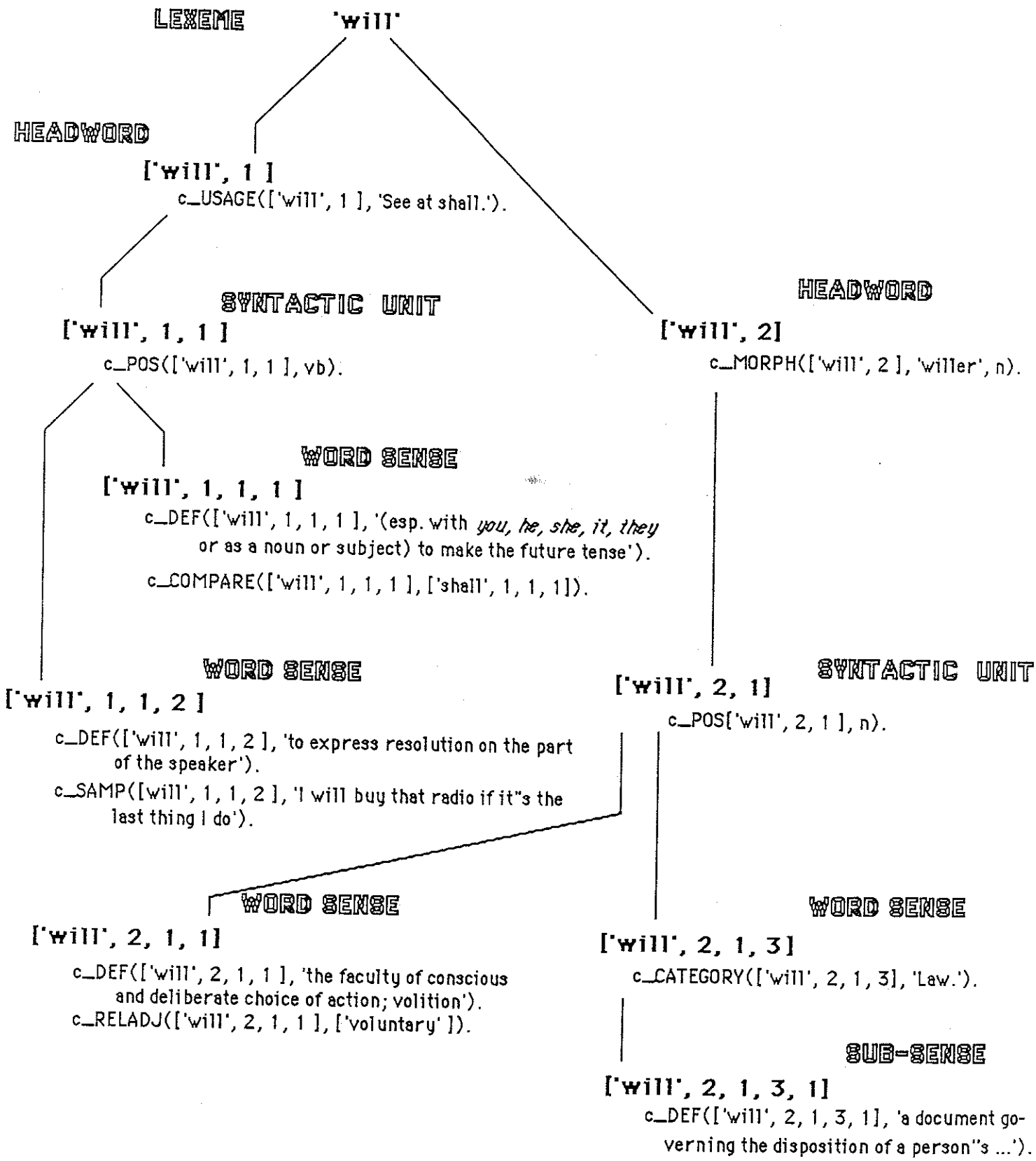


Fig. 4: Hierarchial Structure of the CDEL

c_ABBREV	-	Abbreviation of headword.
c_ALSO_CALLED	-	Headword is also commonly called this.
c_CATEGORY	-	Category (semantic label) of headword.
c_COMPARE	-	Compare to another headword and sense(s).
c_DEF	-	Definition of the headword.
c_DEF_NUM	-	Number of (up to) 80-character blocks of the definition.
c_HEADWORD	-	Headword entry.
c_MORPH	-	Morphological variant of headword (including part of speech).
c_NLAST	-	Rest (e.g. first/middle name) of a proper noun headword.
c_PAST	-	Past form of headword.
c_PLURAL	-	Plural of headword (sometimes just the ending).
c_POS	-	Part of speech.
c_RELADJ	-	Related adjective to headword.
c_SAMP	-	Example of headword in context.
c_SINGULAR	-	Singular form of headword (sometimes just the ending).
c_SYLL	-	Syllabification of headword.
c_USAGE	-	Usage notes providing guidance on usage of headword.
c_USAGE_NUM	-	Number of (up to) 80-character blocks in the usage note.
c_VAR_SPELL	-	Variant spelling(s) (if any).
c_VAR_SYLL	-	Syllabification of variant spelling(s).

Table 1: Relations abstracted from the *CDEL* tapes.

c_ABBREV ([word, homnum, defnum], relword).

c_ALSO_CALLED ([word, homnum, defnum], relword).

c_CATEGORY ([word, homnum, defnum], relword).

c_COMPARE ([word, homnum, defnum], [relword, homnum, posnum]).

c_DEF ([word, homnum, defnum, subnum], relword, blocknum).

c_DEF_NUM ([word, homnum, defnum, subnum], relword, totalblocks).

c_HEADWORD ([word, homnum]).

c_MORPH ([word, homnum], relword).

c_NLAST ([word, homnum, defnum], relword, pos).

c_PAST ([word, homnum, defnum], relword).

c_PLURAL ([word, homnum], relword).

c_POS ([word, homnum], pos).

c_RELADJ ([word, homnum, defnum], [relword, homnum]).

c_SAMP ([word, homnum, defnum, subnum], relword).

c_SINGULAR ([word, homnum, defnum], relword).

c_SYLL ([word, homnum], syllword).

c_USAGE ([word, homnum], relword, blocknum).

c_USAGE_NUM ([word, homnum], relword, totalblocks).

c_VAR_SPELL ([word, homnum], relword).

c_VAR_SYLL ([word, homnum], relword).

where:

word	=	'lexeme' (in single quotes)
homnum	=	headword (integer)
relword	=	'word' or 'phrase' (in single quotes)
posnum	=	syntactic level (integer)
pos	=	part of speech (atomic element, no quotes)
blocknum	=	number of current block (integer)
totalblocks	=	total number of clocks to contain text (integer)
syllword	=	'syllabified word' (syllables connected by (_))
defnum	=	sense level (integer)
subnum	=	sub-sense level (integer)

Table 2: Prolog Syntax of CDEL Relations.

	CDEL			W7		
ITEM	FREQ	FREQ/HD	% USAGE	FREQ	FREQ/HD	% USAGE
Definitions	161693	1.98	36.8	140500	2.04	45.0
Parts Of Speech	96218	1.18	21.9	67809	0.99	21.7
Headwords	81561	1.00	18.6	68766	1.00	22.0
Morphological Variants	27219	0.33	6.2	9957	0.14	3.2
Categories	26765	0.33	6.1	11990	0.17	3.8
Sample Usages	16779	0.21	3.8	–	–	–
Plurals	9366	0.11	2.1	6320	0.09	2.0
Comparisons	7301	0.09	1.7	4025	0.06	1.3
Alternative Names	4562	0.06	1.0	572	0.01	0.2
Variant Spellings	4540	0.06	1.0	2460	0.04	0.8
First Names	2445	0.03	0.6	–	–	–
Abbreviations	525	0.01	0.1	–	–	–
Related Adjectives	141	*	x	–	–	–
Usage Notes	84	*	x	–	–	–
Singular Forms	43	*	x	–	–	–
Past Forms	7	*	x	–	–	–
TOTAL	439494	5.39	100.0	312399	4.54	100.0

Table 3: Comparative Frequencies of Items in CDEL and W7

CATEGORY	FREQ	CATEGORY	FREQ
Informal	2457	Linguistics	93
Archaic	1531	History	90
Brit.	1162	Property law	90
U. S.	1056	Bible	86
Slang	796	Statistics	83
Rare	780	Meteorol.	78
Chiefly U. S.	757	Judaism	76
Law	717	Brit. dialect	75
Music	596	Commerce	70
Obsolete	550	Psychiatry	70
Nautical	547	Finance	67
Chiefly Brit.	488	New Testament	67
Physics	457	Chess	65
Maths.	411	Economics	63
Pathol.	408	Geom.	62
Anatomy	379	Formal	61
Austral.	369	Now rare	60
Biology	354	Bridge	59
Chem.	348	Golf	58
Botany	345	Canadian	56
Greek myth.	296	Derogatory	56
Med.	276	Surgery	56

Table 4: Frequencies of Categories used in CDEL

CATEGORY	FREQ	CATEGORY	FREQ
French	254	Angling	53
Grammar	226	Tennis	50
Brit. informal	221	Ecology	49
Printing	218	Genetics	48
Psychol.	214	Theatre	48
Military	199	Psychoanal.	47
Old Testament	186	Cards	46
U.S. slang	186	Taboo slang	46
R.C. Church	182	Astrology	43
Poetic	177	Criminal law	43
Austral. informal	175	Films	43
U.S. informal	175	Hinduism	43
Zoology	175	Metallurgy	43
Brit. slang	172	Chiefly R.C. Church	42
Logic	172	Chiefly Scot.	42
Literary	164	German	42
Phonetics	161	Mining	41
Philosophy	160	Vet. science	40
dialect	159	Rhetoric	38
Trademark	154	Billiards	37
Austral. slang	151	Norse myth.	37
Architecture	146	Archaeology	36
Electronics	144	Boxing	36

Table 4 (cont'd)

CATEGORY	FREQ	CATEGORY	FREQ
Scot.	144	Chiefly Austral.	36
Christianity	143	Caribbean	35
Cricket	141	Baseball	34
Astronomy	131	English history	34
Physiol.	127	S. African	34
Latin	121	Soccer	34
Geology	120	Stock Exchange	34
Prosody	120	Not standard	33
Sport	120	Classical myth.	32
Computer technol.	107	Rugby	32
Photog.	103	Sociol.	32
Northern Brit.	101	Ecclesiast.	30
Heraldry	97	Embryol.	30
Theol.	95	Surveying	30
Biochem.	94		

Table 4 (cont'd)

	CDEL		W7	
Homograph #	FREQ	FREQ/HD	FREQ	FREQ/HD
1	79222	0.97	60079	0.87
2	1928	0.02	6542	0.10
3	313	*	1427	0.02
4	81	*	475	0.01
5	15	*	164	*
6	2	*	54	*
7	–	–	17	*
8	–	–	5	*
9	–	–	3	*
TOTAL	81561	1.00	68766	1.00

Table 5: Comparative Frequencies of Homograph Numbers in CDEL and W7

	CED		W7		MWPD	
PART OF SPEECH	FREQ	% USE	FREQ	% USE	FREQ	% USE
noun	61820	64.2	42610	62.8	15166	58.8
plural_noun	1226	1.3	-	-	-	-
noun_total	63046	65.5	42610	62.8	15166	58.8
verb_transitive	7078	7.4	4976	7.3	3	x
verb	4303	4.5	2425	3.6	5029	19.5
verb_intransitive	3247	3.4	1363	2.0	-	-
verb_imperative	-	-	10	x	-	-
verbal_auxiliary	-	-	4	x	-	-
verb_impersonal	-	-	2	x	-	-
verb_phrase	-	-	-	-	2	x
verb_total	14628	15.2	8780	12.9	5034	19.5
adjective	12661	13.2	13435	19.8	4677	18.1
adjective_postpos	291	0.3	-	-	-	-
adjective_prenominal	82	0.1	-	-	-	-
adjective_total	13034	13.5	13435	19.8	4677	18.1
combining_form	503	0.5	516	0.8	-	-
noun_combining_form	134	0.1	150	0.2	-	-
adj_combining_form	44	x	-	-	-	-
adv_combining_form	1	x	-	-	-	-
verb_combining_form	1	x	2	x	-	-

Table 6: Part of Speech Entries for CDEL, W7, and WPD

	CED		W7		MWPD	
PART OF SPEECH	FREQ	% USE	FREQ	% USE	FREQ	% USE
combining_form_total	683	0.7	668	1.0	-	-
suffix_forming_nouns	83	0.1	107	0.2	36	0.1
suff_form_pl_prop_nns	6	x	1	x	-	-
suffix	60	0.1	2	x	1	x
suffix_forming_adjs	30	x	50	0.1	26	0.1
suffix_forming_advs	5	x	7	x	4	x
suffix_forming_verbs	-	-	12	x	10	x
suffix_form_interjs	-	-	1	x	-	-
suffix_total	184	0.2	180	0.3	77	0.3
abbreviation	2423	2.5	-	-	-	-
adverb	1247	1.3	1468	2.2	549	2.1
interjection	336	0.3	94	0.1	22	0.1
preposition	130	0.1	164	0.2	120	0.5
prefix	116	0.1	74	0.1	6	x
determiner	96	0.1	-	-	-	-
pronoun	85	0.1	108	0.2	82	0.3
conjunction	62	0.1	96	0.1	55	0.2
symbol_for	61	0.1	-	-	-	-
sentence_connector	42	x	-	-	-	-
sentence_substitute	40	x	-	-	-	-
connecting_vowel	2	x	-	-	-	-
sentence_modifier	1	x	-	-	-	-

Table 6 (cont'd)

	CED		W7		MWPD	
PART OF SPEECH	FREQ	% USE	FREQ	% USE	FREQ	% USE
modifier	1	x	-	-	-	-
trademark	-	-	121	0.2	-	-
indefinate_article	-	-	3	x	-	-
definate_article	-	-	2	x	2	x
overall total	96242	100.0	67809	100.0	25790	100.0

- no occurrences.
x less than .05

Note: Sum of primary and secondary parts of speech is used for W7 data.

Table 6 (cont'd)

	CED			W7		
SENSE	FREQ	FREQ/HD	% USE	FREQ	FREQ/HD	% USE
1	94780	1.16	61.9	82604	1.20	58.8
2	29249	0.36	19.1	33867	0.49	24.1
3	11841	0.15	7.7	11956	0.17	8.5
4	5788	0.07	3.8	5204	0.08	3.7
5	3309	0.04	2.2	2682	0.04	1.9
6	2117	0.03	1.4	1559	0.02	1.1
7	1407	0.02	0.9	872	0.01	0.6
8	984	0.01	0.6	532	0.01	0.4
9	718	0.01	0.5	335	*	0.2
10	553	0.01	0.4	233	*	0.2
11	423	0.01	0.3	183	*	0.1
12	340	*	0.2	143	*	0.1
13	274	*	0.2	97	*	0.1
14	221	*	0.1	66	*	x
15	178	*	0.1	47	*	x
16	147	*	0.1	29	*	x
17	121	*	0.1	22	*	x
18	92	*	0.1	19	*	x
19	77	*	0.1	13	*	x
20	67	*	x	5	*	x

Table 7: Comparative Frequencies of Sense Numbers in CDEL and W7

	CED			W7		
SENSE	FREQ	FREQ/HD	% USE	FREQ	FREQ/HD	% USE
21	53	*	x	9	*	x
22	42	*	x	5	*	x
23	35	*	x	5	*	x
24	31	*	x	4	*	x
25	27	*	x	1	*	x
26	25	*	x	1	*	x
27	18	*	x	-	-	-
28	15	*	x	-	-	-
29	15	*	x	-	-	-
30	12	*	x	-	-	-
31	11	*	x	-	-	-
32	11	*	x	-	-	-
33	9	*	x	-	-	-
34	9	*	x	-	-	-
35	7	*	x	-	-	-
36	7	*	x	-	-	-
37	7	*	x	-	-	-
38	6	*	x	-	-	-
39	5	*	x	-	-	-
40	5	*	x	-	-	-

Table 7 (cont'd)

	CED			W7		
SENSE	FREQ	FREQ/HD	% USE	FREQ	FREQ/HD	% USE
41	4	*	x	-	-	-
42	3	*	x	-	-	-
43	3	*	x	-	-	-
44	3	*	x	-	-	-
45	3	*	x	-	-	-
46	3	*	x	-	-	-
47	3	*	x	-	-	-
48	1	*	x	-	-	-
49	1	*	x	-	-	-
50	1	*	x	-	-	-
51	-	-	-	-	-	-
totals	153061	1.88	100.0	140493	2.04	100.0

- no occurrences
* less than 0.005
x less than 0.05

Table 7 (cont'd)

	CED			W7		
SUB-SENSE #	FREQ	FREQ/HD	% USE	FREQ	FREQ/HD	% USE
1	4945	0.06	47.3	16659	0.24	42.2
2	4950	0.06	47.4	16578	0.24	42.0
3	470	0.01	4.5	4307	0.06	10.9
4	58	*	0.6	1214	0.02	3.1
5	18	*	0.2	409	0.01	1.0
6	3	*	x	153	*	0.4
7	-	-	-	65	*	0.2
8	-	-	-	30	*	0.1
9	-	-	-	12	*	x
10	-	-	-	10	*	x
11	-	-	-	4	*	x
12	-	-	-	4	*	x
13	-	-	-	1	*	x
14	-	-	-	2	*	x
15	-	-	-	-	-	-
totals	10444	0.13	100.0	39448	0.57	100.0

- no occurrences
x less than .05
* less than .005

Table 8: Comparative Frequencies of Sub-sense Numbers in CDEL and W7

Vowels

CDEL	CDEL Prolog Fact File	CDEL	CDEL Prolog Fact File
i	i	ɛ	$e
I	$I	æ	@@
u:	u$:	ʌ	^
a:	$@$:	ɐ	$A
ɔ:	@$	v	$u
a	@	e	e
ə	@	θ	#Gg

Consonants

CDEL	CDEL Prolog Fact File	CDEL	CDEL Prolog Fact File
p	p	t	t
k	k	l	l
d	d	g	g
tʃ	t#s	dʒ	d#f
m	m	n	n
ŋ	&n	r	r
b	b	f	f
v	v	s	s
z	z	ʃ	#s
ʒ	#f	h	h
j	j		

Stress

CDEL	CDEL Prolog Fact File	CDEL	CDEL Prolog Fact File
!	'	~	,

Table 9: Pronunciation Codes in Dictionary and Computer File

Conventions

Capitalized abbreviations are non-terminals.
Punctuation marks in single quotes are terminal characters.

The following abbreviations are used:

ADVDEF = whole definition

USAGE = usage information

ONEDEF = segment of whole definition set off by ';' or '!'

STR, appended to any other variable, indicates a string of terms of that variable type.

P, appended to any other variable indicates a phrase of that variable type.

ADJ = adjective
ADV = adverb
DET = determiner
N = noun
P = preposition
COMP = term used for comparison
V = verb primitive
VINF = infinitive
VING = present participle
VED = past participle
LM = left-hand modifier string
RM = right-hand modifier string
QUANT = quantifier
SUBCONJ = subordinating conjunction
CLAUSE = string with subject and/or verb

* = 0 or more repetitions
() = optional term
| = "or", for alternate terms
_____ = any string of characters

Grammar

```
ADVDEF --> (USAGE) (ONEDEF (';' ONEDEF | '!' ONEDEF)*) ('.') (USAGE)

USAGE  --> '(' (often) VED|foll'.'|VINGP  PP ')'
       --> '(' (esp'.') PP)  ONEDEF  '(' (esp'.')  PP ')'
       --> CAPITAL_LETTER_____'.'
       --> '(' intensifier ')'
       --> ':' _____
       --> _____':'
```

Table 10: Grammar for CDEL Adverb Definitions

--> _____'.'

--> '(' foll'.' by NP|PREPSTR ')'

--> '(' NP|ADJP|ADVP|VINFP|VEDP|VP ')'

--> used in NP

--> used to V NP

--> used with NP

--> as modifier

--> a less common spelling|word of|for ADV

--> a variant (spelling) of ADV

--> a euphemistic word for ADV#

--> a euphemism for NP

--> a parenthetic filler used VINFP

--> that|as

--> ',' esp'.' VINGP

--> an archaic word for ADV

--> an emphatic form of ADV

--> an exclamation used to ONEDEF

--> an informal word for ADV

--> an intensive form of ADV

--> an obsolete word for ADV

--> another term|word ('('_____')') for ADV (',' usually VEDP)

--> dialect

--> short for NP

--> the comparative of ADV

--> the superlative of ADV

Table 10 (cont'd)

```
ONEDEF --> null
       --> PP ((',') PP)* ((',') ADVP)
       --> PP (or) VEDP
       --> PP VINGP
       --> PP or PP (ADVP)
       --> (PP)* and NP
       --> ADVP
       --> ADVP or PP
       --> ADJP (PP)*
       --> NP
       --> VINGP (SUBCONJ) (VINGP)*
       --> VINFP
       --> VP
       --> VEDP

PP  -->  PREPSTR
    -->  PREPSTR NP (PREPSTR) (',' NP (PREP))* ((',') or NP (PREP))
             (VEDP)
    -->  PREPSTR VINGP (or VINGP)
    -->  PREPSTR ADVP (NP)
    -->  PREPSTR or COMP SUBCONJ
    -->  PREPSTR LM PREPSTR NP
    -->  PREPSTR ADVP PREPSTR
    -->  PREPSTR or VINGP

ADJP  -->  (ADVSTR)* ADJSTR (PREPSTR)
      -->  ADVSTR VEDP
```

Table 10 (cont'd)

```
ADVP  -->  ADVSTR (PREPSTR) (VEDP|ADJP|ADVSTR) (VINGP) (VINFP)
             (or ADVSTR (PP))* (PP)* (NP)

VP  -->  VINGP VSTR
    -->  VSTR (PP) (ADJP) (NP)
    -->  or VSTR PSTR


VEDP  -->  (ADVSTR) VEDSTR (ADVSTR or (PP)*) (ADVP) (PP)* (VINGP)
             (ADVP) (NP)

VINGP  -->  (ADVP) VING (ADVP) (NP) (VEDP)
       -->  VING ADJSTR VING|VINF (PP) (ADVP)
       -->  VING (VP) PP
       -->  VING NP PP ADVP
       -->  VING SUBCONJ

VINFP  -->  VINFSTR (ADVSTR) (ADJSTR) (NP) (VEDP) (PP)*

NP  -->  CLAUSE
    -->  (LM) NSTR (RM)

CLAUSE  -->  NSTR|SUBCONJ V (ADVSTR) (NP) (VINFP) (that)
        -->  SUBCONJ ADJP
        -->  SUBCONJ NP
        -->  V N ADVP PP
```

Table 10 (cont'd)

```
LM  -->  (QUANT) (DET) (QUANT) (ADJP|VEDP) (USAGE)

RM  -->  (ADVP)* (CLAUSE) (PP)
    -->  VEDP
    -->  (PP)* (VINF)
    -->  VINGP


PREPSTR --> PREP (PREP)
        --> PREP (PREP) or PREP (PREP)
        --> PREP (PREP) (',' PREP (PREP))* ',' or PREP (PREP)
        --> PREP and PREP

ADJSTR --> ADJ
       --> ADJ (',' ADJ)* ',' or ADJ
       --> ADJ or ADJ
       --> ADJ and ADJ


ADVSTR --> ADV (ADV)* ((',' ADV)* ',') (or ADV (',' etc ('.')))
       --> QUANT|ADJ COMP
       --> as ADJ|ADV as
       --> ADV and ADV

VEDSTR --> VED (or|but (ADV) VED)
       --> VED (',' VED)* ',' or VED

NSTR  -->  N ((',' N)* ',' etc (.))
```

Table 10 (cont'd)

```
VSTR  -->  V (',' V)* ((',') or V)

VINF -->  to V


QUANT --> all|every|such|any|what|many|much
SUBCONJ --> if|while|which|when|that
DET --> a|an|the|what
COMP --> than|as|so
ADV --> not|many|much|more|most|so|as|when {and all -ly words and others
          defined as adverbs}
ADJ --> more|most|time|many|much| {and all words defined as adjectives}
VED --> withdrawn|dealt|cut| {all past participles of verbs}
PREP --> aboard|at|of|to|by| etc.
N --> best|time|distance|degree| etc.
V --> bake|be|come|run|say|have|is| etc.
VED --> covered|directed|placed|used| etc.
VING --> being|coming|indicating|sailing|using| etc.
CAPITAL_LETTER --> A..Z
```

Table 10 (cont'd)

defining formula	possible type	complement type
PP		
aboard	location	null|NP = location
across	direction	NP|null
after	time/location	NP = time or location
along	location	NP = location
as a|an	manner	NP = a similar thing
as if	manner	VED = manner
as well	transition	null
as ADVP as	degree/manner	ADJ|NP|VP|ADVP = degree or limit
at	location/time/ manner/state transition/ direction/ degree	ADVP = manner NP|ADVP|NP than VED
before	time/location	NP = time boundary
behind	location/state	NP = location or ref. to state ("time")
beneath	location	NP = location
between	time/location	NP = range of time or distance
by	manner or degree	NP | VINGP | or PP = means or degree
by and large	transition	null
by or before	time	NP
by or up to	degree	NP
during	time	(or PREP) NP = span
except for	state(?)	null
for	purpose/state/ time	NP|ADV (NP = context, ADV = state)
for, to, or from	degree (allot- ment)	NP = source or recipi- ent
from	time or direction(ab)	NP
from NP|ADJP to NP	direction/time	NP(','ADVP) = location or time span
in	location/ state/time/ manner	NP (PP)* | NP of
in any case	transition	null

Table 11: Defining Formulas in CDEL Adverb Definitions

Note: Abbreviations here follow conventions of Table 10

in or during	time	NP = time span
in accordance with	manner	NP
in addition to that	transition	null
in, at, or to	location	NP = location
in, from, or towards this direction	direction	null (headword = direction)
in or into	state/location	NP (adj's = SYNs)
in or to|towards	location/ degree	NP = location or degree
in a|an|the ADJP manner|way|fashion	manner	null
in the manner of NP	manner	NP
in, to, towards, or from	direction	NP = direction
into	state/direction	NP = state or place
like	manner	NP = similar item
of course	transition	null
of or concerning	usage	NP = context
of or for	location	NP = source or recipient
on	state/manner/ location	NP | null
on or by	manner	NP = means
on or outside	location	NP = location
on or to	location	NP = location
on the other hand	transition	null
out of	state or	NP|VING|PP
outside	location	NP = not location
over	location	NP = spread of area
throughout	time/location	NP = time span or area
so as	manner	VINFP
to	direction/ location/ degree/state	(ADV) NP
to all appearances	transition	null
to all practical purposes	transition	null
to or in|into	state/degree location /	NP = location or degree
to or towards	time/ manner	NP = location goal
towards	direction	NP = location goal
towards or onto|on	direction	NP = location goal
up	direction	NP|PP|ADVP (NP = location; PP|ADVP=time)
with reference to	manner/usage	NP = context
with respect to	manner/usage	NP = context
with regard to	manner/usage	NP = context
with	manner/state	NP (VED) (PREP) (w/o NP = ANT)

Table 11 (cont'd)

within	location	NP = location
without	manner/state	NP (w NP = ANT)
VED		
covered	state	PP
directed	state	PP
placed or kept	location	ADV | PP (PP)*
used	usage info.	INF NP (NP = SYN)
used in	usage info.	NP to indicate NP (last NP = meaning)
VING		
being	state	NP | ADJ , VINGP
coming	time	ADV = time
continuing	direction	PP
floating	state	PP | null
following	time/degree	NP
having	manner	NP | VED or (PP)*
indicating	state / direction	NP
making	manner	NP
occurring	time/location	(PP)*
remaining	location	PP (location = same as before)
using	manner	NP = means
ADJ		
active and prominent	state	PP|NP
free	state	PP (state = w/o NP)
separate	location	PP
together	state	PP
ADV		
any	degree/time	ADV|PP = range
away	location/direc	PP
every	time	NP
more than	state/degree	ADV
near in	state/degree	NP = measurement
not	negation	most anything
*ly	state	VED or ADJ
VINF		
to be	state/manner	VEDP|NP|ADJP|PP (VED = SYN, last ADV = degree)
NP		
all in all	transition	null

Table 11 (cont'd)