COMPUTER AIDED INTERPRETATION OF LEXICAL COOCCURRENCES

Paola Velardi (*)
Maria Teresa Pazienza (**)

(*)*University of Ancona, Istituto di Informatica*, via Brecce Bianche, Ancona
(**)*University of Roma, Dip. di Informatica e Sistemistica*, via Buonarroti 12, Roma

ABSTRACT

This paper addresses the problem of developing a large semantic lexicon for natural language processing. The increasing availability of machine readable documents offers an opportunity to the field of lexical semantics, by providing experimental evidence of word uses (on-line texts) and word definitions (on-line dictionaries).

The system presented hereafter, PETRARCA, detects word cooccurrences from a large sample of press agency releases on finance and economics, and uses these associations to build a case-based semantic lexicon. Syntactically valid cooccurences including a new word W are detected by a high-coverage morphosyntactic analyzer. Syntactic relations are *interpreted* e.g. replaced by case relations, using a a catalogue of patterns/interpretation pairs, a concept type hierarchy, and a set of selectional restriction rules on semantic interpretation types.

Introduction

Semantic knowledge codification for language processing requires two important issues to be considered:

1. *Meaning representation.* Each word is a world: how can we conveniently circumscribe the semantic information associated to a lexical entry?
2. *Acquisition.* For a language processor · to implement a useful application, several thousands of terms must have an entry in the semantic lexicon: how do we cope with one such a prohibitive task?

The problem of meaning representation is one which preoccupied scientists of different disciplines since the early history of human culture. We will not attempt an overall survey of the field of semantics, that provided material for many fascinating books; rather, we will concentrate on the computer science perspective, i.e. how do we go about representing language expressions on a computer, in a way that can be useful for natural language processing applications, e.g. machine translation, information retrieval, user-friendly interfaces.

In the field of computational linguistics, several approaches were followed for representing semantic knowledge. We are not concerned here with semantic languages, which are relatively well developed; the diversity lies in the meaning representation principles. We will classify the methods of meaning representations in two categories: conceptual (or deep) and collocative (or surface). The terms "conceptual" and "collocative" have been introduced in [8]; we decided to adopt an existing terminology, even though our interpretation of the above two categories is broader than for their inventor.

1. *Conceptual Meaning* Conceptual meaning is the cognitive content of words; it can be expressed by *features* or by *primitives*. Conceptual meaning is "deep" in that it expresses phenomena that are deeply embedded in language.
2. *Collocative meaning.* What is communicated through associations between words or word classes. Collocative meaning is "superficial" in that does not seek for "the deep sense" of a word, but rather it "describes" its uses in everyday language, or in some sub-world

language (economy, computers, etc.). It provides more than a simple analysis of cooccurrences, because it attempts an explanation of word associations in terms of *conceptual relations* between a lexical item and other items or classes.

Both conceptual and collocative meaning representations are based on some subjective, human-produced set of primitives (features, conceptual dependencies, relations, type hierarchies etc.) on which there is no shared agreement at the current state of the art. As far as conceptual meaning is concerned, the quality and quantity of phenomena to be shown in a representation is subjective as well. On the contrary, surface meaning can rely on the solid evidence represented by word associations; the interpretation of an association is subjective, but valid associations are an observable, even though vast, phenomenon. To confirm this, one can notice that different implementations of lexicons based on surface meaning are surprisingly similar, whereas conceptual lexicons are very dishomogeneous.

In principle, the inferential power of collocative, or surface [18] meaning representation is lower than for conceptual meaning. In our previous work on semantic knowledge representation, however, [10] [18] [12] we showed that a semantic dictionary in the style of surface meaning is a useful basis for semantic interpretation.

The knowledge power provided by the semantic lexicon (limited to about 1000 manually entered definitions) was measured by the capability of the language processor DANTE [2] [18] [11] to answer a variety of questions concerning previously analyzed sentences (press agency releases on finance and economics). It was found that, even though the system was unable to perform complex inferences, it could successfully answer more than 90% of the questions [12][1]. In other terms, surface semantics seems to capture what, at first glance, a human reader understands of a piece of text.

In[26] , the usefulness of this meaning representation method is demonstrated for

TRANSALTOR, a system used for machine translation in the field of computers.

An important advantage of surface meaning is that makes it easier the acquisition of the semantic lexicon. This issue is examined in the next section.

Acquisition of Lexical Semantic Knowledge.

Acquiring semantic knowledge on a systematic basis is quite a complex task. One needs not to look at metaphors or idioms to find this; even the interpretation of apparently simple sentences is riddled with such difficulties that makes it hard even cutting out a piece of the problem. A manual codification of the lexicon is a prohibitive task, regardless of the framework adopted for semantic knowledge representation; even when a large team of *knowledge enters* is available, consistency and completeness are a major problem. We believe that automatic, or semi-automatic acquisition of the lexicon is a critical factor in determining how widespread the use of natural language processors will be in the next few years. '

Recently a few methods were presented for computer aided semantic knowledge acquisition. A widely used approach is accessing on-line dictionary definitions to solve ambiguity problems [3] or to derive type hierarchies and semantic features [24].

The information presented in a standard dictionary has in our view some intrinsic limitation:

- definitions are often *circular* e.g. the definition of a term A may refer to a term B that in turn points to A;
- definitions are *not homogeneous* as far as the quality and quantity of provided information: they can be very sketchy, or give detailed structural information, or list examples of use-types, or attempt some conceptual meaning definition;
- a dictionary is the result of a conceptualization effort performed by some human specialist(s); this effort may not be consistent with, or

```
ex1 (from [8]):
boy = + animate -adult + male

ex2 (from [25]):
help =
Y carrying out Z, X uses his resources W in order for W to help
Y to carry out Z; the use of resources by X and the carrying out of Z
by Y are simultaneous

ex2 (from [16]):
throw =
actor PROPELs and object from a source LOCation to a
destination LOCation
```

Figure 1. Examples of conceptual meaning representation in the literature

suitable for, the objectives of an application for which a language processor is built.

A second approach is using *corpora* rather than human-oriented dictionary entries. Corpora provide an experimental evidence of word uses, word associations, and language phenomena as metaphors, idioms, and metonymies.

The problem and at the same time the advantage of corpora is that they are raw texts whereas dictionary entries use some formal notation that facilitates the task of linguistic data processing.

No computer program may ever be able to derive formatted data from a completely unformatted source. Hence the ability of extracting lexical semantic information form a corpus depends upon a powerful set of *mapping rules* between phrasal patterns and human-produced semantic primitives and relations. We do not believe that a semantic representation framework is "good" if it mimics a human cognitive model; more realistically, we believe that a set of primitives, relations and mapping rules is "fair", when its coverage over a language subworld is suitable for the purpose of some useful language processing activity. Corpora represent an 'objective' description of that subworld, against which it is possible to evaluate the power of a representation scheme; and they are particularly suitable for the acquisition of a collocative meaning based semantic lexicon.

Besides our work [19], the only knowledge acquisition system based on corpora (as far as we know) is described in [7]. In this work, when an unknown word is encountered, the system uses pre-existing knowledge on the context in which the word occurred to derive its conceptual category.

The context is provided by on line texts in the economic domain. For example, the unknown word *merger* in "another merger offer" is categorized as *merger-transaction* using semantic knowledge on the word *offer* and on pre-analyzed sentences referring to a previous offer event, as suggested by the word *another*. This method is interesting but relies upon a pre-existing semantic lexicon and contextual knowledge; in our work, the only pre-existing knowledge is the set of conceptual relations and primitives.

PETRARCA: a method for the acquisition and interpretation of cooccurrences

PETRARCA detects cooccurrences using a powerful morphologic and syntactic analyzer [14] [1]; cooccurences are interpreted by a set of phrasal-patterns/ semantic-interpretation mapping rules. The semantic language is *Conceptual Graphs* [17]; the adopted type hierarchy and conceptual relations are described in [10]. The following is a summary description of the algorithm:

For any word W,

1. (A) Parse every sentence in the corpus that uses W.

    Ex: W = *AGREEMENT*

    "Yesterday an agreement was reached among the companies".

187

```
ex1 (from [18]):
agreement =
        is_a decision_act
        participant person, organization
        theme transaction
        cause communication_exchange
        manner interesting important effective ..

ex2 (from [26]):
person =
        isa creature
        agent_of take put find speech-action mental-action
        consist_of hand foot..
        source_of speech-action
        destination_of speech-action
        power human
        speed slow
        mass human
```
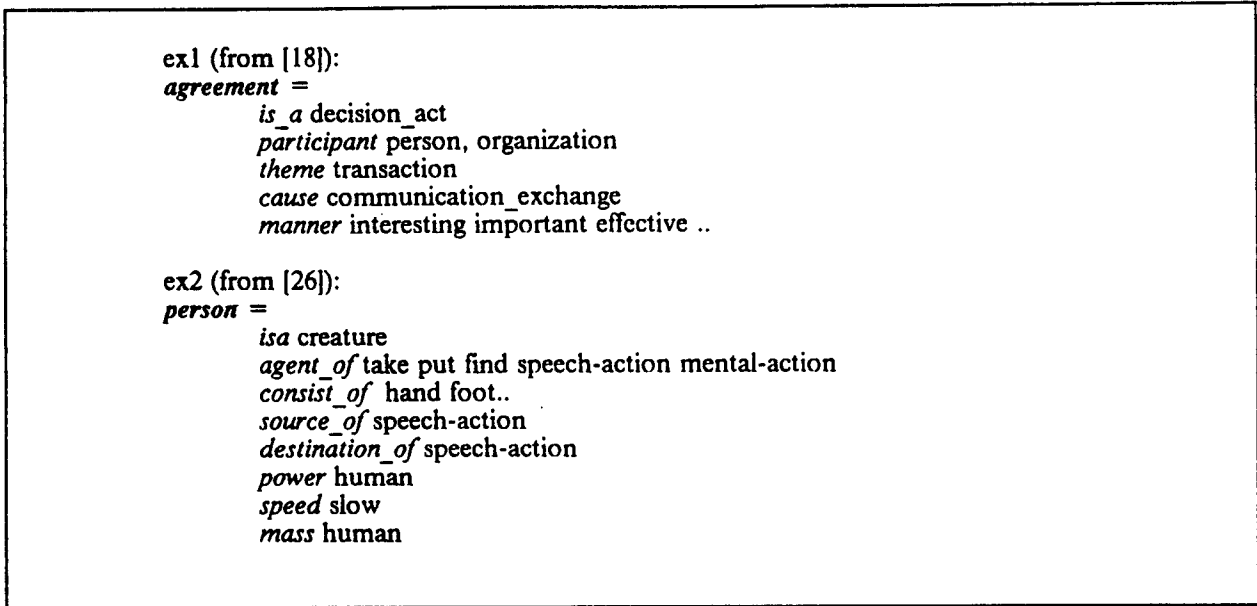
Figure 2. Examples of collocative meaning representation in the literature

2. (A) Determine all syntactic attachments of W (e.g. syntactically valid cooccurrences) Ex:

NP_PP(AGREEMENT,AMONG,COMPANY).
VP_OBJ(TO_REACH,AGREEMENT).

3. (A) Generate a semantic interpretation for each attachment :

Ex:

[AGREEMENT]->(PARTICIPANT)->[COMPANY].

4. (A) Generalize the interpretations.

Ex: Given the following examples:

[AGREEMENT]->(PARTICIPANT)->[COMPANY].
[AGREEMENT]->(PARTICIPANT)->[COUNTRY-ORGANIZATION].
[AGREEMENT]->(PARTICIPANT)->[PRESIDENT].

derive the most general constraint:

[AGREEMENT]->(PARTICIPANT)->[HUMAN_ENTITY]. The above is a new case description added to the definition of AGREEMENT

5. (M) Check the newly derived entry.

Steps marked (A) are automatic; steps marked (M) are manual. The only manual step is the last one: this step is however necessary because of the following:

- step 3 might produce more than one interpretation for a single word pattern, due to the low selectivity of some semantic rule.

- step 3 might fail to produce an interpretation for metonymies and idioms, which violate semantic constraints. Strong syntactic evidence (unambiguous syntactic rules) is used to "signal" the user this type of failure.

Knowledge sources used by PETRARCA

To perform its analysis, PETRARCA uses five knowledge sources:

1. an on line natural corpus (press agency releases) to select a variety of language expressions including a new word W;

2. a high coverage morphosyntactic analyzer, to derive phrasal patterns centered around W;

3. a catalogue of patterns/interpretation pairs, called Syntax-to-Semantic (SS rules);

4. a set of rules expressing selectional restriction on conceptual relation uses (CR rules);

5. a hierarchy of conceptual classes and a catalogue associating to words concept types.

The natural corpus and the parser are used in steps 1 and 2 of the above algorithm; SS rules, CR rules and the word/concept catalogue are used in step 3; the type hierarchy is used in steps 3 and 4

The parser used by PETRARCA is a high coverage morphosyntactic analyzer developed in the context of the DANTE system. The lexical parser is based on a Context Free grammar, the complete set of Italian prefixes and suffixes, and a lexicon of 7000 elementary lemmata (stems without affixes). At present, the morphologic component has an 100% coverage over the analyzed corpus (100,000 words) [14] [13].

The syntactic analysis determines syntactic attachment between words by verifying grammar rules and forms agreement; the system is based on an Attribute Grammar, augmented with lookahead sets [1]; the coverage is about 80%; when compiled, the parsing time is around 1-2 sec. of CPU time for a sentence with 3-4 prepositional phrases; the CPU is an IBM mainframe.

The syntactic relations detected by the parser are associated to possible semantic interpretations using SS rules. An excerpt of SS rules is given below for the phrasal pattern: *noun_phrase(NP) + prepositional_phrase(PP)* ($di = of$).

```
NP_PP('word1,di,'word2) < - rel(POSSESS,di,'word2,'word1).
/*cane di Pietro (the dog of Peter)*/
NP_PP('word1,di,'word2) < - rel(SOC_RELATION,di,'word2,'word1).
/*la madre di Piero (the mather of Peter)*/
NP_PP('word1,di,'word2) < - rel(PARTICIPANT,di,'word1,'word2).
/*riunione dei delegati (the meeting of the delegates)*/
NP_PP('word1,di,'word2) < - rel(SUBSET,di,'word2,'word1).
/*due di noi (two of us)*/
NP_PP('word1,di,'word2) < - rel(PART_OF,di,'word2,'word1).
/*pagine del libro (the pages of the book)*/
NP_PP('word1,di,'word2) < - rel(MATTER,di,'word1,'word2).
/*oggetto di legno (an object of wood)*/
NP_PP('word1,di,'word2) < - rel(PRODUCER,di,'word1,'word2).
/*ruggito dei leoni (the roar of the lions)*/
NP_PP('word1,di,'word2) < - rel(CHARACTERISTIC,di,'word2,'word1).
/*l'intelligenza dell'uomo (the intelligence of the man)*/
```

Overall, we adopted about 50 conceptual relations to describe the set of semantic relations commonly found in language; see [10] for a complete list. The catalogue of SS rules includes about 200 pairs. Given a phrasal pattern produced by the syntactic parser, SS rules select a first set of *conceptual relations* that are candidate interpretations for the pattern.

Selectional restriction rules on conceptual relations are used to select a unique interpretation, when possible. Writing CR rules was a very complex task, that required a process of progressive refinement based on the observation of the results. The following is an example of CR rule for the conceptual relation PARTICIPANT:

participant =

has_participant: meeting, agreement, fly, sail is_participant: human_entity

Examples of phrasal patterns interpreted by the *participant* relation are:
John flies (to New York); the meeting among parties; the march of the pacifists; a contract between Fiat and Alfa; the assembly of the administrators, etc.

An interesting result of the above algorithm is the following: in general, syntax will also accept semantically invalid cooccurrences. In addition, in step 3, ambiguous words can be replaced by the "wrong" concept names. Despite this, selectional restrictions are able to interpret only valid associations and reject the others. For example, consider the sentence: "The party decided a new strategy". The syntax detects the association *SUBJ(DECIDE,PARTY)*. Now, the word "party" has two concept names associated with it: POL_PARTY, and FEAST, hence in step 3 both interpretations are examined. However, no conceptual relation is found to interpret the pattern "FEAST DECIDE". This association is hence rejected.

Simalirily, in the sentence: "An agreement is reached among the companies, the syntactic analyzer will submit to the semantic interpreter two associations:
NP_PP(AGREEMENT,AMONG,COMPANY) and VP_PP(REACH,AMONG,COMPANY) . Now, the preposition *among* in the SS rules, points to such conceptual relations as PARTICIPANT, SUBSET (e.g. "two among all us"), and LOCATION (e.g. "a pine among the trees"), but none of the above relates a MOVE_ACT with a HUMAN_ORGANIZATION. The association is hence rejected.

Future experimentation issues

This section highlights the current limitations and experimentation issues with PETRARCA.

Definition of type hierarchies

PETRARCA gets as input not only the word W, but a list of concept labels CWi, corresponding to the possible senses of W. For each of these CWi, the supertype in the hierarchy must be provided. Notice however that the system knows nothing

189

about conceptual classes; the hierarchy is only an ordered set of labels.

In order to assign a supertype to a concept, three methods are currently being investigated. First, a program may "guide" the user towards the choice of the appropriate supertype, visiting top down the hierarchy. This approach is similar to the one described in [26].

Alternatively, the user may give a list of synonymous or near synonymous words. If one of these was already included in the hierarchy, the same supertype is proposed to the user.

A third method lets the system propose the supertype. The system assumes CW = W and proceeds through steps 1, 2 and 3 of the case descriptions derivation procedure. As the supertype of CW is unknown, CR rules are less effective at determining a unique interpretation of syntactic patterns. If in some of these patterns the partner word is already defined in the dictionary, its case descriptions can be used to restrict the analysis. For example, suppose that the word *president* is unknown in:

The *president* nominated etc.
Pertini was a good *president*'

the knowledge on possible AGENTs for *NOMINATE* let us infer PRESIDENT < HUMAN_ENTITY; from the second sentence, it is possible to further restrict to: PRESIDENT < HUMAN_ROLE. The third method is interesting because it is automatic, however it has some drawbacks. For example, it is slow as compared to methods 1 and 2; a trained user would rather use his experience to decide a supertype. Secondly, if the word is found with different meanings in the sample sentences, the system might never get to a consistent solution. Finally, if the database includes very few or vague examples, the answer may be useless (e.g. ACT, or TOP). It should also be considered that the effort required to assign a supertype to, say, 10.000 words is comparable with the encoding of the morphologic lexicon. This latter required about one month of data entry by 5-6 part-time researchers, plus about 2-3 months for an extensive testing.

The complexity of hierarchically organizing concepts however, is not circumscribed to the time consumed in associating a type label to some thousand words. All NLP researchers experimented the difficulty of associating concept

types to words in a consistent way. Despite the efforts, no commonly accepted hierarchies have been proposed so far. In our view, there is no evidence in humans of primitive conceptual categories, except for a few categories as animacy, time, etc. We should perhaps accept the very fact that type hierarchies are a computer method to be used in NLP systems for representing semantic knowledge in a more compact form. Accordingly, we are starting a research on semi-automatic word clustering (in some given language subworld described by a natural corpus), based on fuzzy set and conceptual clustering theories.

Interpretation of idiomatic expressions

In the current version of PETRARCA, in case of idiomatic expressions the user must provide the correct interpretation. In case of metaphors, syntactic evidence is used to detect a metaphor, under the hypothesis that input sentences to the system are syntactically and semantically correct. At the current state of implementation, the system does not provide automatic interpretation of metaphors. However, an interesting method was proposed in [20]. According to this method, when for example a pattern such as "car drinks" is detected, the system uses knowledge of canonical definitions of the concepts "DRINK" and "CAR" to establish whether "CAR" is used metaphorically as a HUMAN_ENTITY, or "DRINK" is used metaphorically as "TO_BE_FED_BY". An interesting user aided computer program for idiomatic expressions analysis is also described in [23].

Generalization of case descriptions

In PERTRARCA, phrasal patterns are first mapped into "low level" case description; in step 4, "similar" patterns are merged into "high level' case descriptions. In a first implementation, two or three low level case descriptions had to be derived before creating a more general semantic rule. This approach is biased by the availability of example sentences. A word often occurs in dozens of different contexts, and only occasionally two phrasal patterns reflect the same semantic relation. For example, consider the sentences:

The company signs a *contract for new funding*
The ACE stipulates a *contract to increase* its influence

Restricting ourselves to the word "contract", we get the following semantic interpretations of syntactic patterns:

```
1.[SIGN]->(THEME)->[CONTRACT]
2.[CONTRACT]->(PURPOSE)->[FUNDING]
3.[STIPULATE]->(THEME)->[CONTRACT]
4.[CONTRACT]->(PURPOSE)->[INCREASE]
```

In patterns 1 and 3 "sign" and "stipulate" belong to the same supertype, i.e. INFORMATION_EXCHANGE; hence a new case description can be tentatively created for CONTRACT:

```
[CONTRACR]->(THEME)->[INFORMATION_EXCHANGE]
```

Indeed, one can tell, talk about, describe etc. a contract.

Conversely, patterns 3 and 4 have no common supertype; hence two "low level" case descriptions are added to the definition of CONTRACT.

```
[CONTRACT]->(PURPOSE)->[FUNDING]
[CONTRACT]->(PURPOSE)->[INCREASE]
```

Even with a large number of input sentences, the system creates many of these specific patterns; a human user must review the results and provide for case descriptions generalization when he/she feels this being reasonable.

A second approach is to generalize on the basis of a single example, and then retract (split) the rule if a counterexample is found. Currently, we are studying different policies and comparing the results; one interesting issue is the exploitation of counterexamples.

Concluding remarks

Even though PETRARCA is still an experiment and has many unsolved issues, it is, to our knowledge, the first reported system for extensive semantic knowledge acquisition. There is room for many improvements; for example, PETRARCA only detects, but does not interpret idioms; neither it knows what to do with errors; if a wrong interpretation of a phrasal pattern is derived, error correction and refinement of the knowledge base is performed by the programmer. However PETRARCA is able to process automatically raw language expressions and to perform a first

classification and encoding of these data. The rich linguistic material produced by PETRARCA provides a basis for future analysis and refinements. Despite its limitations, we believe this method being a first, useful step towards a more complete system of language learning.

References

[1] F. Antonacci, P. Velardi, M.T. Pazienza, A High Coverage Grammar for the Italian Language, *Journal of the Assoc. for Literary and Linguistic Computing*, in print 1988.

[2] F. Antonacci, M.T. Pazienza, M. Russo, P.Velardi, Representation and Control Strategies for large Knowledge Domains : an Application to NLP, *Journal of Applied Artificial Intelligence*, in print 1988.

[3] JL. Binot and K. Jensen A Semantic Expert Using an On-line Standard Dictionary *Proceedings of the IJCAI* Milano, 1987

[4] K. Dahlgren and J. McDowell Kind Types in Knowledge Representation *Proceedings of the Coling-86* 1986

[5] Heidorn G.E. "Augmented Phrase Structure Grammar" in "Theoretical Issues in Natural Language Processing" Nash-Webber and Schank ,eds,ACL 1975

[6] J. Katz, P. Postal An Integrated Theory of Linguistic Descriptions *Cambridge, M.I.T. Press*, 1964.

[7] P. Jacobs, U. Zernik Acquiring Lexical Knowledge from Text: a Case Study, *Proceedings of the AAAI88*, St. Paul, August 1988

[8] Geoffrey Leech Semantics: The Study of Meaning *second edition, Penguin Books* 1981.

[9] Michalsky R.S., Carbonell J.C., Mitchell T.M. Machine Learning vol I *Tioga Publishing Company* Palo Alto, 1983

[10] M.T. Pazienza and P. Velardi, A Structured Representation of Word Senses for Semantic Analysis *Third Conference of the European*

Chapter of the ACL, Copenhagen, April 1-3 1987.

[11] M.T. Pazienza and P. Velardi, **Integrating Conceptual Graphs and Logic in a Natural Language Understanding System,** *in "Natural Language Understanding and Logic Programming II",* V. Dahl and P. Saint-Dizier editors, North-Holland, 1988.

[12] M.T. Pazienza, P. Velardi, **Using a Semantic Knowledge Base to Support A Natural Language Interface to a Text Database,** *7th International Conference on Entity-Relationship Approach,* Rome, November 16-18 1988.

[13] M. Russo, **A Rule Based System for the Morphologic and Morphosyntactic Analysis of the Italian Language,** *in "Natural Language Understanding and Logic Programming II",* V. Dahl and P. Saint-Dizier editors, North-Holland, 1988.

[14] M. Russo, **A Generative Grammar Approach for the Morphologic and Morphosyntactic Analysis of Italian,** *Third Conference of the European Chapter of the ACL,* Copenhagen, April 1-3 1987.

[15] Shank R.C. **Conceptual Dependency: a theory of natural language understanding.** *Cognitive Psicology, vol 3* 1972

[16] Shank R.C, Goldman, Rieger, Riesbeck **Conceptual Information Processing** *Noth-Holland/american Elsevier* 1975

[17] J.F. Sowa, **Conceptual Structures: Information Processing in Mind and Machine,** *Addison-Wesley,* Reading, 1984.

[18] P. Velardi, M.T. Pazienza and M. DeGiovanetti, **Conceptual Graphs for the Analysis and generation of sentences,** *IBM Journal of Research and Development,* March 1988.

[19] P. Velardi, M.T. Pazienza, S. Magrini **Acquisition of Semantic Patterns from a natural corpus of texts** *ACM-SIGART special issue on knowledge acquisition* in print

[20] E. Way **Dinamic Type Hierachies: An Approach to Knowledge Representation through Metaphor** *PhD dissertation, Dept. of System Science, State Univ. of NY at Binghamton* 1987.

[21] Y. Wilks, **Preference Semantics** *Memoranda from the Artificial Intelligence Laboratory, Stanford University* Stanford, 1973

[22] Y. Wilks, **Deep and Superficial Parsing,** *in "Parsing natural Language"* M. King editor, Academic Press, 1983.

[23] U. Zernik **Strategies in Language Acquisition: Learning Phrases from Examples in Contexts.** *Phd dissertation, Tech. Rept. UCLA-AI-87-1, University of California, Los Angeles* 1987

[24] R. Byrd, N. Calzolari, M. Chodorow, J. Klavans, M. Neff, O. Rizk **Large lexicons for Natural Language Processing: Utilizing the grammar Coding System of LDOCE** *Computational Linguistics, special issue of the Lexicon D. Walker, A. Zampolli, N. Calzolari editors July-December 1987* 1987.

[25] I. Mel'cuk, A. Polguere **A Formal Lexicon in Meaning-Text Theory (or How To Do Lexica with Words)** *Computational Linguistics, special issue of the Lexicon D. Walker, A. Zampolli, N. Calzolari editors July-December 1987* 1987.

[26] S. Nirenburg, V. Raskin **The Subworld Concept Lexicon and the Lexicon Management System** *Computational Linguistics, special issue of the Lexicon D. Walker, A. Zampolli, N. Calzolari editors July-December 1987* 1987.

[27] J. Pustejovsky **Constraints on the Acquisition of Semantic Knowledge,** *Journal of Intelligent Information Systems,* vol 3, n. 3, fall 1988