A methodology for the extraction of semantic knowledge from dictionaries using phrasal patterns

Agirre E., Arregi X., Artola X., Díaz de Ilarraza A., Sarasola K.

Informatika Fakultatea, 649 p.k., 20080 DONOSTIA
Basque Country - Spain

Tel. 34 43 218000
Fax. 34 43 219306
e-mail: jiparzux@si.ehu.es

IDHS (Intelligent Dictionary Help System) is conceived as a computer dictionary system for human use. It 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 analysed using linguistic information from the DDB itself and interpreted in order to be structured as a Dictionary Knowledge Base (DKB); this DKB is the support of the deduction mechanisms. The analysis of the definitions has been done after some empirical studies on the data contained in the DDB. The analysis mechanism is mainly based on hierarchies of phrasal patterns [4] with some extensions. The system has been implemented using KEE knowledge engineering environment.

1 Introduction.

IDHS (Intelligent Dictionary Help System) is conceived as a monolingual (explicative) dictionary system for human use [7, 8]. The fact that it is intended for people instead of automatic processing distinguishes it from other systems dealing with semantic knowledge acquisition from conventional dictionaries. The system provides various access possibilities to the knowledge, allowing to deduce implicit knowledge from the explicit dictionary data. IDHS deals with reasoning mechanisms analogous to those used by humans when they consult a dictionary. User level functionality of the system has been defined and is currently being implemented using KEE knowledge engineering environment.

The starting point of IDHS is a Dictionary Database (DDB) built from an ordinary French dictionary. Meaning definitions have been analyzed using linguistic information from the DDB itself and interpreted to be structured as a Dictionary Knowledge Base (DKB). The intelligent exploitation of the dictionary is supported by the resulting DKB.

This paper describes the semantic knowledge acquisition process performed in order to build the DKB. The analysis of the definitions has been done after some empirical studies on the data contained in the DDB. The analysis mechanism is mainly based on hierarchies of phrasal patterns [4] with some extensions. The parser has been implemented, and integrated with the DDB so that the definitions are directly obtained from the DDB and the different parses result of the analysis are recorded in it. Obviously, the DDB itself has played the role of lexicon for the parser. The methodology used in the process of construction of the hierarchies is briefly explained.

As a result of the parsing different lexical-semantic relations between word senses are established by means of semantic rules (attached to the patterns); this rules are used for the initial construction of the DKB.

In the following section an overview of IDHS is given. Section 3 presents the small dictionary used as basis in this project and gives a summary description of the different procedures that have been done on it. The process of construction of the DKB from the analyzed DDB is described as well. The construction method of the hierarchies and the results obtained from the parsing are fully described in sections 4 and 5.

2 The IDHS dictionary system.

IDHS is a general purpose dictionary help system intended to assist a human user in language comprehension or production tasks. The system provides a set of several functions such as definition queries, search of alternative definitions, differences, relations and analogies between concepts, thesaurus-like word search, verification of concept properties and interconceptual relationships, etc. [2, 6].

The knowledge representation scheme chosen for the DKB of IDHS [3] is composed of three elements, each of them structured as a different knowledge base:

- KB-THESAURUS is the representation of the dictionary as a semantic network of frames, where each frame represents one concept (word-sense) or a phrasal concept (representation of 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 realized by means of meta-linguistic relators, casuals, etc. Other slots contain phrasal, meta-linguistic, and general information.
- KB-DICTIONARY allows access from the dictionary word level to the corresponding concept level in the DKB. Units in this knowledge base represent the entries (words) of the dictionary and are directly linked to their corresponding senses in KB-THESAURUS.
- KB-STRUCTURES contains meta-knowledge about concepts and relations in KB-DICTIONARY and KB-THESAURUS: all the different structures in the whole knowledge base are defined here specifying the corresponding slots and describing the slots by means of facets that specify their value ranges, inheritance modes, etc. Units in KB-THESAURUS and KB-DICTIONARY are subclasses or instances of classes defined in KB-STRUCTURES.

It is interesting to remark that the slots defined in KB-STRUCTURES have associated aspects such as the value class, the inheritance role determining how values in children's slots are calculated, and so on.

Two phases are distinguished in the construction of the DKB. Firstly, information contained in the DDB is used to produce an initial DKB. General information about the entries obtained from the DDB (POS, usage, examples, ...) is conventionally represented —attribute-value pairs in the frame structure— while the semantic component of the dictionary, i.e. the definition sentences, has been analyzed and represented as an interrelated set of concepts. In this stage the relations established between concepts could still be, in some cases, of lexical-syntactic nature. In a second phase, which will not considered in this paper, the semantic knowledge acquisition process is completed using for that the relations established in the initial DKB. The purpose of this phase is to perform lexical and syntactical disambiguation, showing that semantic knowledge about hierarchical relations between concepts can be determinant for this.

Figure 1 gives a partial view of the three knowledge bases with their correspondent units and their inter/intra relationships. The definition of ***spatule I 1:*** *sorte de cuiller plate* (a kind of flat spoon) is represented. This figure also shows the types of concepts used: *spatule I 1* and *cuiller I 1* are noun definitions and considered subclasses of ENTITIES while *plat I 1* (an adjective) is a subclass of QUALITIES. The phrasal concept unit representing the noun phrase *cuiller plate* is treated as a hyponym of its nuclear concept (*cuiller I 1*).

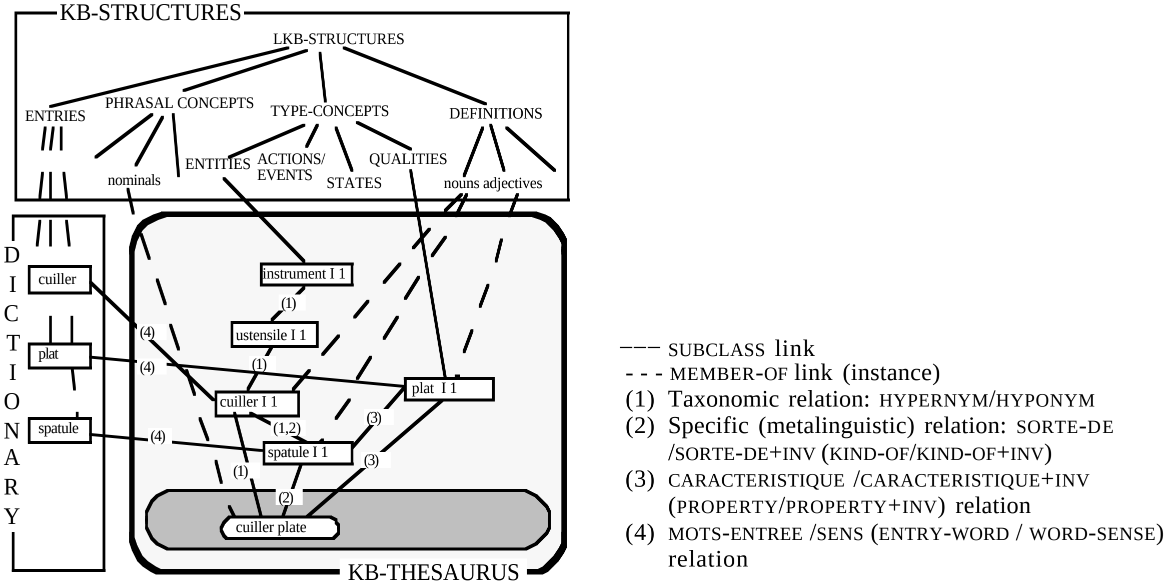
KB-STRUCTURES
LKB-STRUCTURES
ENTRIES
PHRASAL CONCEPTS
TYPE-CONCEPTS
DEFINITIONS
nominals
ENTITIES
ACTIONS/ EVENTS
STATES
QUALITIES
nouns adjectives
DICTIONARY
cuiller
plat
spatule
instrument I 1
ustensile I 1
cuiller I 1
spatule I 1
plat I 1
cuiller plate
(1)
(1,2)
(2)
(3)
(4)
KB-THESAURUS
——— SUBCLASS link
- - - MEMBER-OF link (instance)
(1) Taxonomic relation: HYPERNYM/HYPONYM
(2) Specific (metalinguistic) relation: SORTE-DE /SORTE-DE+INV (KIND-OF/KIND-OF+INV)
(3) CARACTERISTIQUE /CARACTERISTIQUE+INV (PROPERTY/PROPERTY+INV) relation
(4) MOTS-ENTREE /SENS (ENTRY-WORD / WORD-SENSE) relation

Fig. 1. The Dictionary Knowledge Base.

3 Building the Dictionary Knowledge Base.

The starting point of this system is a small monolingual French dictionary (*Le Plus Petit Larousse*, Paris: Librairie Larousse, 1980). This dictionary consists of nearly 23,000 senses related to almost 16,000 entries. Each entry contains the following components: part of speech (POS), meaning definition or cross-references to synonyms, marks of discourse domain usage, examples (14% of entries), etc. The average number of words per definition is 3.27.

The dictionary was recorded in a relational database: the Dictionary Database (DDB). This DDB is the basis of every empirical study that has been developed in order to design the final representation for the intelligent exploitation of the dictionary. The information attached in the DDB to each word occurrence in meaning descriptions was completed following a mainly automatic tagging process. As a result of this process more than 85% of word occurrences were tagged, and in 70% of definitions all words were tagged.

The definition sentences, that is the semantic component of the dictionary, have been analyzed in the process of transformation of the data contained in the DDB to produce the DKB. The analysis mechanism used is based on hierarchies of phrasal analysis patterns [4] including some modifications due mainly to its integration in the environment of the DDB, such as: (a) the definitions are directly obtained from the DDB; (b) the different parses, which are the result of the analysis, are recorded in it; and (c) the DDB itself has played the role of lexicon for the parser supplying the POS's corresponding to words from the DDB.

With regard to the construction of the semantic structure associated to each pattern, we distinguish three types of treatment:

a) Treatment associated to definitions which follow a classic schema. The links between the *definiendum* and the *genus* are of type *subclass* and properties described by the *differentia* are expressed by means of attributes.

b) Treatment associated to synonymic definitions. In this case, an attribute representing the synonymic relation is used.

c) Treatment associated to definitions with a specific formula (specific relators). Different kind of attributes are defined in order to represent the information conveyed by the formula.

Figure 2 gives an example of pattern rule (RN110532) defined in the hierarchy for analysis of noun entry definitions along with its semantic structure construction rule (SSCR). It covers noun entries defined by means of the words "sorte" "de", followed by a noun phrase (GN, *groupe nominal*) and an arbitrary sequence of words which match the wildcard M&&.

```
;;;         Pattern RN110532 (noun definitions)
;;;                   (N "sorte" "de" (GN 1) (M&&))
;;;         Corresponding SSCR:
            `(RN110532
;;;         Units to be created:
                      ((DEFINIENDUM)
                      (CONF1 CONF T (,(CAN '(GN 1 ((NOM 1))))) (NOMINALES))
                      (REF1 REF ,(CAN '(GN 1 ((ADJ 1)))) (QUALITES))
;;;         Value attachment:
                      ((,DEFINIENDUM DEF-SORTEDE ,CONF1)
                       (,CONF1 CARACTERISTIQUE ,REF1))
```

Fig. 2. Simplified example of pattern and semantic structure construction rule (SSCR).

The matching of the definition text of ***spatule I 1***, *sorte de cuiller plate* (see section 2) against the pattern RN110532 will give the following structure to be stored in the database:

```
(RN110532   ((GN 1  2  2    ((NOM 1  2  1)  (ADJ  1  3  1)))))
```

In this list structure the first element represents the identifier of the pattern matched by the definition. The second element is an expression of the bindings established between the match items in the pattern and the actual components of the definition text. In this case the only binding established is that corresponding to the (GN 1) item, matched with the sequence *cuiller plate*.

The semantic interpretation of each pattern is composed of two different parts: the specification of the representation units to be created, and the value attachment to the different attributes of those units. In the example, there are three new units to be created: DEFINIENDUM represents the concept defined, CONF1 stands for the phrasal concept unit representing the phrasal component matched with the (GN 1) item in the pattern, and REF1 -*plate*- represents the concept matched with the (GN 1 ((ADJ 1))) item. CONF1 is created as a noun phrasal concept unit depending hierarchically from its nuclear concept -*cuiller*- represented by (GN 1 ((NOM 1))) (whenever a phrasal concept is created these hierarchical links are automatically established); REF1 is created as a quality entity designed by its corresponding canonical form.

The value attachment part, among other things, establishes the specific relation DEF-SORTEDE between the definiendum and the unit CONF1, and the qualification relation (CARACTERISTIQUE) between CONF1 and REF1.

In the following the building process of the hierarchies for the parser is described. This process is of special relevance to obtain the best results from the analysis mechanism, since "intuition" about the structures to be found in definition sentences doesn't seem to be sufficient to build a well founded hierarchy of patterns.

4 Method for the construction of pattern hierarchies.

The method to construct the pattern hierarchies is based on semantic criteria: the objective is to find different types of syntactic structures used for the different meaning descriptions which are semantically significant. The main goal of the analysis is the characterization of the lexicographic language used in the definition texts in order to: a) specify SSCR's for the most frequent syntactic structures found in definitions, b) specify SSCR's for definitions containing elements able to work as specific relators between concepts [19], although their absolute frequency is not very high, and c) extract partial -but correct- information from those definitions not completely parsable.

Three kinds of statistical measures were calculated in order to check the semantic typology intuitively established, and search for more concrete words and phrase structures associated to each type of meaning description in the context of LPPL. They are the following ones:

- Frequency lists of POS sequences in definition sentences.

  Taking advantage of the DDB structure it was quite easy to obtain a frequency list of different sequences of POS's used in sense definitions.

- Frequency lists of sequences of phrasal structures in definition sentences.

  A specially designed and implemented bottom-up parser obtains a sequence of phrasal components for each definition.

  Frequency data on POS sequences and phrasal structures conveniently sorted were adequate starting points to define pattern rules for the most frequent syntactic schemes.

- Finding specific relators.

  Frequencies of word occurrences in definitions were calculated for noun, verb, adjective, and adverb entries. These frequency data and related work [11, 16, 19] have guided the search for specific meta-linguistic relators in this French dictionary. The specific relators correspond roughly to the notion of "shunters" and "linkers" in several known works [19].

  Particularized research on these elements showed the structure of their syntactic realizations in the definitions and led to the specification of the SSCR's for them. Unlike the case of classical definitions, the semantic interpretation for this kind of definitions is built in terms of these relators. For instance, *MEMBRE-DE* (member of) for noun entries, *FAIRE* (factitive) for verbs, or *QUI* (who) for adjectives are among these specific relators.

The lexical-semantic relations between different concepts extracted from the analysis of the source dictionary are grouped into two classes:

Paradigmatic relations: a) Synonymy and Antonymy; b) Taxonomic relations: Hypernymy / Hyponymy (obtained from definitions of type "genus et differentia"), and Taxonymy expressed by means of specific relators such as *SORTE-DE* or *ESPECE-DE*; c) Meronymy; and d) Others: Gradation (for adjectives and verbs), Equivalence (adjectives with past participle), Factitive and Reflexive (for verbs), Lack and Reference (to the previous sense).

Syntagmatic relations (those that relate concepts belonging to different POS's): a) Derivation; b) Relations between concepts without any morphological relation: case relation, and C) Others: Attributive (for verbs), Lack and Conformity.

5 Evaluation of the analysis: results and problems.

The hierarchies created have already been used for the analysis of all the noun, verb, and adjective definitions in the DDB. In this section a summary description of the created hierarchies is given along with some comments on the results of their application to the analysis of definitions.

The hierarchy devoted to analyze noun definitions is formed with 65 patterns, 49 different patterns have been defined to analyze verb definitions, and 45 for adjectives. Among these very general patterns can be found, i.e. noun phrase based patterns for noun definitions, verb phrase based ones for verbs and so on, along with very specific patterns derived from the empirical studies described in section 4.

To get an idea of the patterns created in the three hierarchies, it can be said that 37 patterns for nouns, 21 for verbs and 37 among the ones devoted to adjectives are specific patterns. Within these can be found patterns like ("partie" "de" (GN 1) (M&&)) for noun definitions, ("commencer" "à" (GV 1)) for verbs or ("sans" (GN 1)) for adjectives. Among these specific patterns also are considered those patterns defined in order to cope with synonymic definitions, which are of great importance in a small dictionary like the LPPL.

Together with pattern hierarchies 15 subsidiary patterns —mainly devoted to describe phrasal component structures— have been defined. These subsidiary patterns have also shown very useful when grouping under the same label different meta-linguistic structures destined to cope with specific relations stated in definitions.

Given that definitions in the LPPL are very short, the aim has been to get the most from the analysis process. Therefore, the wildcard of the pattern-matching (meta-character M&&) has been used in a very limited way placing it only in the last position of patterns. The correctness of partial analyses has not yet been evaluated exhaustively but it has been observed that the indiscriminate use of "everything matching items" in patterns, especially after facultative items, can lead to incorrect parses and hence, incorrect semantic structure assignments.

Although it is a partial parsing procedure, 57.76% of noun definitions, 79.8% of verbs and 69.04% of those corresponding to adjectives have been totally analyzed in this application (no use of M&&). The meta-character M&& matching zero or more words in a definition text has been used more often in patterns devoted to noun definitions.

	NOUN DEFS.		**VERB DEFS.**		**ADJ. DEFS.**	
		%		%		%
Failure	1086	7.9	105	2.00	527	16.35
Group 1	4657	33.88	2625	50.02	2370	73.51
Group 2	8001	58.22	2517	47.97	327	10.14
Totals	**13744**		**5247**		**3224**	

Fig. 3. Results of the analysis.

Number of patterns	**NOUN DEFS.**		**VERB DEFS.**		**ADJ. DEFS.**	
		%		%		%
1	8033	58.45	4322	82.37	2596	80.52
2	4183	88.88	810	97.81	579	98.48
3	1003	96.18	32	98.42	29	99.40
4	464	99.56	75	99.85	20	100.00

Fig. 4. Analysis ambiguity level.

In figure 3 the results of the analysis are presented, grouping the definitions according to the type of pattern they match.

With this technique of partial parsing, the parse is considered successful when an initial phrase structure is recognized, which in general contains the genus or superordinate of the defined sense. This is not so for the case of specific meta-language constructions, whose corresponding semantic structure is built in a specific way and which deserve specific patterns in the hierarchies.

As is seen, the failure rate is quite low, especially for verb definitions. This fact indicates, in the opinion of the authors, that the set of syntactic structures used by the lexicographer to write verb definitions is more reduced than those of other parts of speech definitions. The low failure rate also seems due to the fact that verb definition sentences are shorter in average than those of nouns.

Group 1 includes definition sentences that have been recognized by patterns based on certain meta-language level structures. Synonymic definitions are also considered to be in this level. In fact, these definitions are built using some meta-linguistic ways of expression, such as definitions containing a single word, or commas used as boundaries between words belonging to the same part of speech of the definiendum. This group is particularly relevant given that these meta-linguistic features offer a straightforward way to attach to them a suitable semantic interpretation expressed in terms of themselves. In this context, one should note the high percentage of adjective definitions in Group 1. This reveals that the use of synonymic definitions is very frequent in the case of adjectives (33.81%) and also that the utilization of certain fixed definition structures is higher here than for the other parts of speech: 39.7% of adjective definitions have been matched with only 27 patterns of this kind.

Group 2 concerns definitions expressed in the Aristotelian way of *genus et differentia specifica*. The possibility of partial parsing, where some differential aspects of the definition are eventually lost, is highly exploited in this case. The semantic interpretation given to all definitions included in this group takes the head of the first phrase structure (noun, verb or adjective phrase) as the *genus* of the defined word.

Figure 4 presents the level of ambiguity of the results obtained for the analysis of nouns, verbs and adjectives.

The figures indicate the number of definitions matched with, at most, four different patterns along with the accumulative percentage on the right column. This multiplicity of analyses is due either to the possibility that complex categories have different parses at the same pattern or to the fact that patterns placed on different branches of the hierarchy may be matched with the same definition. Obviously, there is no ambiguity with different patterns matched on the same branch. In this case, the parser chooses only the more specific one rejecting all the rest. Although all syntactically correct analyses are recorded in the database, to construct the DKB the first one is chosen as the right interpretation. This means that it is essential to arrange the different analyses so that the semantically correct one be on top. This can be done by arranging carefully the patterns on the same hierarchy level, given that results are ordered by the parser according to the hierarchy. However, results are revised manually before DKB construction, in order to solve remaining incorrect assignments of the semantic interpretation.

6 Conclusion.

A methodology for the extraction of semantic knowledge from a conventional dictionary has been described. This extraction has been founded on a systematic study of dictionary definitions. A parser based on phrasal pattern hierarchies has been implemented and used in that study.

The method followed in the construction of the hierarchies needed by the parser is based on an empirical study on the structure of definition sentences. The results of its application to a real dictionary has shown that the parsing method is particularly suited to the analysis of short definition sentences, as it was the case of the source dictionary.

As a result of this process, the characterization of the different lexical-semantic relations between senses —which is the basis for the proposed DKB representation scheme— has been established.

7 References.

[1] Ageno A., Castellón I., Martí M. A., Rigau G., Ribas F., Rodriguez H., Taulé M., Verdejo V.. SEISD: An environment for extraction of Semantic Information from on-line dictionaries, Proceedings 3rd Conference on Applied Natural Language Processing (Trento, Italia), 253-254. 1992.

[2] Agirre E., Arregi X., Artola X., Díaz de Ilarraza A., Evrard F., Sarasola K.. Intelligent Dictionary Help System. Proc. 9th Symposium on Languages for special Purposes. Bergen (Norway), 1993.

[3] Agirre E., Arregi X., Artola X., Díaz de Ilarraza A., Sarasola, K.. Lexical Knowledge Representation in an Intelligent Dictionary Help System. Proc. of COLING'94, 544-550. Kyoto (Japan), 1994.

[4] Alshawi, H.. Analysing dictionary definitions in B. Boguraev, T. Briscoe eds., 153-169, Computational Lexicography for Natural Language Processing. New York: Longman, 1989.

[5] Amsler, R.A.. A Taxonomy for English Nouns and Verbs, Proc. 19th Annual Meeting ACL, 133-138. 1981.

[6] Arregi X., Artola X., Díaz de Ilarraza A., Evrard F., Sarasola K.. Aproximación funcional a DIAC: diccionario inteligente de ayuda a la comprensión. Boletín SEPLN 11, 127-138. Barcelona, 1991.

[7] Artola X., Evrard F.. Dictionnaire intelligent d'aide à la compréhension, Actas IV Congreso Int. EURALEX´90 (Benalmádena), 45-57. Barcelona: Biblograph, 1992.

[8] Artola X.. HIZTSUA: Hiztegi-sistema urgazle adimendunaren sorkuntza eta eraikuntza / Conception et construction d´un système intelligent dáide dictionnariale (SIAD), Ph.D. thesis, University of the Basque Country, 1993.

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

[10] 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.

[11] Calzolari, N.. Detecting patterns in a lexical data base, Proc. COLING (Standford Univ.), 170-173. 1984.

[12] Calzolari N., Picchi E.. Acquisition of semantic information from an on-line dictionary, Proc. COLING (Budapest), 87-92. 1988.

[13] Chodorow M.S., Byrd R.J.. Extracting semantic hierarchies from a large on-line dictionary, Proc. ACL, 299-304. 1985.

[14] Copestake, A.. An approach to building the hierarchical element of a lexical knowledge base from a machine-readable dictionary, paper read at First Int. Workshop on Inheritance in NLP (Tilburg). 1990.

[15] van den Hurk I., Meijs W.. The dictionary as a corpus: analyzing LDOCE's definition-language, Corpus Linguistics II, 99-125.

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

[17] Pazienza M.T., Velardi P.. A structured representation of word-senses for semantic analysis, Proc. 3rd European Conference ACL (Copenhaguen), 249-257. 1987.

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

[19] 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.