EXTRACTING SEMANTIC INFORMATION FROM ON-LINE SOURCES

Yael Ravin

IBM T.J. Watson Center
Yorktown Heights, New York 10598
U.S.A.

ABSTRACT: In order to achieve broad-coverage semantic capabilities, natural language applications require large lexical databases. This paper describes the work of the Lexical Systems Group at IBM, whose goal is to create such a database, by using (semi-)automatic methods and tools, to extract semantic information from machine-readable dictionaries and thesauri: type-setting tapes are turned into lexical databases representing meaning in hierarchical tree-structures; polysemous terms are disambiguated and indexed with their intended senses; and semantic relations are extracted from the dictionary. The paper concludes with some applications and open problems.

I. INTRODUCTION

In contrast with syntax, computational semantics has not yet achieved broad coverage. Most systems that perform semantic analysis are still restricted to one or few domains for which they are specifically developed. The challenge for the next decade, as we see it at IBM, is to develop systems capable of broad-coverage semantics. The availability of broad-coverage semantics will increase the number of general applications which our systems will be able to address. At the moment, with syntax alone, we have developed one product - a writer's aid called **CRITIQUE**[18,38,39]. With broad-coverage semantics we will also be able to tackle translation of general texts and creation of abstracts, improve the accuracy of information retrieval and enrich the writer's aid program.

Developing broad-coverage semantics may prove quite challenging. The main difficulty is that the field of theoretical semantics is not as developed as that of syntax.

There is less consensus among theoreticians about the nature of the phenomena under study and about the form of representation they should take[11,22,31,37,40]. The question of "what is meaning" is still alive in both linguistics and philosophy in a way that "what is grammar" has never been.

The various theories are incompatible on many accounts but there is one assumption which most of them share - the compositionality of meaning. Intuitively, the meaning of paragraphs is composed of the meanings of their sentences, which in turn, are composed of the meanings of their lexical items - words and idioms. Therefore, an important part of any semantic theory is the semantics of individual lexical items. If we are interested in achieving broad-coverage semantics, we should also develop a large database of lexical items. The creation of such a database is our goal at the Lexical Systems Group at IBM and the topic of this paper. Section II discusses the nature of the lexical database that we are proposing. Sections III to VII are each devoted to one component of the database and represent the work of various members of the Lexical Systems Group.[1]

II. A BROAD-COVERAGE LEXICAL DATABASE

Rather than adopt a definitive theoretical position, we attempt to build a lexical database in small steps, opting for practical solutions as we encounter problems, keeping the various semantic theories in mind, and, most importantly, testing our success against the usefulness the database will have for applications, specifically for machine translation. Whatever theoretical position we will end up adopting, we are presently led by pragmatic decisions and their consequences.

Our major pragmatic decision has been to derive our lexical information from existing lexical resources such as thesauri, dictionaries and text corpora.[2] Good reasons for relying on these resources are that they are compiled by lexicographers - experts in the meaning of words, who also keep track of how words are used in various contexts, through citations sent in by readers or found in corpora. Existing lexical resources have stood the test of time, being edited and revised periodically and thus relatively

---

[1] Further references partaining to each of these projects are found at the end of the relevant sections.

[2] The work on text corpora is not discussed in this paper. For a discussion see [21].

consistent and complete. We have neither the experience of lexicographers nor their means for the labor-intensive task of compiling lexical information on our own. Another important factor is that most of these resources are available in machine-readable form, as type-setting tapes, and thus most suitable for the creation of an on-line database. We are fortunate at IBM to have access to many monolingual, bilingual and learners' dictionaries, as well as thesauri and encyclopedias [12-17,28,30,44].

Our decision to use available lexicographic resources and to rely on the lexicographic judgement of their compilers commits us to accept the content of theses sources as given. Indeed, we are very reluctant to change anything in them other than obvious misspellings and other typographical errors. The content of dictionaries and thesauri determines for us the nature of lexical meaning and its scope. We look at the nature of dictionaries and thesauri as consisting of networks of words, or more accurately, word senses, captured in textual form and in natural language. Every word-sense defined in these sources (i.e. every headword) can be said to stand in some relation to every other word(-sense) mentioned in its entry. Thus senses are interconnected by a variety of semantic (and other) relations, such as synonymy, hypernymy, predicate-argument association, and translation equivalence. Our task becomes to translate this network with its relations from natural language text into a more computationally conventional representation. The task is not easy, first and foremost because of the way information is presented in lexicographical sources.

There are several difficulties associated with using publishers' type-setting tapes as sources of information. Published dictionaries are written with severe space constraints to reduce costs. Their information is highly condensed, elliptical and often implicit. It is meant to be read and relies on visual clues, such as font changes and relative placement of various pieces of information. Dictionaries rely on the human mind's ability to handle scope ambiguity and ellipsis. All these must be changed for computer use. We discuss the parsing of type-setting tapes into expanded and explicit texts in Section III.

Another major difficulty with information in dictionaries and thesauri stems from the fact that it is encoded in natural language, which is highly ambiguous. Much of the ambiguity is due to the polysemy of words. The relations found in our sources appear as relations among words, although they are really meant to be relations among word-senses. It is our task to disambiguate the words, that is, to identify for each

word, its intended sense; and index all words that appear in the source with their appropriate sense number. We discuss the disambiguation of a thesaurus in Section IV; the disambiguation of the text of dictionary definitions in Section V; and the disambiguation of hypernyms in Section VI.

After the text of the type-setting tape has been parsed and after the terms appearing in the text have been disambiguated, the next step is to extract the semantic relations embodied in the text. In section VI we discuss the extraction of the hypernym relation as an example. Finally, we manipulate the network connections in order to gain some further semantic knowledge. A set of heuristics for classifying word-senses into semantic categories, labelled by binary features is discussed in Section V.

A guiding principle in our work is to minimize the encoding of information by hand. We strive to perform our lexical tasks of parsing, disambiguating and extracting information automatically. For this purpose, we have been building various software tools that are general enough to operate on our various sources, despite differences in content and format. The use of automatic tools is not only an attempt to save labor; it is also reflective of our approach to lexical information. The use of tools allows us to perform the same operations again and again, thus refining our emerging database with time. As we learn more about the nature of our material, we modify our parsers and heuristics and run them again to produce better results. [2,4,5]

III. DICTIONARY ENTRY PARSING

Machine-readable dictionaries come from publishers in the form of type-setting tapes - that is, flat character streams containing lexical data interspersed with special characters that control font changes. These flat streams have to be analyzed and then converted into complex, hierarchical structures suitable for a lexical database. The tools which are built to analyze the character streams should be general so that they can be used for any of the various type-setting tapes, with their different content and formatting conventions. They should also be specific enough in order to fully analyze all the information found in each tape. This double goal is achieved by having a general parser - the Dictionary Entry Parser (**DEP**) - run with one of several grammars, each of which captures the conventions of a particular dictionary.

The Structure of Type-setting Tapes

The structure of entries in type-setting tapes is quite complex. Font codes divide an entry into different fields of information, such as pronunciation, definition, grammatical category, style and use comments. The fields are highly compacted with a variety of abbreviation devices. Recurring identical elements are often omitted in order to save space. For example:

---

in.cu.ba.tor ... a machine for **a** keeping eggs warm until they HATCH
b keeping alive babies that are too small to live ...

Figure 1. Definition of "incubator" in Longman's Dictionary of Contemporary English (LDOCE) [29]

Here the initial part of the definition text pertains to both definitions **a** and **b**. Elements may have a double function. The capitalization of HATCH, for example, in Fig. 1 above signals that it is conceptually closely related to **incubator**.

We transform each entry into a hierarchy of a complex sort whose attributes may contain other relations, not just simple scalar values. An entry in the lexical database forms an iterative template as shown in Fig. 2.

title ... *n* (**a**) Titel *m (also Sport); (of chapter)* Uberschrift *f; (Film)* Untertitel *m; (form of address)* Anrede *f.* **what -- do you give a bishop?** ... ... (**b**) *(Jur) (right)* (Rechts)anspruch (*to* auf + *acc*), Titel *(spec) m;* ...

```
entry
 +-hdw: title
 +-superhom
  +-pronunc: ...
  +-hom
   +pos: n
   +sens
    +-sensnum: a
    +tran_group
     +-tran
      +-word: Titel
      +-gender: m
      +-domain: also Sport
      ....

    +-sens
     +-sensnum: b
     +-domain: Jur
     +-tran_group
      +-usage note: right
      +-tran
       +-word: Rechtsanspruch
       +-word: anspruch
       ....
```

Figure 2. Tree template for "title" in Collins English German Dictionary

In order to produce these iterative templates from the tape entries, the grammar formalism needs to be mildly context sensitive. It needs to handle fields that look similar but differ because of their local or global context. For example, **pos**, **domain**, **usage-note** and **style** in Fig. 2 all appear in italics but are differentiated by their relative position within the entry. The formalism should also be able to handle tree-transformations *during* parsing because of scoping phenomena of various kinds, such as in Fig. 2, where the gender marker (**m**) is common to both **Titel** and **anspruch**. A gender node should be duplicated for **anspruch**, but this is discovered and resolved only when the gender marker for **Titel** is reached. Finally the formalism has to be able to handle without failing information which it cannot yet parse. The information is gathered under a tree node without being further analyzed. This mechanism of "graceful failure" fits with our philosophy of incremental processing, allowing the grammars to grow in the future, until complete coverage is achieved.

The Parser

DEP takes as input entries from a type-setting tape, consults a grammar specific to the particular dictionary and produces explicit structural representations, which are either displayed on the screen or stored. The system includes a rule compiler, a parsing engine and a dictionary-entry template generator, all written in PROLOG. The compiler accepts 3 types of rules: tokenization, retokenization and parsing proper. Tokenization rules specify a one-to-one mapping from a character substring to some token, taking account of immediate context only. These rules transform the entry into a sequence of tokens and strings. The former serve as delimiters; the latter contain lexical information. Retokenization rules specify substitutions that are sensitive to local context. They remove superfluous tokens, form whole strings out of hyphen-delimited sequences of syllables, and handle various errors in the input format. The result of tokenization and retokenization is shown below. The tape stream corresponding to the **LDOCE** entry

F < autistic < F > au{*80}tis{*80}ticP < C:"tIstIK M < adj < S < 0000 < D < suffering
from {*CA}autism{*CB}R < 01 < R < -ally < R < > < adv < Wa4 <

is converted into the following token list:

```
head_marker                 . "autistic"
fld_sep . pf_marker         . "au-tis-tic"
pron_marker                 . "C:"tIstIK"
pos_marker                  . "adj"
...
```

Parsing rules make use of unification and backtracking to identify segments by context. Parsing rules operate top-down and depth-first. They remove the tokens introduced in the (re)tokenization stage(s) , assign labels to string segments and move them to their appropriate places in the entry-tree. The rules have the capability of arbitrary tree transformations (node raising, splitting and deletion) which is necessary in order to handle the complex and compact information encountered in dictionaries. Other important aspects of the parser are the optionality operator, the ability to constrain rule application by arbitrary tests, and the ability to escape to PROLOG when necessary.

With the **DEP** and respective grammars, we have parsed to date Collins German-English (98% of the entries), Collins English-German, English-French and French-English (95%), Webster's 7th (100%) and **LDOCE** (93%). We have parsed others in lesser detail. The remaining unparsed entries pose a variety of vexing problems:

inconsistencies in the type-setting tape; very long entries or ones with particularly complex structures. These require further enhancements to either the parser, the grammar or the present storage capacity. [32,33,34]

IV. SYNONYM DISAMBIGUATION

DEP transforms type-setting tapes into hierarchical structures that can be traversed and manipulated by our programs to yield semantic links and construct semantic networks. A main obstacle, though, to the formation of accurate networks is the polysemy of natural language. Most words linked in a dictionary or thesaurus are polysemous; that is, have more than one sense [7]. Typically, semantic links hold between word-senses and not between words. Thus, when word **A** is said in the thesaurus to be synonymous with word **B**, it is actually one sense of **A**, **Ai**, that is in this relation with one sense of **B**, **Bj**. Keeping record of which senses of a word are involved in a particular semantic relation is important in order to avoid forming spurious connections among unrelated senses of the same words.

Synonymy in CT

CT is an alphabetically arranged thesaurus with entries consisting of a headword, separated into different senses, and synonym lists for each of the senses. The links between synonyms in the thesaurus can be characterized according to their degree of symmetry and transitivity. We say that the link between **a** and **b** is **symmetric** if **a** points to **b** and **b** points to **a**; that is, if the headword **a** has **b** in its synonym list and the headword **b** has **a** in its list. We say that the link between **a** and **b** is **transitive** if for every word **c**, if **b** points to it then **a** points to it too; that is, if all the synonyms found in **a**'s synonym list are also found in **b**'s list (with the exception of **a** and **b** themselves, of course). Thus, if links were symmetric and transitive throughout the thesaurus, all words would partition into disjoint sets. Each member of the set would be a synonym of every other member. But there are only 27 sets of words in CT which exhibit completely symmetric and transitive links among their members. Most of the synonymy links in **CT** are different. 62% are asymmetric (e.g., **part** has **department** as a synonym, but **department** does not have **part**); and 65% are non-transitive (e.g., **part** has **piece** as a synonym; **piece** has **chunk** as a synonym; but **part** does not have **chunk** as a synonym).

Sense Disambiguation

Every entry in **CT** is broken into the different senses of its headword, as can be seen in the entry of **house**, given below, which contains 6 senses.

1. abode, building, domicile, dwelling, edifice, habitation, home, homestead, residence
2. family, household, ménage
3. ancestry, clan, dynasty, family tree, kindred, line, lineage, race, tribe

...

The synonyms listed for each sense, however, are not marked for their intended sense. Thus, it is not explicitly marked which sense of **abode**, for example, is linked to **house1**. We have tried two automatic methods of sense marking (i.e. sense disambiguation): disambiguation by symmetry and disambiguation by intersection.

In an alphabetically arranged thesaurus such as **CT**, an entry **a** may have word **b** listed as a synonym of its nth sense, and entry **b** may have word **a** listed as a synonym of its mth sense. We can mark **b** in entry **a** as the mth sense of **b**, and **a** in entry **b** as the nth sense of **a**. An example of this type of one-to-one mapping in **CT** is given below.

dense (adj) 1. ... condensed ... solid ....
2. ... dull ... stupid ...

dull (adj) 1. dense .... stupid ....
2. ... callous ... unsympathetic
.
.
.
7. drab ... muted ....

Here, sense 1 of **dull** is synonymous with sense 2 of **dense**. 37% of the 287,000 synonym tokens show this type of symmetry. Of course, there are also mappings of the one-to-many variety (for example, only the first sense of **feeble** has **faint** as its synonym, whereas both senses 1 and 2 of **faint** have **feeble**), but they account for only .5% of the tokens. By this method of disambiguation-by-symmetry, we could automatically mark the senses of all synonyms in one-to-one and one-to-many relations. The third type of mapping, many-to-many, accounts for just .5% of the total, but it poses a problem for the strategy outlined above. This can best be seen by considering an example. Senses 1 and 2 of **institution** list **establishment** as a synonym, and senses 1

and 2 of **establishment** list **institution**. Is sense 1 of **institution** synonymous with sense 1 of **establishment** or with sense 2? The distribution of the terms **institution** and **establishment** cannot answer the question.

The problem of many-to-many mappings and the large percentage of asymmetric **CT**-synonyms led us to another method. Consider again the case of **dense** and **dull**. Evidence for linking sense 2 of **dense** with sense 1 of **dull** comes from the symmetric distribution of the two words in the entries. There is however another piece of evidence for linking sense 2 of **dense** with sense 1 of **dull**, and that is the co-occurrence of the word **stupid** in their synonym lists. Thus, the intersections of synonym lists serve as the basis for an automatic disambiguation of the many-to-many mappings, and, for that matter, for the disambiguation of the whole **CT**. This is similar to Lesk's suggestion for disambiguating hypernyms [27]. The intersection method disambiguated more entries than the symmetry method, but it, too, left a certain percentage of ambiguous words. In some cases, the intersection of two words was null. In other cases, there was a tie. For example, **ripe2** has equal-size intersections with both **perfect1** and **perfect4**. No disambiguation resulted in either of these cases. The results obtained with each method are shown in the following table:

by symmetry:	
sense disambiguated:	103,648 (46.7%)
ties:	1,662 (0.7%)
remainder:	116,647 (52.5%)
Total number of synonyms available for processing:	221,957
by intersection:	
sense disambiguated:	179,126 (80.7%)
ties:	6,029 (2.7%)
remainder:	36,802 (16.6%)
Total number of synonyms available for processing:	221,957

Figure 3. Disambiguation Results

The quantitative advantage of the intersection method is evident. To determine the qualitative difference, we studied cases where the symmetry and the intersection methods conflicted. We compared fifty randomly selected entries. Of the approximately 900 synonyms listed in the entries, 337 were disambiguated by both methods. Of these, there were 33 pairs for which the two methods disagreed. 20 were symmetric

ties, disambiguated by the intersection method. 5 were intersection ties, disambiguated by the symmetry method. The remaining 8 were given to two human reviewers. In 3 out of the 8, the reviewers themselves could not determine which of the methods provided better disambiguation. To conclude, the best disambiguation algorithm would be a combination of the two methods. We are currently studying more cases where the methods disagree in order to determine how they should be combined[10,43].

V. DISAMBIGUATION OF DICTIONARY DEFINITIONS

Natural language ambiguity is not only due to the polysemy of words in isolation. It is also found in the way different words combine; for example, in the relationship between heads of phrases and their modifiers. We are interested in resolving this ambiguity as it exists in the text of dictionary definitions. Compared with free text, dictionary text is somewhat easier, since the style is fairly regular, but the vocabulary is vast enough to present a real challenge.

We have chosen to concentrate initially on definitions of the form "to VERB with NP" in **W7**[30]. Disambiguating these definitions consists of identifying the appropriate sense of "with" (that is, the type of semantic relation linking the VERB to the NP) and choosing, if possible, the appropriate senses of the VERB and the NP-head from among all their **W7** senses. For example, the disambiguation of the definition of **angle**(3,vi,1), "to fish with a hook", determines that the relation between **fish** and **hook** is use of instrument. It also determines that the intended sense of **fish** is (vi,1)-"to attempt to catch fish" and the intended sense of **hook** is (n,1)-"a curved or bent implement for catching, holding, or pulling". **W7** lists 4 senses for intransitive **fish** and 4 for the noun **hook**. Together with the five senses of **with** (described in the next section), these yield 80 possible sense combinations for "to fish with a hook".

To resolve the ambiguity, we follow the approach proposed by Jensen and Binot [20] and consult the dictionary definitions of the words involved. Our Disambiguation Module (henceforth **DM**) selects the most appropriate sense combination(s) in two parts: first, it tries to identify the semantic categories or types denoted by each sense of the VERB and the NP-head. It checks if the VERB denotes change, affliction, an act of covering, marking or providing. It tests whether the NP-head refers to an implement, a part of some other entity, a human being or group, an animal, a body part, a feeling, state, movement, sound, etc. We have defined 16 semantic categories for

nouns, so far. A most relevant question is how many such categories need to be stipulated. For the purpose of the work reported here, these 16 categories suffice. Others, however, will be needed for the disambiguation of other prepositions and other forms of ambiguity. Having tested for semantic categories of the NP-head, **DM** then tries to identify the semantic relation holding between the VERB and NP-head. In the constructions we are interested in, the semantic relation between the two terms depends not only on their semantic categories but also on the semantics of **with**, which we discuss in the following section.

The Meaning of WITH

Dictionaries and other lexicographical works typically explain the meaning of prepositions in a collection of senses, some involving semantic descriptions and others expressing usage comments. **W7** lists a total of 12 senses for **with** and various subsenses. **LDOCE** lists 20. Others attempt to group the variety of meanings under a few general categories [20]. After reviewing the different characterizations of the meanings of **with** against a small corpus of verb definitions containing **with**, we have arrived at a set of five senses for it, corresponding to five semantic relations that can hold between the VERB and the NP-head in "to VERB with NP". They are USE, MANNER, ALTERATION, CO-AGENCY/PARTICIPATION, and PROVISION, each including several smaller sub-classes. Each sense is characterized by a description of the states of affairs it refers to and by some criteria which test it. As can be expected, however, the criteria are not always conclusive. There exist both unclear and overlapping cases. The caracterization of USE and ALTERATION are provided below.

USE - examples are "to fish with a hook"; "to obscure with a cloud"; and "to surround with an army". **With** in this sense can usually be paraphrased as "by means of" or "using". The states of affairs in this category involve three participants: an agent (usually the missing subject of the definition), a patient (the missing object) and the thing used (the referent of "with NP"). The agent usually manipulates, controls or uses the NP-referent and the NP-referent remains distinct and apart from the patient at the end of the action. The sub-classes of USE are USE -OF-INSTRUMENT, -OF-SUBSTANCE, -OF-BODYPART, -OF-ANIMATE_BEING, -OF-OBJECT.

ALTERATION - examples are "to mark with bars"; "to impregnate with alcohol"; "to fill with air"; and "to strike with fear". In some cases, this sense can be paraphrased

with "make" and an adjective (e.g., "make full", "make afraid"); in others, with "put into/onto" (e.g., "put air into"; "put marks onto"). The states of affairs are ones in which change occurs in the patient and the NP-referent remains close to the patient or even becomes part of it. The sub-classes are ALTERATION -BY-MARKING, -BY-COVERING, -BY-AFFLICTION, and CAUSAL ALTERATION. Cases of overlap between ALTERATION and USE are abundant. "To spatter with some discoloring substance" is an example of creating a change in the patient while using a substance. The definition of **spatter** itself indicates this overlap: "to splash with or as if with a liquid; also to spoil in this way".

In addition to the five semantic meanings, there is also one purely syntactic function, PHRASAL, which **with** fulfills in verb-preposition combinations, such as "invest with authority".

The **DM** disambiguates a given string by classifying it as an instance of one of these six categories, and thus selecting the appropriate sense combination of the words in the string. The process of disambiguation is a function of interdependencies among the senses of the VERB, the NP-head and **with**, as we show in the next section.

The Disambiguation Process

The first step in the disambiguation process is parsing the ambiguous string (e.g., "to fish with a hook") by PEG, our syntactic parser, [19] and identifying the two relevant terms, the VERB and NP-head (**fish** and **hook**). Next, each of these terms is looked up in **W7**, its definitions are retrieved and also parsed by PEG. Heuristics then apply to the parsed definitions of the terms to determine their semantic categories. The heuristics contain a set of lexical and syntactic conditions to identify each semantic category. For example, the INSTRUMENT heuristic for nouns checks if the head of the parsed definition is "instrument", "implement", "device" ,"tool" or "weapon"; if the head is "part", post-modified by an **of**-pp, whose object is "instrument", "implement", etc.; if the head is post-modified by the participial "used as a weapon"; etc.. If any of these conditions apply, that sense of the noun is marked +INSTRUMENT. The heuristics apply to each definition in isolation, retrieving information that is static and unchanging. In the future, we intend to apply the heuristics to the whole dictionary and store the information in our lexical database.

Next, each of the possible **with**-relations is tried. Let us take USE as a first example. To determine whether a USE relation holds in a particular string, the **DM** considers the semantic category of the NP-head. The most typical case is when the NP-head is +INSTRUMENT, as in "to fish with a hook". In this case, the relationship of USE is further supported by a link established between the NP-head definition and the VERB definition through **catch**: a hook is an "... implement for **catching**, holding, or pulling" and to fish is "to attempt to **catch** fish". (See [20] for similar examples and discussion.) Such a link, however, is rarely found. In many other USE instances, it is the meaning of the NP-head alone that determines the relation. Thus, **DM** determines that USE applies to "to attack with bombs" based on **bomb**(n,1)-"an explosive **device** fused to **detonate** under specified conditions", although no link is established between **attack** and **detonate**.

USE is also applied regardless of the VERB when the NP-head is +BODYPART and certain syntactic conditions (a definite article or a 3rd-person possessive pronoun) hold of the string, as in "to strike or push with or as if with the head" and "to write with one's own hand". USE is similarly assigned if the NP-head is +SUBSTANCE: "to rub with oil or an oily substance" or "to kill especially with poison".

Since the heuristics for each semantic relation are independent of each other, conflicting interpretations may arise. There are cases of unresolved ambiguity, when different senses of one of the terms give rise to different interpretations. For example, "to write with one's own hand" receives a USE (-OF-BODYPART) interpretation but also a USE (-OF-ANIMATE_BEING), which is incorrect but due to several **W7** senses of **hand** which are marked +HUMAN ("one who performs or executes a particular work"; "one employed at manual labor or general tasks"; "worker, employee", etc.). A general heuristic can be added to prefer a +BODYPART interpretation over a +HUMAN one, since this ambiguity occurs with other body parts too. Other instances of ambiguity, however, are more idiosyncratic.

Results

We have developed our **DM** heuristics based on a training corpus of 170 strings - 148 transitive and 22 intransitive verb definitions extracted randomly from the letters **a** and **b** of **W7** using **QT** (see Section VI). The syntactic forms of the strings vary as can be seen from the following examples: "to suffer from or become affected with blight";

"to contend with full strength, vigor, craft, or resources"; "to prevent from interfering with each other (as by a baffle)". However, since we submit the strings to the PEG parser and retrieve the VERB and NP-head from the parsed structures, we are able to abstract over most of the variations.

The **DM** results can be summarized as follows: The correct semantic relation, based on the appropriate semantic category (of the NP-head or VERB), is assigned to 113 out of the 170 strings. Here are a few examples:

sever with an ax
  USE(-OF-INSTRUMENT)
wet with blood
  USE(-OF-SUBSTANCE)
inter with full ceremonies
  (ACTION-AS-) MANNER
dispute with zeal
  (ATTITUDE-AS-) MANNER
ornament with ribbon
  ALTERATION (BY-COVERING)
clothe with rich garments
  ALTERATION (BY-COVERING)
equip with weapons
  PROVISION

We consider these 113 results to be completely satisfactory.

In a second group of cases, the correct semantic relation, based on the appropriate semantic category, is one of 2 (and rarely of 3) semantic relations assigned to the string. There are 15 such cases. For example:

harass with dogs
  USE(-OF-ANIMATE_BEING) correct
  USE(-OF-INSTRUMENT) incorrect

The second interpretation is due to **dog**(n,3,a)-"any of various usually simple mechanical devices for holding, gripping, or fastening consisting of a spike, rod, or bar". We consider this second group of cases, which are assigned two interpretations, to be partial successes, since they represent an improvement over the initial number of possible sense combinations even if they do not fully disambiguated them.

In 37 cases, **DM** is unable to assign any interpretation. Failure to assign any interpretation does not, of course, count as success; but it does not produce much harm either. Far more dangerous than no assignment is the assignment of one incorrect interpretation, since incorrect interpretations cannot be differentiated from correct ones

in any general or automatic way. Out of the set of 170 strings, only 5 are assigned a single incorrect interpretation.

Since results obtained with the training corpus were promising, we ran **DM** on a testing corpus: 132 definitions of the form "to VERB with NP" not processed by the program before. The results obtained with the testing corpus are compared below with those of the training corpus. The first column lists the total number of strings; the second, the number of strings assigned a single, correct interpretation; the third, the number of strings assigned two interpretations, one of which is correct; the fourth column shows the number of strings for which no interpretation was found, and the last column lists the number of strings assigned one or more incorrect interpretations (but no correct ones).

	TOT	COR	1/2	0	INC
TRAINING	170	113	15	37	5
TESTING	132	75	13	22	22

To measure the reliability of **DM**, we calculate the ratio of correct interpretations to incorrect ones:

	COR-TO-INC RATIO
TRAINING	113/133 (or 85%)
TESTING	75/110 (or 68%)

If we include in the correct category those strings for which two interpretations were found (only one of which is correct), the reliability measure increases:

	COR + 1/2-TO-INC RATIO
TRAINING	128/133 (or 96.2%)
TESTING	88/110 (or 80%)

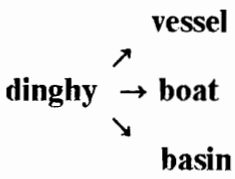
As expected, reliability for the testing material is lower than for the training set. This is due to the several iterations of fine-tuning to which the training corpus has been subjected. The examination of the testing results suggests some further fine-tuning, which is currently being implemented, and which will reduce the number of incorrect interpretations. [20,42]

VI. HYPERNYM EXTRACTION AND DISAMBIGUATION

Dictionary definitions form an implicit taxonomy of concepts. A headword **a** (hyponym) is typically defined as a **b** (hypernym), followed by some other modification (differentia). For example

boat n 1.1 a small **vessel** propelled by oars or paddles or by sail...

The relations between **a** and **b** is known as hypernymy, and chains of hypernyms define a taxonomy of concepts - from the most general to the most specific. Thus we obtain chains like



Semantic features are inherited from hypernyms to the hyponyms they define. By extracting all hypernym links in the dictionary, we create a taxonomy dictionary that reflects this important semantic information.

Identifying Hypernyms

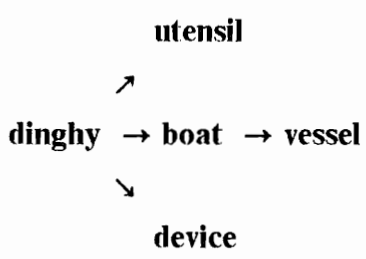
The first step is the identification of the hypernym(s) in the definition string. Definition strings are parsed with our broad-coverage syntactic parser, **PEG**, which analyzes definition strings as noun phrases or verb phrases and outputs their structure in the form of parse trees (in contrast with [8]). A program called **QT** (Querying Trees) is used to retrieve the heads of these phrases from among all the other nodes on the tree. **QT** finds target nodes in parser output without requiring the user to generate tree-walking programs. It is interactive and can be refined by the user as misses and false hits are encountered. **QT** is again a tool that can be easily modified to work on any tree-structure, thus appropriate for the retrieval of information from syntactic parses (as here) or from **DEP**-trees which display the structure of entire dictionary entries (as described in Section II).

The extraction of hypernyms was performed on noun and verb definitions of **W7** (and to a lesser extent on the definitions of **LDOCE**). From 68,000 noun definitions, 78,200 hypernym tokens were extracted (9300 hypernym types). From the 24,500 verb

definitions, 49,000 tokens were retrieved (7200 types). The success rate for nouns was 96.2%; for verbs - 98%. Failure was caused by the **DEP** parser or the **PEG** parser.

Hypernym Disambiguation

An attempt to construct hypernymy chains from the dictionary as it is may lead to the formation of spurious links, as from **dinghy** to **utensil** below, due to the ambiguity of **boat**:



In order to identify the intended sense of the ambiguous hypernyms, we use two techniques, similar to the ones described in [27] and in section III here. They both embody the principle that two senses are related in meaning if their definitions have a number of words in common. An exception list consists of function words and others that should not be counted for this purpose. We look for common words in the hyponym definition and in the definitions of each hypernym sense. The intended hypernym sense is the one (or several) sharing a number of common words with the definition of the hyponym. For example,

launder	1.0	a box **conduit** conveying middlings or tailings suspended in water in ore dressing
conduit	1.1	a natural or artificial channel through which water or other fluid is conveyed
	1.3	a pipe, tube, or tile for protecting electric wires or cables

The definition of **launder** shares two words in common with the definition of **conduit 1.1** ("convey" and "water"); therefore the latter is its disambiguated hypernym.

The results obtained for a sample of 86 hyponym-hypernym pairs, evaluated by hand, are as follows:

Number of pairs:	86
Number of mappings:	75
Number of successes:	7 (9%)
Number of partial successes:	24 (32%)
Number of omissions:	31 (41%)
Number of failures:	13 (18%)

Since there are several hypernym senses and since more than one can be intended as the appropriate sense in the hyponym definition, there is a large number of partial successes - that is, pairs where at least one correct hypernym sense was picked out. Sometimes other correct choices were missed; at others, incorrect choices were made in addition. In the majority of cases, no common words were found, and consequently, no hypernym sense was picked out at all.

The second technique that we used to pick out the intended hypernym sense involves common synonyms. We compare the synonym lists offered in **CT** for the hyponym with the lists offered for the hypernyms of each hypernym sense. This is better explained with an example:

law	1.1b3 the **agency** of or an agent of established law
agency	1.1 the **capacity**, **condition**, or **state** of acting or of exerting power: **operation**
	1.2 a **person** or **thing** through which power is exerted or an end is achieved: **instrumentality**
	1.3a the **office** or **function** of an agent
	...

To determine which sense of **agency** is intended in the definition of **law**, we compare the synonyms of **law** including **law** itself first with the synonyms of **capacity**, **condition**, **state** and **operation**; then with those of **person**, **thing** and **instrumentality**; etc. If any common synonyms are found, the relevant sense of **agency** is picked out.

Based on our small sample, common synonyms seem a more accurate test for disambiguation than common definition elements, but many words (especially concrete nouns) lack synonyms and are not found in the thesaurus at all. We found synonyms for only 31 pairs out of our sample of 86. We are now considering ways to refine these methods and improve our results. [9,24,25]

VII. APPLICATIONS

Although the work on lexical databases described above is still going on, some applications already benefit from our early results.[3]

Machine Translation

Experimental work is currently being done towards expanding our English-to-German machine translation system - **LMT** [29] - to have broad-coverage semantic capabilities. This work involves augmenting the information available in a small, hand-built lexicon with lexical information obtained from our various **DEP**-parsed sources by automatic querying. The main lexical problem in machine translation is that there usually are various translation candidates for each source word to be translated but only one term is semantically appropriate in the context. Our assumption is that some lexical information in the source context, such as the complement structure of the source word or the semantic category of its arguments, uniquely identifies the proper translation term. These identifiers are usually available in bilingual dictionary entries. Our task is two-fold: a) to extract this information from the dictionaries and manipulate it so that **LMT** can make use of it; b) to try to match the identifiers in the dictionary entry for a source-word with its actual context in the translated text.

The lexical access component of **LMT** extracts argument- and complement-structure information and constructs slot frames for the source word and for all of its possible target translations. After consulting two hand-coded lexicons, it turns to two machine-readable sources: several fields in **CEG** (the English-German dictionary) and features in **UDICT**, a large encoded lexicon containing syntactic features. **UDICT** was our first broad-coverage lexical database. It contains syntactic information only derived from **LDOCE** and augmented by analysis programs and hand-edited lists. [26] The lexical component of **LMT** currently retrieves information for verbs and nouns. For verbs, it gathers information from the feature field of **CEG** to determine transitivity; from the **comp** field, to obtain information about obligatory complements, such as direct objects, indirect objects and some prepositional-phrase complements; and from the **colloc** field, for further information about non-obligatory and more complex complements. **Colloc**

---

[3] For reasons of space limitations, we will not discuss the area of textual sense-disambiguation. See [3] and [1] for more information.

fields contain either common collocations or example sentences. Both need to be parsed by a simple grammar that recognizes prepositions and dictionary place-holders, such as **sb.** and **sth.**, for **somebody** and **something** . The information is extracted by means of queries, performed with the Lexical Query Language, an access method developed specifically for lexical databases [5,35].

The result of this lookup are source-translation pairs. During transfer, tests are performed on the textual context of the source-word in order to determine which sense is intended, and hence, which translation term is appropriate. Current experimental work consists of improving the selection process of the appropriate translation term by adding semantic information to the syntactic tests. For example, in order to choose between **fress** and **ess** in German, information is needed about the animacy of the subject of the source verb, **eat**, in context.

Lexicographical Tools

Converting the information found in published dictionaries and thesauri into a network and manipulating it on-line can be of significant use to the editors and compilers of these lexicographical resources in their efforts to revise and improve their content. We have looked at ways in which to assist the lexicographers of **CT** and found that asymmetric links in **CT** are good indication of problems. In particular, terminal nodes should be looked at. Terminal nodes are words that are offered as synonyms but do not occur as headwords themselves. They account for 36% of the total of asymmetric links in **CT**. Reviewing all terminal nodes by hand is extremely labor-intensive but a simple computer program generates a list of all terminal nodes in descending order of frequency, together with all the entries in which they occur. Lexicographers are thus able to choose how many and which of these terms they wish to analyze.

A small percent of the terminal nodes in CT is due to vocabulary inconsistencies. For example, **record** has **annals**, **archives** and **diary** as synonyms; whereas **annals** and **archives** have the plural **records**; and **diary** has the phrase **daily record**. This inconsistency results in both **records** and **daily record** becoming terminal nodes whereas, it would seem that they should not be, since they are equivalent to the main entry **record**. Identifying this category of terminals is particularly important because its correction involves changes in several entries.

Another indication of problems in the lexicographical data are **intransitive** links. We identify those by constructing synonym trees with a process called **SPROUTing** [8] Of particular importance to lexicographers are the nodes that point back to *different* senses of nodes already encountered. For example, the following branch of the **house1** tree points to a problem:

house1 → building1 → construction1 → building2

We have noticed that in most such loops, the problem lies in poor sense separation in the original **CT** entries. The sprouting mechanism may be useful when extensive changes are entertained for a family of word senses. [43]

VIII. CONCLUSION

The different projects described in this paper overlap and complement each other in interesting ways - for example, the disambiguation of hypernym senses in **W7** follows the method carried out for the disambiguation of synonym senses in **CT**; and the completion of hypernym disambiguation goes hand in hand with the disambiguation of modifiers (such as prepositional phrases) in definition texts. Implicit in this overlap and complementation loom most difficult problems, for which we do not yet have answers. One such problem is the **mapping** problem. As could be noticed, throughout the paper, we refer to several lexical databases, extracted from bilingual dictionaries, monolingual ones and thesauri. Our ultimate goal, however, is to create only one database that will encompass all the information presently found in multiple databases, by mapping information from one source onto that of the others. We have attempted to approach the mapping problem with the techniques described in this paper. For example, in order to map two monolingual dictionaries, such as **W7** and **LDOCE**, we looked at words shared in common. However, we have encountered problems in attempting the mapping of small samples. [23] This was due to the fact that the various sources differ in their sense distinctions. Thus, they vary in the number of senses they accord to the same words, and in the content of these senses. This is not surprising, given the open theoretical questions in semantics mentioned in Section I. Together with the mapping problem which remains open for our further study is a related problem of the structure of the comprehensive database we would like to have. Should word meaning continue to be represented in the form of hierarchical trees like

the ones generated by the **DEP** parser or is a tree structure too limited to express all lexical knowledge?

Another well-known problem in semantics is the adequacy of semantic features. So far, we have found them very useful for the disambiguation of terms in definition texts and for the selection of appropriate translation terms, but it is not clear how much information should be encoded in features. Finally, there remains the unclear separation between lexical information which should be encoded as static and relating to words in isolation and contextual information which should be inferred dynamically as words are encountered together in a text. Obviously, we do not have answers to these questions, but it is our hope that our developing work along the lines described in this paper will bring us closer to a satisfactory pragmatic position in relation to them.

REFERENCES

[1] J-L Binot, K. Jensen, "A Semantic Expert Using an Online Standard Dictionary", Proceedings of IJCAI-87, Milan, Italy, 1987.

[2] B. Boguraev, R. J. Byrd, J. L. Klavans, M. S. Neff, "From Structural Analysis of Lexical Resources to Semantics in a Lexical Knowledge Base", paper presented at IJCAI, Detroit, 1989. Also IBM Research Report 68655.

[3] L. Braden-Harder, W. Zadrozny, "Lexicons for Broad Coverage Semantics", in Lexical Acquisition: Using On-Line Resources to Build a Lexicon, U. Zernik, Ed., Lawrence Erlbaum, Hillsdale, New Jersey, to appear. Also IBM Research Report 15568, 1989.

[4] R. J. Byrd, "Discovering Relationships among Word Senses", Dictionaries in the Electronic Age: Proceedings of the Fifth Annual Conference of the University of Waterloo Centre for the New Oxford English Dictionary, Sept. 1989. Also IBM Research Report 14799.

[5] R. J. Byrd, "LQL User Notes, An Informal Guide to the Lexical Query Language", IBM Research Report 14853, 1989.

[6] R. J. Byrd, N. Calzolari, M. S. Chodorow, J. L. Klavans, M. S. Neff, O. A. Rizk, "Tools and Methods for Computational Lexicology", Computational Linguistics, vol. 13, no. 3-4, pp. 219-240, 1987. Also IBM Research Report 12642.

[7] M. S. Chodorow, "Making Sense of Word Senses: Detecting and Analyzing Systematic Polysemy in Noun Definitions", IBM Research Report, 1990.

[8] M. S. Chodorow, R. J. Byrd, G. E. Heidorn, "Extracting Semantic Hierarchies from a Large On-line Dictionary", Proceedings of the Association for Computational Linguistics, Chicago, Illinois, 1985, pp. 299-304.

[9] M. S. Chodorow, J.L. Klavans, "Locating Syntactic Patterns in Text Corpora", IBM Research Report, 1990.

[10] M. S. Chodorow, Y. Ravin, H. E. Sachar, "A Tool for Investigating the Synonomy Relation in a Sense Disambiguated Thesaurus", Proceedings of the 2nd ACL Conference on Applied NLP, Austin, Texas, 1988, pp. 144-151.

[11] N. Chomsky, Knowledge of Language: Its Nature, Origin and Use, Praeger, New York, 1986.

[12] Collins Spanish Dictionary: Spanish-English. English-Spanish, Collins Publishers, Glasgow, 1971.

[13] Collins Robert French Dictionary: French-English. English-French, Collins Publishers, Glasgow, 1978.

[14] Collins German Dictionary: German-English. English-German, Collins Publishers, Glasgow, 1980.

[15] Collins Sansoni Italian Dictionary: Italian-English. English-Italian, Collins Publishers, Glasgow, 1980.

[16] The New Collins Thesaurus, Collins Publishers, Glasgow, 1984.

[17] Cobuild, English Language Dictionary, Collins Publishers, London, 1987.

[18] G. E. Heidorn, K. Jensen, L. A. Miller, R. J. Byrd, M. S. Chodorow, "The EPISTLE Text-Critiquing System", IBM Systems Journal no. 21, pp. 305-326, 1982.

[19] K. Jensen, "PEG 1986: A Broad-coverage Computational Syntax of English", Unpublished paper, 1986.

[20] K. Jensen, J-L Binot, "Disambiguating Prepositional Phrase Attachments by Using On-Line Definitions", Computational Linguistics, vol. 13, no. 3-4, pp. 251-260, 1987.

[21] J. Justeson, S. Katz, "Antonymy, Co-occurrence, and Sense Disambiguation", 2nd I.T.L. on Natural Language Processing, IBM Scientific Center, Paris, 1990, pp. 363-74.

[22] J. Katz, Semantic Theory, Harper and Row, New York, 1972.

[23] J.L. Klavans, "Building a Computational Lexicon using Machine Readable Dictionaries", BUDALEX '88 Proceedings, Papers from the Euralex 3rd International Congress, Budapest, Hungary, 1988. Also IBM Research Report 14501.

[24] J. L. Klavans, R. Byrd, N. Wacholder, M. S. Chodorow, "Taxonomy and Polysemy", 2nd I.T.L. on Natural Language Processing, IBM Scientific Center, Paris, 1990, pp. 387-400.

[25] J.L. Klavans, M.S. Chodorow, N. Wacholder, "From Dictionary to Knowledge Base via Taxonomy", Proceedings of the Conference of the Centre for the New OED on Electronic Text Research, University of Waterloo, Canada, 1990.

[26] J.L. Klavans, N. Wacholder, "Documentation of Features and Attributes in UDICT", IBM Research Report 14251, 1989.

[27] M. Lesk, "Automatic sense disambiguation using machine-readable dictionaries: How to tell a pine cone from an ice cream cone", Proceedings of 1986 SIGDOC Conference, Canada, 1986.

[28] Longman Dictionary of Contemporary English, Longman Group, London, 1978.

[29] M. C. McCord, "Design of LMT: A Prolog-Based Machine Translation System", Computational Linguistics, vol. 15, pp.33-52, 1989.

[30] Webster's Seventh New Collegiate Dictionary G.&C. Merriam, Springfield, Massachusetts, 1963.

[31] R. Montague, Formal Philosophy: Selected Papers of Richard Montague, R.H. Thomason, Ed., Yale University Press, New Haven, Connecticut, 1974.

[32] M. S. Neff, B. K. Boguraev "Dictionaries, Dictionary Grammars and Dictionary Entry Parsing", Proceedings of the 27th Annual Meeting of the ACL, Vancouver, B.C., June, 1989, pp. 91-101.

[33] M. S. Neff, B. K. Boguraev, "Dictionary Entry Parser - A Reference Guide". IBM Research Report, 1989.

[34] M. S. Neff, B. K. Boguraev "From Machine-Readable Dictionaries to Lexical Data Bases", IBM Research Report 16080, 1991.

[35] M. S. Neff, R. J. Byrd, O. A. Rizk, "Creating and Querying Hierarchical Lexical Data Bases", Proceedings of the Second ACL Conference on Applied NLP, Austin, Texas, 1988, pp. 84-92.

[36] M. S. Neff, M. C. McCord, "Acquiring Lexical Data from Machine-Readable Dictionary Resources for Machine Translation", The Third International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Language, Linguistics Research Center, The University of Texas at Austin, June 1990. Also IBM Reasearch Report 15905.

[37] W. Quine, Word and Object, MIT Press, Cambridge, Massachusetts, 1960.

[38] Y. Ravin, "The Interaction Between the Lexicon and the Parser in the PLNLP System", Proceedings of the IBM ACIS University AEP Conference (Discipline Symposia), Boston, Massachusetts, 1987.

[39] Y. Ravin, "Grammar Errors and Style Weaknesses in a Text-Critiquing System", IEEE Transactions on Professional Communications, vol. 31, no. 3, pp. 108-115, 1988.

[40] Y. Ravin, Lexical Semantics without Thematic Roles, Oxford University Press, Oxford, England, 1990.

[41] Y. Ravin, "Synonymy from a Computational Point of View", in Frames, Fields and Contrasts: New Essays in Lexical and Semantic Organization, E. F. Kittay, A. Lehrer, Eds., Erlbaum, Hillside, New Jersey (in press). Also IBM Research Report 14962 (1989).

[42] Y. Ravin, "Disambiguating and Interpreting Verb Definitions", Proceedings of the 28th Meeting of the Association for Computational Linguistics, Pittsburgh, Pennsylvannia, 1990, pp. 260-267.

[43] Y. Ravin, M. S. Chodorow, H. E. Sachar, "Tools for Lexicographers Revising an Online Thesaurus", in BUDALEX '88 Proceedings, Papers from the Euralex 3rd International Congress. Budapest, Hungary, 1988.

[44] Roget's II: The New Thesaurus, Houghton Mifflin, Boston, Massachusetts, 1980.