INTELLIGENT DICTIONARY HELP SYSTEM

E. Agirre, X. Arregi, X. Artola, A. Díaz de Ilarraza, F. Evrard (*), K. Sarasola

Informatika Fakultatea, p.k. 649. 20080 DONOSTIA
(Basque Country - Spain)
e-mail: xabier@si.ehu.es Tel: 34 43 218000
(*) ENSEEIHT (Toulouse)

1. Introduction.

In this work dictionary definitions are considered as a special purpose language. The specific language used by the lexicographer to describe word senses has been analyzed and characterized. Different kinds of definitions have been represented by means of their corresponding relations in a knowledge base in order to provide several knowledge accessing capabilities. This knowledge base allows the deduction of the implicit knowledge conveyed by definitions.

IDHS (Intelligent Dictionary Help System) is conceived as a computer dictionary system for human use. IDHS supports reasoning mechanisms analogous to those used by humans when consulting a dictionary.

The starting point of IDHS is a Dictionary Database (DDB) built from an ordinary monolingual (explanatory) French dictionary. Meaning definitions have been analyzed using linguistic information from the DDB itself and interpreted in order to be structured as a Dictionary Knowledge Base (DKB). The intelligent exploitation of the dictionary is supported by the resulting DKB. The system has been implemented on a symbolic architecture computer using KEE knowledge engineering environment.

The DKB is based on the representation of the dictionary as a semantic network of frames, where each frame represents one concept. Frames are interrelated by attributes representing lexical-semantic relations such as taxonymy, synonymy, meronymy, and specific relations established by the lexicographic metalanguage used in definitions.

The system provides a set of several functions that have been inspired from the different reasoning processes a human user follows when consulting a dictionary, such as definition queries, search for alternative definitions, differences, relations and analogies between concepts, thesaurus-like word search, and so on.

Currently a multilingual environment is being designed on the basis of different dictionaries. MLDS (MultiLingual Dictionary System, an extension of IDHS) conceived as an intelligent help system for human translators, where two monolingual dictionaries (French and Basque) constitute the knowledge base along with a bilingual dictionary that establishes equivalence-links among concepts from the monolingual dictionaries.

Following is given a general motivation of the IDHS dictionary system. Section 3 presents the dictionary used as source in this project and gives a summary description of the construction of the Dictionary Knowledge Base (DKB). The knowledge representation model designed for the DKB is described in section 4. Finally the functionality of IDHS and MLDS is shown.

2. General motivation.

In this project, the dictionary is seen as a help system, as a vast reference handbook of the lexicon of a language. The user looks up words in order to know their meanings, find synonyms or similar words, confirm intuitions about different aspects, etc. The main objective of a dictionary is to help the user in language comprehension (reading), as well as in language production (writing) tasks. We are specially interested in the semantic aspect of the dictionary, that is, the definitions.

The importance of the lexicon in natural language processing is increasing**.** There is a need to make the process of construction of lexical components in NLP systems automatic, using for that actual dictionaries (Machine Readable Dictionaries, MRD).

IDHS is a dictionary help system for human users. The main objectives followed in its design and implementation are the following ones:

- To extract lexical-semantic knowledge from conventional dictionaries.
- To make a proposal for dictionary knowledge representation.
- To design the exploitation mechanisms needed to make explicit the knowledge implicit in dictionary structures.
- To specify a basic functionality set taking into account a wide variety of users.
- To integrate the system in a help context.

All the knowledge represented in IDHS system has been acquired from a conventional dictionary by means of parsing dictionary definitions using NLP techniques. Two different phases were distinguished to build the DKB. First the extraction of the information from the dictionary and its recording into a relational database: the Dictionary Database (DDB). This DDB was the starting point in order to create, in phase 2 (see figure 1), the object oriented Dictionary Knowledge Base, that is, in fact, the support of our deduction system.

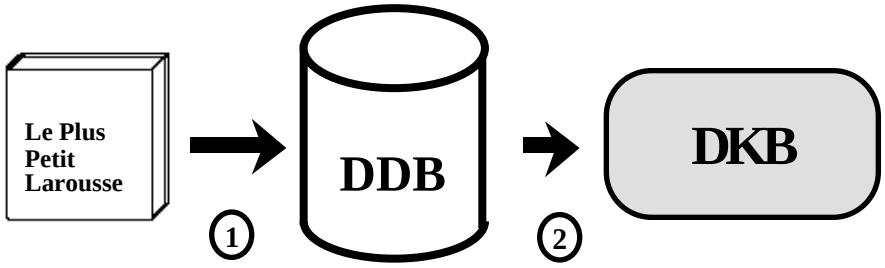

Fig. 1- From the MRD to the DKB

3. Building the Dictionary Knowledge Base.

The dictionary used as first source has been *Le Plus Petit Larousse* (Paris: Librairie Larousse, 1980), a French explanatory dictionary that contains the following fields for each entry: orthography, phonetics, part of speech, usage label, definition of the different senses, examples and others.

Definitions are quite short in this small dictionary. The average length of them is 3.27 words, 74.57% of definitions containing less than 5 words. There are 15953 entries, 70.09% of them with a unique sense and 21.29% with two senses, giving a total amount of 22899 senses. That is, the average number of senses per entry is 1.44. Besides, 1980 inflected forms are presented as entries. As there was no MRD version, the dictionary was recorded directly into a relational database: the DDB.

The method to parse dictionary definitions is based on pattern hierarchies as defined by (Alshawi, 89). The DDB itself has played the role of lexicon when parsing the definition sentences. Special attention has been paid to the method for building the patterns. As intuition may not be reliable enough, it has been done systematically. The main objective of this method is the semantic characterization of each different type of dictionary definition.

The method to characterize and parse dictionary definitions follows the next steps:

1) POS tagging and lexical disambiguation of words in definitions.
2) Statistical analysis of words in definitions.
3) Compilation of frequency lists of POS sequences in definition sentences.
4) Compilation of frequency lists of phrasal structure sequences in definition sentences.
5) Empirical research of stereotyped definition formulae. Finding specific relators such as "type of", "act of", or "kind of" (Vossen et al., 89).
6) Taken as basis the data obtained in steps 2 to 5, the hierarchy of patterns is built and the definitions parsed. The results of the parsing are added to the DDB.
7) After assigning to each pattern a semantic structure construction rule the DKB is generated automatically.

4. Structure of knowledge in IDHS.

The knowledge representation scheme chosen for the DKB of IDHS is composed of three elements (see figure 2), each of them structured as a distinct knowledge base:

- THESAURUS, a concept network where word senses are linked by means of lexical-semantic relationships.
- DICTIONARY allows access from dictionary word entries to their corresponding senses in THESAURUS.
- STRUCTURES contains meta-knowledge about concepts and relations in DICTIONARY and THESAURUS: all the different structures in the whole knowledge base are defined here hierarchically specifying the corresponding slots and describing them by means of facets that specify their value ranges, inheritance modes, etc.

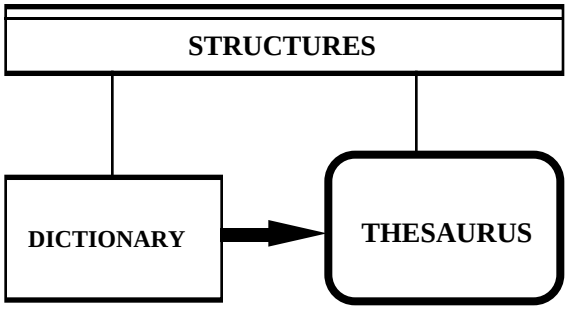

Fig. 2.- General schema of the DKB

4.1. THESAURUS knowledge base.

THESAURUS is the representation of the dictionary as a semantic network of frames, where each frame represents a one-word concept (word-sense) or a phrasal-concept (phrase structures associated to the occurrence of concepts in meaning definitions). Frames —or units— are interrelated by slots representing lexical-semantic relations such as synonymy, taxonomic relations (hypernymy, hyponymy, and taxonymy itself), meronymic relations (part-of, element-of, set-of, member-of), specific relations expressed by means of meta-linguistic relators, casuals, etc. Those relations have been implemented by means of reference attributes which point to concepts. Hypernymy and hyponymy have been made explicit (establishing a concept taxonomy) and implemented using the hierarchical relationship of the programming environment in order to get inheritance. Other slots contain phrasal, meta-linguistic, and general information.

4.2. DICTIONARY knowledge base.

This knowledge base is the link between each dictionary entry and its senses. The following example illustrates the link between the word *plante* and its corresponding senses.

|plante|
SENS: |plante I 1|, |plante I 2|

4.3. STRUCTURES knowledge base.

Four are the main object classes in the DKB: ATTRIBUTES, DEMONS, INFERENCE-RULES and DICTIONARY-STRUCTURES. The last one defines the data types as a taxonomy of units belonging to DICTIONARY and THESAURUS knowledge bases. The main dictionary data types are: ENTRIES (dictionary entries), DEFINITIONS (senses classified according to part of speech), REFERENCES (concepts created in THESAURUS due to their occurrence in definitions of other concepts), and CONCEPTS (dictionary senses and other conceptual units).

Three different classes of conceptual units are distinguished:

- TYPE-CONCEPTS. They are similar to Quillian´s "type nodes" (Quillian, 68). It is the superclass under which every concept of THESAURUS is placed. Those concepts are classified into: ENTITIES, ACTIONS/EVENTS, QUALITIES and STATES.
- PHRASAL-CONCEPTS. They correspond to Quillian´s "tokens", that is, occurrences of type concepts in definition sentences. They represent syntagmatic structures which are composed by several concepts with semantic content, e.g. |plante I 1#3| represents the noun phrase *une plante d'ornement*
- AMBIGUOUS-CONCEPTS. Which correspond to not completely disambiguated concepts.

There are two kinds of ATTRIBUTES:

- *Representational attributes* that reflect the surface (definitory) level representation of the definition of each sense (morphosyntax features like determination, verb mode, time, etc. are represented by means of facets).
- *Relational attributes* that are used to give the relational view of the lexicon. They support the deductive behaviour of the system.

4.4. Examples.

In order to represent the following definition:

géranium I 1: une plante d'ornement

two new conceptual units have to be created in the THESAURUS KB, the one which corresponds to the definiendum and the phrasal concept representing the noun phrase of the definition, as well as the units which represent *plante* and *ornement*, if they have not been previously created. Let us suppose that three new units are created: |*géranium I 1*|, |*plante I 1#3*| and |*ornement I 1*|. Their definitory level of representation is the following (slots are in capitals, facets or properties of slots are in smaller letters):

|géranium I 1|
MEMBER.OF: NOMS
GROUPE-CATEGORIEL: NOM
CLASSE-ATTRIBUT: INFO-GENERALE
TEXTE-DEFINITION: "une plante d'ornement"
CLASSE-ATTRIBUT: INFO-GENERALE
DEF-CLASSIQUE: |plante I 1#3|
CLASSE-ATTRIBUT: DEFINITOIRES
DETERMINATION: UN
GENRE: F
RELATIONNELS-CORRESPONDANTS: DEFINI-PAR

|plante I 1#3|
SUBCLASS.OF: |plante I 1|
MEMBER.OF: NOMINALES
TEXTE: "plante d'ornement"
CLASSE-ATTRIBUT: INFO-GENERALE
DE: |ornement I 1|
CLASSE-ATTRIBUT: SYNTAGMATIQUES
RELATIONNELS-CORRESPONDANTS: ORIGINE, POSSESSEUR, MATIERE, OBJECTIF
OBJECTIF: 0.9

|ornement I 1|
MEMBER.OF: REFERENCES

The knowledge structure resulting from the enrichment processes performed on the initial DKB by executing some deductive procedures (e.g. inverse relationships and taxonomy formation) is shown in figure 3. Note that, at this level, a OBJECTIF/OBJECTIF-INV relation has been deduced between |*géranium I 1*| and |*ornement I 1*| , on the basis (see above slot DE of unit |*plante I 1#3*|) that the preposition "de" was deemed to mean the relation "objectif" with certainty 0.9.

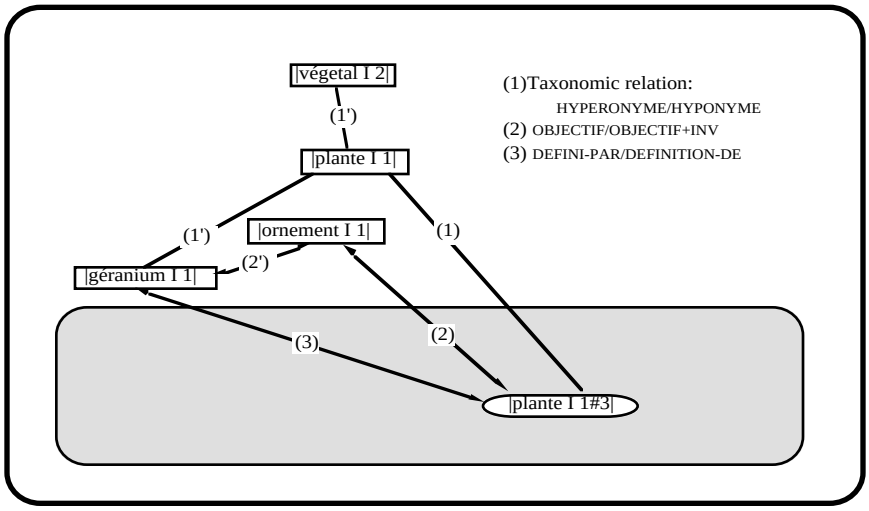

Fig.3.- Relational view of the concept |*géranium I 1*| (in the THESAURUS network). Phrasal concepts are inside the shaded box, type concepts outside.

Figure 4 shows the links among the three knowledge bases and the relations between the units created or referenced during the construction of the DKB corresponding to the following definition:

pansement I 1: *action de panser une plaie*

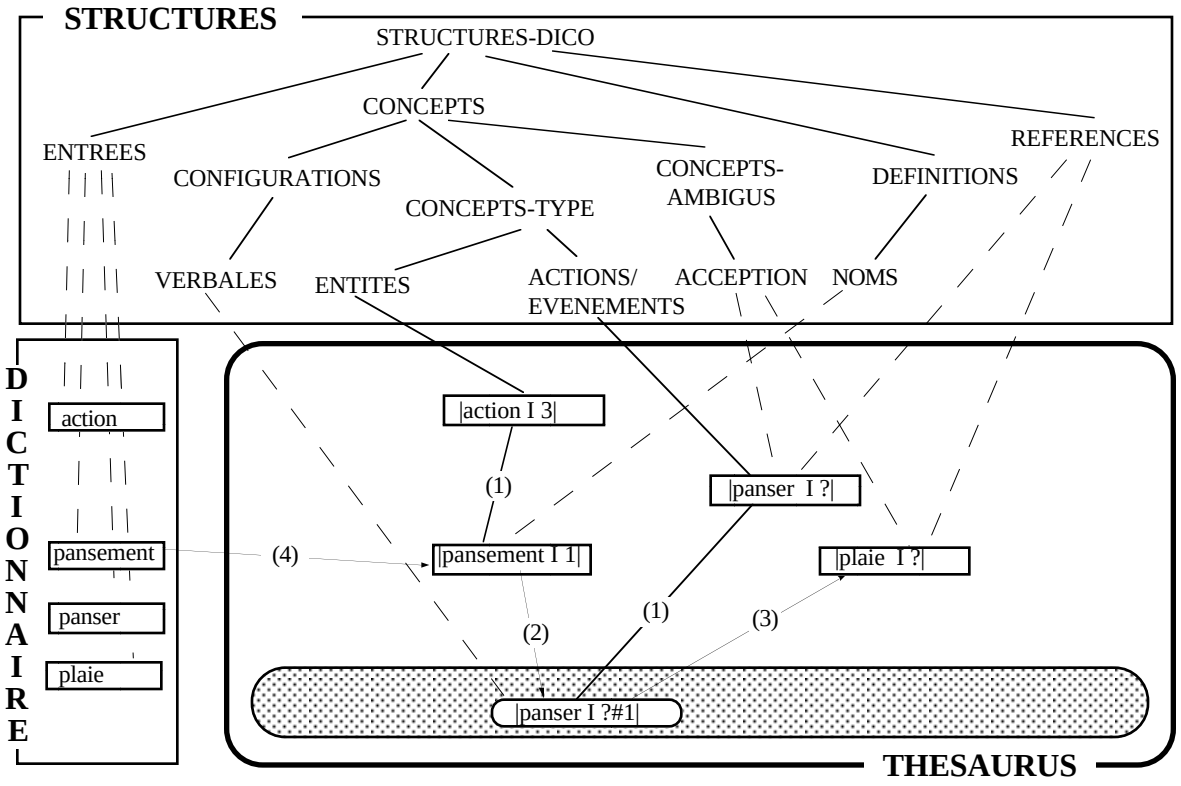

Fig. 4.- Links among the three knowledge bases

(1) Taxonomic relation: HYPERNYM/HYPONYM
(2) DEF-ACTIOND ("Act of" definition mode)
(3) OBJET (Object)
(4) SENS (Sense)
———— CLASS/SUBCLASS link
- - - - - - MEMBER.OF link

5. Functionality.

Before presenting the functionality of the system, MLDS will be introduced. MLDS is based on IDHS, incorporating more than one language and more than one dictionary per language with the purpose of helping human translators in the use of dictionaries.

Traditionally three different methods have been used in the analysis of dictionary use: a) free invention that relies only on intuition, b) questionnaires posed to human users –where it is difficult to distinguish between what the user answers and what he really does when using dictionaries–, and finally, c) direct observation, currently the most used method. Our method comprises direct observation along with protocols and personal interviews to professional translators.

The functions below result from our analysis of translators' needs. The monolingual version of some of these functions were already included in IDHS. They have been classified according to three main activities: source text understanding, object text generation, and search for translation equivalents.

5.1 Source text understanding.

There are three main functions in this activity: definition request (DDEF), reformulation of a definition (RDEF) and property-value request for a concept (DPRO).

For instance, DDEF takes as input a concept, an explanatory-level, a dictionary and a language, giving as output a definition. The following examples are definition queries for the meaning of *wasp* in the LPPL French dictionary, but the requested explanatory levels are different: *textual* in the first example, *local* in the second one (its result is the internal representation of the textual definition), and *inherited* in the third one (its result is the internal representation of the textual definition plus other relations deduced from the concept hierarchy).

Translator.-
DDEF (|guêpe I 1|, textual, LPPL, French, ?D)
*Definition of **wasp** in French with "textual" as explanatory-level*
System.- D= **'insecte hyménoptère à aiguillon'**

T.- DDEF (|guêpe I 1|, local, LPPL, French, ?D)
*Definition of **wasp** in French with "local" as explanatory-level*
S.- D= (and (|guêpe I 1| HYPERONYME |insecte I 1|)
(|guêpe I 1| CARACTERISTIQUE |hyménoptère I 1|)
(|guêpe I 1| POSSESSION |aiguillon I 1|))
Wasp is an **hymenopterous insect** with **sting**.

T.- DDEF (|guêpe I 1|, inherited, LPPL, French, ?D)
*Definition of **wasp** in French with "inherited" as explanatory-level.*
S.- D= (and (|guêpe I 1| HYPERONYME |insecte I 1|)
(|guêpe I 1| CARACTERISTIQUE |hyménoptère I 1|)
(|guêpe I 1| CARACTERISTIQUE |articuler I 1#m|)
(|guêpe I 1| POSSESSION |aiguillon I 1|)
(|guêpe I 1| POSSESSION |patte I 1#n|)
(|guêpe I 1| HYPONYME |frelon I 1|)
(|guêpe I 1| POSSESSEUR |guêpier I 1|))
Wasp is an **articulated hymenopterous insect** with **sting** and **legs**, a **bumblebee** is a wasp, and a **wasp's nest** has **wasps**.

5.2. Object text generation.

For this activity translators' needs are captured by the following functions: thesaurus-like search of concepts (RTHS), search of relations between two concepts (DRAP), request of differences between two concepts (DDIF), definition verification (VDEF), and property verification for a concept (VPRO).

For instance, *DRAP* gives the path relating two different concepts. RTHS takes as input a restriction-expression, a dictionary, and a language, and returns the list of concepts that meet the restrictions stated. Examples follow:

T.- RTHS((and (?X HYPERONYME |instrument I 1|)
(?X OBJECTIF |mesurer I 1|)),
LPPL, French, ?X, ?LC)
The user asks for nouns in French tools used for measurement
S.-
LC=(|baromètre I 1||dynamomètre I 1||télémètre I 1|...)

T.- RTHS((and (?X HYPERONYME |consumer I 1|)
(?X AGENT |feu I 1|)),
LPPL, Basque, ?X, ?LC)
The user asks for verbs in Basque for to consume *with agent* fire
S.- LC=(|izeki I 1|, |kiskali I 1|)
to burn, to blacken.

5.3. Search for translation equivalents.

There are some well known problems with lexical gaps when (a) there is no single word in the target language to express the source concept, which can be solved giving *phrasal concept equivalents*, and when (b) the source concept does not appear as an entry in the bilingual dictionaries; in this case, in order to express that the concept in the result is *more general* or *more specific* than the source concept, set operators as ? and • can be used.

In the first two examples below there is no problem when translating the concept |*accusatif I 1*| or |*coup_de_bec I 1*| from French into Basque. In the third and fourth examples |*pattar I 1*| and |*txakolin I 1*| are not in the bilingual dictionary, so the system gives the closest concept from the monolingual dictionary and indicates whether it is more or less specific. In the last example there is no single word to say *abere* (domestic animal) in French, therefore a phrasal concept is returned .

T.- EQUIV ((|accusatif I 1|, ,), Basque, gram, ?LP)
S.- LP = ((|akusatibo I 1|, ,))
T.- EQUIV ((|coup_de_bec I 1|, ,), Basque, common, ?LP)
S.- LP = ((|mokokada I 1|, ,))
T.- EQUIV ((|pattar I 1|, ,), French, common, ?LP)
S.- LP = ((•,|eau-de-vie I 1|, ,))
T.- EQUIV ((|txakolin I 1|, ,), French, common, ?LP)
S.- LP = ((?,|vin I 1|, ,))
T.- EQUIV (|(abere I 1|, ,), French, common, ?LP)
S.- LP = ((|animal I 1#n|, ,))
where |animal I 1#n| represents "domestic animal".

6. Conclusion.

The starting point of this project has been the semantic characterization of the different types of dictionary definitions that determine the sublanguage used in a conventional dictionary. This characterization lead to the automatic parsing of definitions, and to their representation in a knowledge base that provides several knowledge accessing capabilities.

IDHS and MLDS have been presented as two different systems in the context of intelligent dictionary help systems, IDHS in a monolingual environment and MLDS in a multilingual one with the following relevant aspects: extraction of knowledge from conventional dictionaries, a proposal for dictionary knowledge representation, deductive capabilities to make explicit the knowledge implicit in dictionary structures, and the specification of a basic functionality set.

A prototype of IDHS has been implemented on a Symbolics Lisp machine using KEE (Knowledge Engineering Environment).

References

Alshawi, H. "Analyzing dictionary definitions" in B. Boguraev, T. Briscoe eds., 153-169, *Computational Lexicography for Natural Language Processing.* New York: Longman, 1989.

Agirre, E., Arregi , X., Artola, X., Díaz de Ilarraza, A., Evrard, F., Sarasola, K. "Sistema diccionarial multilingüe: aproximación funcional". SEPLN congress, Santiago, Spain. 1993.

Artola, X. "HIZTSUA: Hiztegi-sistema urgazle adimendunaren sorkuntza eta eraikuntza / Conception d'un système intelligent d'aide dictionnariale (SIAD)" Ph.D. Thesis. UPV-EHU. 1993.

Artola, X., Evrard, F. "Dictionnaire intelligente d'aide à la compréhension". Actas IV Congreso International EURALEX'90 (Benalmádena), 45-57. Barcelona: Bibliograf., 1992

Boguraev, B., Briscoe, T. eds. *Computational Lexicography for Natural Language Processing.* New York:Longman**,** 1989.

Brustkern, J., Hess, K.D. "The BONNLEX lexicon system" in J. Goetschalckx, L. Rolling eds., 33-40, *Lexicography in the electronic age*. Luxembourg: North-Holland, 1982.

Byrd, R.J., Calzolari, N., Chodorow, M.S., Klavans, J.L., Neff, M.S., Rizk, O.A. "Tools and Methods for Computational Lexicography" *Computational Linguistics* 13, 3-4, 219-240. 1987.

Markowitz, J., Ahlswede, T., Evens, M. "Semantically significant patterns in dictionary definitions", *Proc. 24th Annual Meeting ACL* (New York), 112-119. 1986.

Quillian, M.R. "Semantic Memory" in M. Minsky ed., 227-270, *Semantic Information Processing*. Cambridge (Mass.): MIT Press, 1968.

Vossen, P., Meijs W., den Broeder, M. "Meaning and structure in dictionary definitions" in B. Boguraev, T. Briscoe eds., 171-192, *Computational Lexicography for Natural Language Processing*. New York:Longman, 1989.