Machine Tractable Dictionaries
as Tools and Resources
for Natural Language Processing

Yorick WILKS, Dan FASS, Cheng-ming GUO,
James E. MCDONALD, Tony PLATE, and Brian M. SLATOR

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

ABSTRACT

This paper discusses three different but related large-scale computational methods for the transformation of machine readable dictionaries (MRDs) into machine tractable dictionaries, i.e., MRDs converted into a format usable for natural language processing tasks. The MRD used is *The Longman Dictionary of Contemporary English.*

1 Introduction

Machine readable dictionaries (MRDs) contain knowledge about language and the world essential for tasks in natural language processing (NLP). However, this knowledge, collected and recorded by lexicographers for human readers, is not presented in a principled enough manner for MRDs to be used directly as tools for such tasks. What is badly needed is machine **tractable** dictionaries (MTDs): MRDs transformed into a format usable for NLP tasks.

This paper discusses three different but related large-scale computational methods for the transformation of MRDs into MTDs. The MRD used is *The Longman Dictionary of Contemporary English* (LDOCE). The three approaches differ in the amount of knowledge they start with and the kinds of knowledge they produce. All begin with some hand-coding of initial information but are largely automatic. Approach I, a connectionist approach, uses the least hand-coding but then generates data for the co-occurrence of words, which is the simplest form of semantic information produced by any of the approaches. Approach II requires the hand-coding of a grammar and semantic patterns used by its parser, but not the hand-coding of any lexical material. This is because the approach builds up lexical material from sources wholly within LDOCE. Approach III employs the most hand-coding because it develops and builds lexical entries for a very carefully controlled defining vocabulary of 3,600 word senses (1,200 words). The payoff is that the approach will produce a MTD containing highly structured semantic information.

The three approaches are all processes: tools for transforming MRDs into MTDs. Such tools will be applicable to MRDs other than LDOCE. The products of these tools are MTDs which are resources useful not just for NLP tasks but for artificial intelligence (AI) generally.

2 Background: The Value of Machine Readable Dictionaries

Dictionaries are language texts whose subject matter is language. The purpose of dictionaries is to provide definitions of senses of words and, in so doing, they supply knowledge about not just language, but the world. For years, researchers in computational linguistics (CL) and AI have viewed dictionaries (a) with theoretical interest as a means of investigating the semantic structure of natural language, and (b) with practical interest as a resource for overcoming the knowledge acquisition bottleneck in AI. The knowledge acquisition bottleneck has been viewed by most researchers as too hard a problem to tackle at present. However, more optimistic researchers have recently begun to seek methods to overcome it, and have had some success. This difference in attitudes regarding the knowledge acquisition bottleneck is reflected in a long-standing difference between two alternative methods of lexicon building: the demo approach and the book approach (Miller 1985; a similar distinction appears in Amsler 1982).

The demo approach, which has been the dominant paradigm in natural language processing (and AI in general) for the last two decades, does not address the knowledge acquisition bottleneck. This approach is to hand-code a small but rich lexicon for a system that analyses a small number of linguistic phenomena. This is an expensive method as each entry in the lexicon is prepared individually. Every entry is constructed with foreknowledge of its intended use and hence of the knowledge it should contain. Being designed with only a specific purpose in mind, the knowledge representation runs into problems when scaled up to cover additional linguistic phenomena.

The alternative, the book approach, confronts the problem of the knowledge acquisition bottleneck. This approach attempts to develop methods for transforming the knowledge within dictionaries or encyclopaedias into some format usable for CL and AI tasks, usually with the aim of covering as large a portion of the language as possible. The problem with dictionary and encyclopaedia entries is that, although they are constructed in a principled manner over many years by professional lexicographers and encyclopaedists, they are designed for human use.

Recently, interest in the book approach has greatly expanded because a number of MRDs have become available, each causing a flurry of research interest, e.g., *The Merriam-Webster New Pocket Dictionary* (Amsler and White 1979; Amsler 1980, 1981), *Webster's Seventh New Collegiate Dictionary* (Evens and Smith 1983; Chodorow, Byrd, and Heidorn 1985; Markowitz, Ahlswede, and Evens 1986; Binot and Jensen 1987), and *The Longman Dictionary of Contemporary English* (Michiels, Mullenders, and Noel 1980; Michiels and Noel 1982; Walker and Amsler 1986; Boguraev, Briscoe, Carroll, Carter, and Grover 1987; Boguraev and Briscoe 1987; Wilks, Fass, Guo, McDonald, Plate, and Slator 1987).

The big advantage of MRDs is that now both theoretical and practical concerns are investigable by large-scale computational methods. Some of the above research has been into the underlying semantic structure of dictionaries (e.g., Amsler and White 1979; Amsler 1980, 1981; Chodorow, Byrd, and Heidorn 1985). The remainder of the research has been seeking to develop practical large-scale methods to extract syntactic information from MRD entries (e.g., Boguraev and Briscoe 1987) and transform that information into a format suitable for other users. This latter research has the effect of transforming a MRD into a limited MTD. We use the word "limited" because such a MTD has only syntactic information presented in a format usable by others; semantic information remains buried in the MRD though this semantic information is the knowledge about language and the world that is needed as a resource for many CL and AI tasks. The next step is therefore to develop large-scale methods to extract both the syntactic and semantic information from MRD entries and present that information as a data base of acceptable format for potential users of that information.

Within the book approach there are a number of ways one can construct such a MTD. One way is to automatically extract the semantic information and build the MTD. We firmly advocate automatic extraction. A second way is to extract the semantic information manually and hand-code the entire MTD, as is being attempted in the CYC Project (Lenat, Prakash, and Shepherd 1986; Lenat and Feigenbaum 1987). The main problem with this approach is the volume of effort required: the CYC Project aims to hand-code one million entries from an encyclopaedia, which will take an estimated two person-centuries of work. We believe this is a mistaken approach because it wastes precious human resources and makes dubious theoretical assumptions, despite Lenat's claims that their work is theory free (see section 5).

Which ever form of the book approach is taken, there are two sets of issues that must be faced by those developing methods for the transformation of MRDs into MTDs. One set of issues concerns the nature of the knowledge in MRDs. The second set of issues concerns the design of the database format of a MTD. Both sets of issues rest on understanding the structure and content of the knowledge that is both explicitly and implicitly encoded in dictionaries, but such understanding rests on certain key issues in semantics. We examine some of these issues in the next section.

3 Background: The State of Semantic Theory

There are obstacles to the development of methods (whether manual or automatic) for the transformation of the semantic information from MRDs into a MTD that have not been present for those developing methods for syntactic analysis only. The main obstacle is that, compared to syntactic theory, understanding of semantic theory is much less advanced, as shown by the lack of consensus about even some of the general underlying principles of semantics. Nevertheless there is some understanding and some local consensus on semantics that can allow work to proceed.

Positions on certain basic issues in semantics affects one's stance concerning what semantic information should be extracted from a MRD and represented in a MTD. In developing our own methods for the transformation of MRDs into MTDs, we have adopted a certain approach from computational semantics. Examples of this approach are Preference Semantics (Wilks 1973, 1975a, 1975b) and Collative Semantics (Fass 1986, 1987, 1988). The main assumptions of this approach are the inescapable problem of the **word sense** and **the inseparability of knowledge and language.**

Lexical ambiguity is pervasive in language: the lexical ambiguity of words has been a problem since before the advent of dictionaries and is particularly apparent when translating between languages; tasks such as translation cannot be modelled by computer without some representation of lexical ambiguity. Furthermore, lexical ambiguity is pervasive in most forms of language text, including dictionary definitions: the words used in dictionary definitions of words and their senses are themselves lexically ambiguous and must be disambiguated.

We also believe that it is acceptable for a semantics to be based on the approach to lexical ambiguity taken by traditional lexicography that constructs dictionaries. The major problem with the approach comes from its arbitrariness in the selection of senses for a word. This arbitrariness appears between dictionaries in different sense segmentations of the same word. It is also observable within a single dictionary when the sense-distinctions made for the definition of a word do not match with the uses of that word in the definitions of other words in the dictionary. These problems can be overcome by methods that reconcile different sense selections of a word within and across dictionaries by extending (or reducing) the coverage of individual word senses.

Our position on the inseparability of knowledge and language is that common principles underlie the semantic structure of language text and of knowledge representations, and that some kinds of language text structures are a model for knowledge structures (Wilks 1978). Examples of such knowledge structures include the planes of Quillian's Memory Model (1967, 1968), pseudo-texts from Preference Semantics, sense-frames from Collative Semantics, and integrated semantic units or ISUs (Guo 1987). Supporting evidence comes from comparisons between the semantic structure of dictionaries and the underlying organisation of knowledge representations, which have observed similarities between them (Amsler 1980; Chodorow, Byrd, and Heidorn 1985).

These positions on semantics suggest the following for those engaged in transforming MRDs into MTDs. First, the problem of lexical ambiguity must be faced by any approach seeking to extract semantic information from a MRD to build a MTD. Because lexical ambiguity exists in the language of dictionary definitions and in language generally, it follows that the language in MRD definitions needs to analysed to the word sense level and must be represented in the format of the MTD. Second, the format of the MTD, while being of principled construction, should be as language-like as possible.

Next, we focus attention onto some basic issues in transforming MRDs concerning the nature and accessibility of the knowledge in dictionaries.

4 The Analysis of MRDs

We hold that those who advocate the extraction (both manual and automatic) of semantic information from dictionaries (and even encyclopaedias) have made certain assumptions about the extent of knowledge in a dictionary, about where that knowledge is located, and how that knowledge can be extracted from the language of dictionary definitions. These are not assumptions about semantics, but rather, are assumptions about the extraction of semantic information from text. These assumptions are **methodological** assumptions because they underlie the decisions made in choosing one method for semantic analysis rather than another. These assumptions are about (a) **sufficiency**, (b) **extricability**, and (c) **bootstrapping.**

Sufficiency addresses whether a dictionary is a strong enough knowledge base for English, specifically as regards the linguistic knowledge and, above all, the knowledge of the real world needed for subsequent text analysis. Sufficiency is of general concern, including hand-coding projects like CYC, where they attempt to make explicit (a) the facts and heuristics which one would need in order to understand sentences, (b) generalisations of those facts and heuristics, and (c) inferences that fill inter-sentential gaps (Lenat and Feigenbaum 1987, p.1180).

Extricability is concerned with whether it is possible to specify a set of computational procedures that operate on a MRD and extract, through their operation alone and without any human intervention, general and reliable semantic information on a large scale, and in a general format suitable for, though independent of, a range of subsequent text analysis processes.

Bootstrapping refers to the process of collecting the initial information that is required by a set of computational procedures that are able to extract semantic information from the sense definitions in an MRD. The initial information needed is commonly linguistic information, notably syntactic and case information, which is used during the parsing of sense-definitions into an underlying representation from which semantic information is then extracted.

Bootstrapping methods can be **internal** or **external.** Internal bootstrapping methods obtain the initial information needed for their procedures from the dictionary itself and use the procedures to extract that information. This is not as circular as it may seem. A process may require information for the analysis of some sense-definition (e.g., some knowledge of the words used in the definition) and may be able to find that information elsewhere in the dictionary. By contrast, external bootstrapping methods obtain the initial information for their procedures by some method other than the use of those procedures. The initial information may be from some source external to the dictionary or may be in the dictionary but impossible to extract without the use of the very same information. For example, the word 'noun' may have a definition in a dictionary but the semantic information in that definition cannot be extracted without prior knowledge of a sentence grammar that contains knowledge of syntactic categories, including what a noun is.

Those who advocate hand-coding presumably have pessimistic views about extricability and bootstrapping.

5 The Production of MTDs

The main issue here concerns the format that MTDs should have. One thing is clear: the format must be versatile if a variety of consumers in CL and AI are to use it. The most likely initial consumers are those that place a considerable emphasis on the availability of words, such as spelling correction, and those that already use large lexicons, such as machine translation (MT) and word processing (Amsler 1982). Within CL, two primary consumers are the semantics mentioned in section 3, Preference Semantics and Collative Semantics.

These consumers need a variety of semantic information. To meet these needs MTD formats should be clean, unambiguous, preserve much of the semantic structure of natural language, and contain as much information as is feasible. However, this does not mean that the format of a MTD must consist of just a single type of representation because it is possible that different kinds of information require different types of representation. For example, two kinds of information about word use are: (a) the use of senses of words in individual dictionary sense definitions, and (b) the use of words throughout a dictionary. It is not clear that a single representation can record both (a) and (b) because (a) requires a frame-like representation of the semantic structure of sense definitions that records the distinction between genus and differentia, the subdivision of differentia into case roles, and the representation of sense ambiguity, whereas (b) requires a matrix or network-like representation of word usages that encodes the frequency of occurrence of words and the frequency of co-occurrence of combinations of words. Hence, a MTD may consist of several representations, each internally uniform.

Given the arguments presented in section 3, we believe that the first of these representations should be modelled on natural language though it should be more systematic and without its ambiguity. Hence, this component representation should be as language text-like as possible and should represent word senses, whether explicitly or implicitly.

Other approaches to the building of representations that contain semantic information extracted from dictionaries and encyclopaedias (e.g., Binot and Jensen 1987; Pustejovsky and Bergler 1987; CYC) separate knowledge and language and overlook the problem of the lexical ambiguity of the words in dictionary definitions (these are the underlying theoretical assumptions made by these approaches).

The other representation form of representation can be construed as a connectionist network representation, based on either localist (e.g., Cottrell and Small 1983; Waltz and Pollack 1985) or distributed approaches to representation (e.g., Hinton, McClelland and Rumelhart 1986; St.John and McClelland 1986). Like our position on semantics, connectionism emphasises the continuity between knowledge of the language and the world and many connectionist approaches have paid special attention to representing word senses, especially the fuzzy boundaries between them (e.g., Cottrell and Small 1983; Waltz and Pollack 1985; St.John and McClelland 1986). Localist approaches assume symbolic network representations whose nodes are word senses and whose arcs are weights that indicate the relatedness of the word senses at the ends of the arcs. An interesting new approach, which we shall outline shortly in section 6.1, uses a network whose nodes are words and whose arc weights indicate co-occurrence data between words. Although this approach initially appears to be localist, it is being used to derive more distributed representations which offer ways of avoiding some serious problems inherent in localist representations. Such frequency-of-association data is not represented in standard knowledge representation schemes, is complementary to the knowledge in such schemes, and may be useful in its own right for CL tasks such as lexical ambiguity resolution and spelling correction.

To summarise so far, we have outlined: (1) some basic theoretical assumptions about semantics and our position regarding those assumptions (inseparability of language and knowledge, tackling the problem of the word sense), (2) some basic methodological assumptions about the extraction of semantic information from dictionaries (sufficiency, extricability, bootstrapping), and (3) some basic theoretical assumptions regarding the format of a MTD (language-like format, inclusion of different kinds of semantic information, notably lexical ambiguity).

6 Three Approaches to the Transformation of MRDs into MTDs

At CRL, we are pursuing three approaches to the automatic translation of the information in *The Longman Dictionary of Contemporary English* (Proctor et al 1978) into a MTD. LDOCE is a full-sized dictionary designed for learners of English as a second language that contains over 55,000 entries in book form and 41,100 entries in machine-readable form (a type-setting tape). The preparers of LDOCE claim that entries are defined using a "controlled" vocabulary of about 2,000 words and that the entries have a simple and regular syntax. We have analysed the machine-readable tape of LDOCE and found that about 2,219 words are commonly used.

The three CRL approaches all fall within the general position on computational semantics outlined above and are extensions of fairly well established lines of research. All three approaches also pay special attention to their underlying methodological assumptions concerning the extraction of semantic information from dictionaries. With respect to sufficiency and extricability, all three approaches assume that dictionaries do contain sufficient knowledge for at least some CL applications, and that such knowledge is extricable. But the approaches differ over bootstrapping, i.e., over what knowledge, if any, needs to be hand-coded into an initial analysis program for extracting semantic information.

The three approaches differ in the amount of knowledge they start with and the kinds of knowledge they produce. All begin with a degree of hand-coding of initial information but are largely automatic. In each case, moreover, the degree of hand-coding is related to the source and nature of semantic information sought by the approach. Approach I, a connectionist approach, uses the least hand-coding but then the co-occurrence data it generates is the simplest form of semantic information produced by any of the approaches. Approach II requires the hand-coding of a grammar and semantic patterns used by its parser, but not the hand-coding of any lexical material. This is because the approach builds up lexical material from sources wholly within LDOCE. Approach III employs the most hand-coding because it develops and builds lexical entries for a very carefully controlled defining vocabulary of 3,600 word senses (1,200 words). The payoff is that the approach will produce a MTD containing highly structured semantic information.

6.1 Approach I: Obtaining and Using Co-Occurrence Statistics from LDOCE (Tony Plate)

Our first approach extracts semantic information from text (specifically LDOCE) that does not require any semantic information to bootstrap it. Central to this technique is that all sentences that contain a word are used as sources of information about the use of that word, rather than just the definition of the word. This technique is based on some experimental findings that the frequency of co-occurrence of a pair of words provides a reasonable measure of the strength of the semantic relationship between them.

This approach bears some resemblance to Sparck Jones's (1964) investigation into the semantic classification of the uses of words. Her underlying linguistic assumption was that the uses of words may be distinguished, described, or analyzed by the semantic relations which hold between them and the vocabulary of a language has a semantic structure determined by these relations. Of twelve possible semantic relations, synonymy was chosen as the fundamental feature of natural language.

Despite some surface similarities to Sparck Jones's technique there are many differences, some of which are discussed below. First, Sparck Jones's data collection method is much more laborious than the co-occurrence method (see Wilks, Fass, Guo, McDonald, Plate, and Slator 1987).

Second, Sparck Jones's method requires that the data must contain all the senses of words that need to be considered. In the co-occurrence method, it is not necessary that the text contain examples of all senses, because the sense definitions are used to provide information about the senses. The text need only use enough senses of words to define all words, but should make frequent use of the senses it does use.

The approach proposed here finds much more distant and general relationships than synonymy, and which involve the combination of many semantic relations. Co-occurrence data for the LDOCE controlled vocabulary has been collected. This data contains nearly two-and-a-half million frequencies of co-occurrence (the triangle of a 2200 by 2200 matrix). This is too much data to examine in raw form, so we have used two techniques to convert the data into a more understandable format.

We have written a program called BROWSE which can manipulate the entire co-occurrence matrix and can select groups of words based on whether the values of various probability functions pass selected thresholds. These groups of words can be manipulated as sets, and one technique we are using is to build sets of words that are either related to a certain word or to a certain sense of a word.

The other technique involves using BROWSE to extract submatrices which are then given to the PATHFINDER program (Schvaneveldt, Durso, and Dearholt 1985). This program was designed to discover the network structure of psychological data and it reduces the total amount of information while not eliminating much of useful information. We have applied this program to LDOCE co-occurrence data with some success; it produces sparsely connected networks which are easy to examine by eye and which appear to contain much useful world knowledge.

In both formats (groups of words and PATHFINDER networks) the data is a potentially useful resource for a number of applications. Of particular interest is the possibility of sense disambiguation. To investigate this, we have written a number of processes that use the co-occurrence data. One process we are studying involves rating the coherence of particular sense assignments for sentences, based on the set overlap of the groups of words related to each of the assigned senses. Another process we are studying is how activation spreading from the nodes in a network produced by the PATHFINDER program can select the appropriate senses of words in context.

The work has strong links to connectionism, and indeed we are investigating how this work can proceed within the connectionist paradigm. We are developing a theory of representation, utilisation, and learning of networks within distributed connectionist models. In addition, we have been developing a connectionist simulator for the Intel hypercube; this work is well under way (see Plate 1987).

6.2 Approach II: A Lexicon-Producer (Brian M. Slator)

While the first approach begins with no prior knowledge needed at all, the other two approaches begin with certain kinds of external information supplied. The second approach (Slator and Wilks 1987; Slator 1988) hand-codes a grammar, some semantic patterns, and a list of the 2,219 words of the LDOCE controlled vocabulary. The approach seeks to build dictionary entries for the words of the controlled vocabulary and the other words in LDOCE using semantic information extracted from not only the dictionary entries of LDOCE, as in the other two approaches, but also from the box and pragmatic codes found on the machine readable version of LDOCE (though not in the book). The box codes use a special set of primitives such as "abstract," "concrete," and "animate," organised into a type hierarchy. The primitives are used to assign type restrictions on nouns and adjectives, and type restrictions on the arguments of verbs. The pragmatic codes (also called "subject" codes but referred to here as "pragmatic" codes to avoid confusion with the grammatical subject) use another special set of primitives organised into a hierarchy. The hierarchy consists of main headings such as "engineering" and subheading like "electrical." The primitives are used to classify words by their subject, for example, one sense of 'current' is classified as "geography:geology" while another sense is marked "engineering/electrical."

The semantic information is extracted from LDOCE dictionary entries using a large-scale parser that isolates the genus and differentia terms in each entry, expanding upon other similar work (e.g., Chodorow, Heidorn, and Byrd 1985; Alshawi, Boguraev, and Briscoe 1985; Boguraev and Briscoe 1987; Binot and Jensen 1987).

The dictionary entries that are built for individual word senses are frame-based lexical semantic structures, intended for subsequent use in knowledge based parsing. The process of building a frame for a word sense begins by first assigning the box and pragmatic code information from LDOCE for that word sense. The parser then analyses the definition of that word sense from LDOCE.

The parser is a chart parser (taken from Slocum 1985) which is left-corner and bottom-up with top-down filtering and early constituent tests. Chart parsing was selected because of its utility as a grammar testing and development tool. The parser is driven by a context free grammar of over 100 rules and a lexicon composed of the 2,219 words from the LDOCE controlled vocabulary. It must be emphasised that this chart parser is not a parser for English -- it is a parser for just the language of LDOCE definitions. The grammar is still being tuned, but currently covers over 90% of the language used in LDOCE definitions of content words.

The parser produces a phrase-structure tree of an LDOCE word sense definition which is passed to an interpreter for pattern matching and inferencing. The interpreter extracts the dominating phrase, reorganises the phrase into genus and differentia components, and attempts to infer and fill in case roles that subdivide the differentia information. The interpreter then accesses the pre-existing frame for the word sense, which already contains the relevant box and pragmatic code information for the word sense, and enriches the frame by adding the genus and differentia information extracted from its definition.

Consider, for example, how the frame for 'ammeter' is built. From the box and pragmatic codes, the following hierarchical information is extracted and used to create an initial frame for 'ammeter': from the box code, that an ammeter is of type "solid," and from the pragmatic code, that an ammeter is classified under the subject "engineering/electrical."

Next, the chart parser is used to analyse the LDOCE definition of an 'ammeter', which is that it "is an instrument for measuring ... electric current." The definition is parsed into a phrase-structure tree which is passed to the interpreter. The interpreter adds to the frame for 'ammeter' that 'instrument' is its genus and "for measuring electric current" is its differentia information. Furthermore, the interpreter notes the phrase "for measuring" and creates the case role slot PURPOSE, i.e., that the purpose of an ammeter is for measuring electric current.

6.3 Approach III: Building a MTD from a Key Defining Vocabulary (Cheng-ming Guo)

The third approach, unlike the first and the second, argues that a small amount of hand-coding of world knowledge is necessary before the bootstrapping process can begin. The amount of hand-coding required, though more than the other approaches, is still relatively small because over 95% of its MTD is built automatically. The prior world knowledge that requires hand-coding is a set of 1,200 words, called the Key Defining Vocabulary (KDV), which has been found to define the controlled vocabulary of LDOCE, and thence all 27,758 words defined in LDOCE. The senses of all the words in LDOCE can be defined by the KDV in a series of four "defining cycles."

When a candidate word enters a defining cycle, the stems of the words used in the definitions of the first three senses of that candidate word are examined. If all the word stems in those three sense definitions occur in the KDV, then the candidate word is put into a "success" file and added to the KDV at the end of the defining cycle; if not, the word is put into a "fail" file and addition of the word to the KDV is postponed until a later cycle. In this way, the size of the KDV expands with each cycle until, after three cycles, all the words from the LDOCE controlled vocabulary are accounted for. The remaining words in LDOCE is expected to be defined in the next defining cycle.

The discovery of the KDV and the use of defining cycles is valuable for a number of reasons. First, in building a MTD, a KDV reduces the initial number of knowledge structures for dictionary entries that have to be hand-coded before such structures can be constructed automatically by some bootstrapping process. The knowledge structures used in this particular study are called "integrated semantic units" or ISUs. Though the preliminary study reported here uses a KDV of around 1,200, the number can probably be reduced to about 1,000.

Second, the use of defining cycles helps to identify vacuous circular definitions. Circular definitions that use circles of just two words pose special problems for building a MTD from a MRD. For example, in LDOCE a "trip" is defined as a "journey", and a "journey" as a "trip". A MTD built from a MRD should be free of such circular definitions. One way to overcome such circular definitions is to try and include just one of the words involved as a KDV word, but not the other. The word selected for the KDV will be the one whose first three senses fulfil the criteria of a defining cycle given earlier.

Thirdly, when constructing a MTD, use of the defining cycles ensures that all definitions of words and their senses that are built contain only words that already have definitions. In the case of LDOCE, use of the defining cycles sorts out words in the LDOCE controlled vocabulary whose definitions include words outside of that vocabulary. This has proved to be not uncommon in LDOCE definitions.

Fourthly, in building a MTD, the main senses of these empirically found KDV words are taken as the "semantic primitives" of the MTD. The use of defining cycles ensures that a set of primitives that best suit a particular MRD can be found empirically.

An estimated 3,600 ISUs for an average of three basic senses of the 1,200 KDV words are to be hand-coded to start the bootstrapping process (the bootstrapping process is shown schematically in Figure 2 of the Appendix, p.14). A language analyzer and learner (LAAL) carries out the bootstrapping process according to a bootstrapping schedule (as with approach II, any grammar rules or semantic patterns used by the LAAL will have to be hand-coded). The bootstrapping schedule is concerned with which word senses are to be processed first, and which later. The necessity for the bootstrapping schedule stems from the fact that the ISUs for the basic senses of the words in the definition of a word sense have to be in the ISU database before that definition can be analysed and its ISU produced. After the ISUs for the basic word senses of the words from the LDOCE controlled vocabulary are built into the database, the non-basic senses of these words will be processed. When all of the controlled vocabulary words are finished, words from outside the controlled vocabulary will be attended to. Following the bootstrapping schedule, the LAAL system processes word sense definitions to produce more and more ISUs until the entire LDOCE is turned into a full MTD of ISUs.

Further details about the three approaches may be found in the Appendix (Wilks, Fass, Guo, McDonald, Plate, and Slator 1987). The MTDs produced by these approaches are fed into a number of consumers: a Lexicon-Consumer (Slator and Wilks 1987) and Collative Semantics.

7 Summary

We do not expect to produce a single format for representing the knowledge extracted from LDOCE because the three approaches use different sources of knowledge and different processes. The formats produced by approaches II and III are notationally the most alike but the knowledge they contain is different. Unlike the others, the format of approach I contains co-occurrence data. The format of II contains box and pragmatic code information not present in the format of approach III; but the underlying organisation of the knowledge in approach III is very systematic, unlike the equivalent knowledge in approach II. We expect that the comparison of formats will be very fruitful, as will the comparison of underlying approaches to the extraction of semantic information, and will produce clearer understanding for future work on transforming MRDs into MTDs.

8 References

Alshawi, Hiyan, Bran Boguraev, and Ted Briscoe (1985). Towards a Dictionary Support Environment for Real Time Parsing. In *Proceedings of the European Conference on Computational Linguistics*, Pisa, Italy.

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

Amsler, Robert A. (1981) A Taxonomy of English Nouns and Verbs. In *Proceedings of the 19th Annual Meeting of the Association for Computational Linguistics*, Stanford, Ca, pp.133-138.

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

Amsler, Robert A. (1986) Deriving Lexical Knowledge-Base Entries from Existing Machine-Readable Information Sources. Unpublished Ms.

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

Binot, Jean-Louis, and Karen Jensen (1987) A Semantic Expert Using an Online Standard Dictionary. *Proceedings of the 10th International Joint Conference on Artificial Intelligence (IJCAI-87)*, Milan, Italy, pp.709-714.

Boguraev, Bran, and Ted Briscoe (1987) Large Lexicons for Natural Language Processing: Exploring the Grammar Coding System of LDOCE. *Computational Linguistics*, 13.

Boguraev, Branimir K., Ted Briscoe, John Carroll, David Carter, and Claire Grover (1987) The Derivation of a Grammatically Indexed Lexicon from the Longman Dictionary of Contemporary English. *Proceedings of the 25th Annual Meeting of the ACL*, Stanford University, Stanford, CA, pp.193-200.

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

Cottrell, Garrison W., and Steven L. Small (1983) A Connectionist Scheme for Modelling Word-Sense Disambiguation. *Cognition and Brain Theory*, 6, pp. 89-120.

Evens, M., and R.N. Smith (1983) Determination of Adverbial Senses from Webster's Seventh Collegiate Definitions. Paper presented at Workshop on Machine Readable Dictionaries, SRI-International, April 1983.

Fass, Dan C. (1986) Collative Semantics: An Approach to Coherence. Memorandum in Computer and Cognitive Science, MCCS-86-56, Computing Research Laboratory, New Mexico State University, New Mexico.

Fass, Dan C. (1987) Semantic Relations, Metonymy, and Lexical Ambiguity Resolution: A Coherence-Based Account. In *Proceedings of the 9th Annual Cognitive Science Society Conference*, University of Washington, Seattle, Washington, pp.575-586.

Fass, Dan C. (1988) Collative Semantics: A Semantics for Natural Language Processing. Memorandum in Computer and Cognitive Science, MCCS-88-118, Computing Research Laboratory, New Mexico State University, New Mexico.

Guo, Cheng-ming (1987) Interactive Vocabulary Acquisition in XTRA. *Proceedings of the 10th International Joint Conference on Artificial Intelligence (IJCAI-87)*, Milan, Italy, pp.715-717.

Lenat, Douglas, B., Mayank Prakash, and Mary Shepherd (1986) CYC: Using Common Sense Knowledge to Overcome Brittleness and Knowledge Acquisition Bottlenecks. *AI Magazine*, 7, (4), pp.65-85.

Lenat, Douglas, B., and Edward A. Feigenbaum (1987) On The Thresholds of Knowledge. *Proceedings of the 10th International Joint Conference on Artificial Intelligence (IJCAI-87)*, Milan, Italy, pp.1173-1182.

Hinton, Geoff E., James L. McClelland, and David E. Rumelhart (1986) Distributed Representations. In James L. McClelland, David E. Rumelhart, and the PDP Research Group (1986) *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Volume 2: Psychological and Biological Models*, MIT Press/Bradford Books: Cambridge, MA, chapter 3, pp.77-109.

McClelland, Jay, David E. Rumelhart, and the PDP Research Group (1986) *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Volume 2: Psychological and Biological Models*, MIT Press/Bradford Books: Cambridge, MA.

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

Michiels, A., J. Mullenders, and J. Noel (1982) Exploiting a Large data Base by Longman. *Proceedings of the 8th International Conference on Computational Linguistics (COLING-80)*, Tokyo, Japan, pp.374-382.

Michiels, A., and J. Noel (1982) Approaches to Thesaurus Production. *Proceedings of the 9th International Conference on Computational Linguistics (COLING-82)*, Prague, Czechoslovakia, pp.227-232.

Miller, George A. (1985) Dictionaries of the Mind. In *Proceedings of the 23rd Annual Meeting of the Association for Computational Linguistics*, Chicago, pp.305-314.

Plate, Tony (1987) HYCON: A Design for the Simulation of Connectionist Models on Coarse Grained Parallel Computers. Memorandum in Computer and Cognitive Science, MCCS-87-106, Computing Research Laboratory, New Mexico State University, New Mexico.

Procter, Paul, Robert F. Ilson, John Ayto, et al (1978). *Longman Dictionary of Contemporary English*, Longman Group Limited: Harlow, Essex, England.

Pustejovsky, James, and Sabine Bergler (1987) The Acquisition of Conceptual Structure for the Lexicon. *Proceedings of the 6th National Conference on Artificial Intelligence (AAAI-87)*, Seattle, Wa., pp.556-570.

Quillian, M. Ross (1967) Word Concepts: A Theory and Simulation of Some Basic Semantic Capabilities. *Behavioral Science*, 12, pp.410-430. Also reprinted in Ronald J. Brachman and Hector J. Levesque (Eds.) (1985) *Readings in Knowledge Representation*, Morgan Kaufmann: Los Altos, CA, pp.98-118.

Quillian, M. Ross (1968) Semantic Memory. In Marvin Minsky (Ed.) *Semantic Information Processing*, Cambridge, Mass: MIT Press, pp.216-270.

St.John, Mark F., and James L. McClelland (1986) Reconstructive Memory for Sentences: A PDP Approach. Ohio University Inference Conference.

Schvaneveldt, Roger W., Frank T. Durso, and Don W. Dearholt (1985) Pathfinder: Scaling with Network Structure. Memorandum in Computer and Cognitive Science, MCCS-85-9, Computing Research Laboratory, New Mexico State University, New Mexico.

John Sinclair, Patrick Hanks, Gwyneth Fox, Rosamund Moon, Penny Stock, et al (1987) *Collins COBUILD English Language Dictionary*, William Collins and Sons: Glasgow, Scotland.

Slator, Brian M. (1988) Lexical Semantics and a Preference Semantics Parser. Memorandum in Computer and Cognitive Science, MCCS-88-116, Computing Research Laboratory, New Mexico State University, New Mexico.

Slator, Brian M. and Yorick A. Wilks (1987) Toward Semantic Structures from Dictionary Entries. In *Proceedings of the Second Annual Rocky Mountain Conference on Artificial Intelligence*, Boulder, Colorado, pp.85-96. Also Memorandum in Computer and Cognitive Science, MCCS-87-96, Computing Research Laboratory, New Mexico State University, New Mexico.

Slocum, Jonathan (1985) Parser Construction Techniques: A Tutorial. Tutorial held at the 23rd Annual Meeting of the Association for Computational Linguistics, Chicago.

Sparck Jones, Karen (1964) Synonymy and Semantic Classification. Ph.D. Thesis, University of Cambridge, England. Published in Edinburgh Information Technology Series (EDITS), Sidney Michaelson and Yorick A. Wilks (Eds.), Edinburgh University Press: Edinburgh, Scotland, 1986.

Walker, Donald E., and Robert A. Amsler (1986) The Use of Machine-Readable Dictionaries in Sublanguage Analysis. In Ralph Grishman and Richard Kittredge (Eds.) *Analyzing Language in Restricted Domains*, Lawrence Erlbaum: Hillsdale, NJ.

Waltz, David L., and Pollack, Jordan B. (1985) Massively Parallel Parsing: A Strongly Interactive Model of Natural Language Interpretation. *Cognitive Science*, 9, pp.51-74.

Wilks, Yorick A. (1973) An Artificial Intelligence Approach to Machine Translation. In Roger C. Schank and Kenneth M. Colby (Eds.) *Computer Models of Thought and Language*, San Francisco: W.H. Freeman, pp.114-151.

Wilks, Yorick A. (1975a) A Preferential Pattern-Seeking Semantics for Natural Language Inference. *Artificial Intelligence*, 6, pp.53-74.

Wilks, Yorick A. (1975b) An Intelligent Analyser and Understander for English. *Communications of the ACM*, 18, pp.264-274.

Wilks, Yorick A. (1978) Making Preferences More Active. *Artificial Intelligence*, 11, pp.197-223.

Wilks, Yorick A., Dan C. Fass, Cheng-ming Guo, James E. McDonald, Tony Plate, and Brian M. Slator (1987) A Tractable Machine Dictionary as a Resource for Computational Semantics. Memorandum in Computer and Cognitive Science, MCCS-86-105, Computing Research Laboratory, New Mexico State University, New Mexico. To appear in Bran Boguraev and Ted Briscoe (Eds.) *Computational Lexicography for Natural Language Processing*, Longman: Harlow, Essex, England.