DETECTING PATTERNS IN A LEXICAL DATA BASE

Nicoletta Calzolari

Dipartimento di Linguistica - Universita' di Pisa
Istituto di Linguistica Computazionale del CNR
Via della Faggiola 32
56100 Pisa - Italy

ABSTRACT

In a well-structured Lexical Data Base, a number of relations among lexical entries can be interactively evidenced. The present article examines hyponymy, as an example of paradigmatic relation, and "restriction" relation, as a syntagmatic relation. The theoretical results of their implementation are illustrated.

I INTRODUCTION

In previous papers it has been pointed out that in a well-structured Lexical Data Base it becomes possible to detect automatically, and to evidence through interactive queries a number of morphological, syntactic, or semantic relationships between lexical entries, such as synonymy, hyponymy, hyperonymy, derivation, case-argument, lexical field, etc.

The present article examines hyponymy, as an example of paradigmatic relation, and what can be called "restriction or modification" relation, as a syntagmatic relation. By restriction or modification relation, I mean that part of a so-called "aristotelian" definition which has the function of linking the "genus" and the "differentia specifica".

When evidenced in a lexicon, the hyponymy relation produces hierarchical trees partitioning the lexicon in many semantically coherent subsets. These trees are not created once and for all, but it is important that they are procedurally activated at the query moment.

While evidencing the second relation considered, one can investigate as to whether it is possible to discover any correlation between lexical or grammatical features in definitions and particular kinds of "definienda", and thus try to answer questions such as the following: "Are there any connections between these restriction relations and the fundamental ways of definition, i.e. the criterial parameters by which people defines things?"

For both relations, the paper presents the different procedures by which they are automatically recognized and extracted from the natural language definitions, the degree of reliability of their automatic labeling, the use of these labels in interactive queries on the lexical data base, and finally the theoretical results of their implementation in a Machine-Dictionary.

II THE LANGUAGE OF DEFINITIONS AS A SUBLANGUAGE

I am trying to develop and exploit the idea of considering the language of dictionary definitions as a particular sublanguage within natural language. This perspective cannot obviously be adopted for subject matter restrictions in definitions, but only for the purpose of the text, i.e. the specific communicative goal. From this restriction on the purpose of the text, certain lexico-grammatical restrictions do result, which prove to be very useful.

As to the restrictions on the lexical richness of definitions, these are not due to the fact that they relate to a specific domain of discourse, but only to the property of closure (although not satisfied at 100%) that the defining vocabulary should in principle be simpler and more restricted than the defined set of lemmas, i.e. the former should be a proper subset of the latter.

This kind of quantitative restriction on the vocabulary of definitions would not be of any interest in itself, if it were not accompanied by other kinds of constraints both on a) the lexical, and on b) the grammatical side.

a) From the frequency list of the words used in definitions (about 800,000 word-occurrences, and 75,000 word-types), it appears in fact that some words have a much greater importance than in normal language, as evidenced by a comparison with the data of the *Lessico di Frequenza della Lingua Italiana Contemporanea* (Bortolini et al., 1971). These are the defining generic terms

which are traditionally used by lexicographers, such as ACT, EFFECT, PERSON, OBJECT, WHO, PROCESS, CAUSE, etc. It is not by chance that these same concepts are of relevance in many Artificial Intelligence systems.

b) Not only single words, or classes of words, are particularly relevant in the defining sublanguage. There are also lexical patterns and syntactic patterns which occur with great frequency, and which play a very special role in defining sentences.

The combination of these constraints can be and actually is very useful, when trying to exploit the information contained in definitions, and when transforming an archive of natural language definitions into a knowledge base, structured as a network. Some important parts of knowledge are in fact already retrievable in interactive mode from the Italian Lexical Data Base, which has recently been restructured.

Analyses on large corpora of definitions, carried out on many dictionaries (Amsler, 1980; Calzolari, 1983a, 1983b; Michiels, Noel, 1982) have in fact shown that the definitions sublanguage displays several regularities of lexical and syntactic occurrences and patterns. These general lexical classes and the classes of recurrent patterns can be more or less easily captured for instance by pattern-matching rules, and if possible characterized with formal rules.

III HYPONYMY RELATION

Hyponymy is the most important relation to be evidenced in a lexicon. Due to its taxonomic nature, it gives the lexicon, when implemented, a particular hierarchical structure: its result is obviously not a tree, but many tangled hierarchies (Amsler, 1980).

Instead of evidencing and labelling this relation by hand, I have tried to characterize it procedurally. The procedure which automatically coded (with a precision of more than 90% calculated on a random sample of 2000 definitions) true superordinates in all the definitions (approx. 185,000 for 103,000 lemmas), was based almost exclusively on the position of the "genus" term at the beginning of the definitional phrases, giving Nouns, Verbs, and Adjectives as superordinates of defined entries of the same lexical category. Ad hoc subroutines solved exceptional cases where a) quantifiers, or other modifiers preceded the genus term (e.g. ***aletta*** ---> piccolo ***gruppo*** di penne dietro l'angolo dell'ala), or b) more than one genus was present in the definition (e.g. ***assordare*** ---> ***attutire, smorzarsi*** detto di suono), or c) a prepositional phrase, usually of locative type, was at the beginning of the phrase (e.g. ***piazzato*** ---> nel rugby, ***calcio*** al pallone collocato sul terreno).

Even though the first immediate purpose of this procedure is of classificational nature, the ultimate goal is the extraction and formalization of the most relevant relationship between lexical items which is implicitly stored in any standard printed dictionary. It is in fact now possible to retrieve in the lexical data base not only all the definitions in which any possible word-form appears, together with the defined lemmas (e.g. SUONO appears in 328 definitions), but also to retrieve on-line, if desired, only the definitions in which the given word-form is used as a superordinate, therefore with the list of its hyponyms (e.g. the same word SUONO is used as superordinate of only 65 words, i.e. of a subset of the preceding set containing MUSICA, RUMORE, SQUILLO, SUSSURRO, etc.).

The query-language so far implemented for the lexical data base permits therefore to retrieve information on this hierarchical relation, identifying on-line the allowable interconnections within the entire lexicon. The links produced can be analyzed, evaluated, and, if necessary, interactively corrected.

From explorations on the trees thus obtained, we can also try to set up classes and subclasses of superordinates, on the basis of the upper nodes to which many other nodes are connected as descendants. Only as an example, the identification criterion for the noun-class "SET-OF" containing INSIEME, GRUPPO, COLLEZIONE, COMPLESSO, AGGREGATO, etc., among the set of noun-superordinates, is the fact that they are linked one to the other in the tree which results from querying the data base. Their hyponyms will obviously be for the most part collective nouns.

The identification of word-classes like this one leads to the next step in the formalization of the hyponymy relation, which will consist in the insertion of a label indicating a semantic class to these sets of superordinates. It will thus be possible to retrieve, for example, all the nouns generically definable as "SET-OF", independently of the particular word denoting a set used in definitions. Since it is already possible to trace these chains of hyponyms going upwards or downwards for more than one level, one can immediately ask whether, for example, MASSERIA belongs to the set of collectives even if it is defined as MANDRIA, because MANDRIA is defined as BRANCO, which is in turn defined as INSIEME, which finally is one of the nouns belonging to the class "SET-OF".

IV RESTRICTION RELATION

Even though some refinements are still required in order to improve the reliability of the automatic recovery of ISA-related terms chains, this kind of structural relation within the lexicon, that is hyponymy, is at a good stage of implementation in the Italian lexical data base.

Much still remains to be done as far as other very interesting relationships between the entries are concerned. I am now considering what could be called "restriction or modification" relation, since its purpose is to restrict or modify the meaning of the genus term. It is exemplified in the following definitions by the words in italics:

```
stannite     ---> calcopirite contenente stagno
arricciolare ---> modellare a forma di ricciolo
risonatore   ---> dispositivo atto a generare
                  risonanza
```

I wish to evaluate what could be done with respect to this kind of relation, starting from the available definitional data. One of the first aims of this lexicological research is to analyze, by means of computational tools, and to use the information contained in the different definitional formats and structures. The implementation of a number of procedures which convert the natural language information conveyed by definitions into processable formats, made up by structured relational links between lexical items or classes of lexical items, is now taken into consideration.

These formats can be made traceable e.g. in an Information Retrieval system on definitions, like the one actually implemented, on the entire corpus, for the taxonomic part of the lexical structure. But these formatted relational structures can also be used as starting points for a computationally exploitable reorganization of the definitional content. One of the characteristics of the definitional sublanguage, i.e. the presence of recurrent patterns (such as *proprio di, relativo a, prodotto da, originario di*, etc.), enables, at least in certain cases, to produce a constant mapping from certain variable types of more frequently detected definitional phrases to constant underlying relational structures.

Using rather simple pattern-matching procedures some classes and subclasses of definitions can be separated, and a small number of simpler types of definitions have already been converted into a formalized coded format also with regard to this restriction relation. A new virtual Relation is thus added to the original data base. The distinguished elements of a number of simple natural language patterns are mapped into some general structured information formats. Up to now, some of the definitions displaying the following restriction relations have been treated:

REL.FORM (e.g. *a forma di*)
REL.PROV (e.g. *provvisto di*)
REL.APT (e.g. *atto a*)

and the corresponding relational links generated.

Among the lexical variants of REL.PROV there are *fornito di, dotato di, munito di, pieno di, ricco di*, etc.; while REL.FORM groups the following variants of a different type: *in forma di, che ha (la) forma (di), di forma, di forma simile a (quella di), sotto forma di, avente forma di*, etc. It is thus possible, for example, to retrieve, among the 1271 definitions in which the word FORMA appears, only those defining something as "having the shape of something else". The implementation of these links allows to produce another kind of partitioning within the lexical system, and permits to better investigate the internal structure of words.

A procedure of the kind exemplified above, based on pattern-matching, is possible for a good number of definition types; for example, with a different format, for many adjectives:

```
       def               ┌ NP ┐
Adj   ---->>   REL.X    ┤    ├
                         └ VP ┘
```

where several groups of definitions are found to share a common underlying structure in terms of the restriction relation involved, in spite of other lexical and syntactic differences.

V FUTURE PERSPECTIVES

A comparison with the definitional corpora of other dictionaries, also of other languages, will certainly prove to be useful in establishing the set of the most general or primitive Relations, used for definition in lexicographical practice, often overlapping with the primitive Relations stated in many AI systems. These relations, mapped into a formal link in the data base, can then be paraphrased in each language, in the standard language.

The data base structure envisaged does permit both to maintain at a lower level (the starting level), and to eliminate at an upper level, many peculiarities and variations in the linguistic

expression of the same or of similar concepts or relations; their effect is to facilitate the comprehension by the users of the printed dictionary, inhibiting however immediate comprehension by procedural routines in the mechanical processing of dictionary data.

By applying similar methods of automatic conversion and mapping into suitable formats, as extensively as possible throughout the lexicon, many definitional expressions can be submitted to an attempt of standardization, thus achieving major precision, which gives a considerable improvement when performing, for example, information retrieval operations on the content of a dictionary.

This more structured, but, in another sense, simplified version of definitions, which also accounts for their relational nature, provides an excellent basis for testing and studying the "knowledge of the world" which underlies the structure of a dictionary.

VI REFERENCES

Alinei, M., La Struttura del Lessico, Bologna: Il Mulino, 1974.

Amsler, R.A., The Structure of the Merriam-Webster Pocket Dictionary, Ph.D. Thesis, Department of Computer Sciences, University of Texas, Austin, Texas, 1980.

Bortolini, U., Tagliavini, C., Zampolli, A., Lessico di Frequenza della Lingua Italiana Contemporanea, Milano: Garzanti, 1972.

Calzolari, N., "Towards the organization of lexical definitions on a data base structure", COLING82 Abstracts, ed. by E. Hajičovà, Prague: Charles University, 1982, 61-64.

Calzolari, N., "Lexical definitions in a computerized dictionary", Computers and Artificial Intelligence, II(1983a)3, 225-233.

Calzolari, N., "Semantic links and the dictionary", in Proceedings of the 6th International Conference on Computers and the Humanities, ed. by S.K.Burton, D.D.Short, Rockville (Maryland): Computer Science Press, 1983b, 47-50.

Calzolari, N., Ceccotti, M.L., "Organizing a large scale lexical database dictionary", Actes du Congrès Informatique et Sciences Humaines, Liège: L.A.S.L.A., 1981, 155-163.

Clark, E.V., Clark, H.H., "When nouns surface as verbs", Language, 55(1979)4, 767-811.

Evens, M.W., Litowitz, B.E., Markowitz, J.A., Smith, R.N., Werner, O., Lexical-Semantic Relations: a Comparative Survey, Edmonton, Alberta: Linguistic Research Inc., 1980.

Findler, N.V. (ed.), Associative Networks, New York: Academic Press, 1979.

Hendrix, G.G., "Natural-language interface", Proceedings of the Workshop 'Applied Computational Linguistics in Perspective', American Journal of Computational Linguistics, 8(1982)2, 56-61.

Michiels, A., Müllenders, J., Noël, J., "Exploiting a large data base by Longman", COLING80: Proceedings of the 8th International Conference on Computational Linguistics, Tokyo, 1980, 374-382.

Michiels, A., Noël, J., "Approaches to thesaurus production", COLING82: Proceedings of the Ninth International Conference on Computational Linguistics, ed. by J.Horecky, Amsterdam: North-Holland, 1982, 227-232.

Nagao, M., Tsujii, J., Ueda, Y., Takiyama, M., "An attempt to computerize dictionary data bases", COLING80: Proceedings of the 8th International Conference on Computational Linguistics, Tokyo, 1980, 534-542.

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

Smith, R.N., "On defining adjectives: part III", Dictionaries, the Journal of the Dictionary Society of North America, Winter, (1981)3, 28-38.

Smith, R.N., Maxwell, E., "An English dictionary for computerized syntactic and semantic processing", in Computational and Mathematical Linguistics, ed. by A.Zampolli, N.Calzolari, Firenze: Olschki, 1977, 303-322.

Walker, D.E., Amsler, R.A., Proposal to the National Science Foundation on an Invitational Workshop on Machine-Readable Dictionaries, SRI, 1982 (mimeo).

Zingarelli, N., Vocabolario della lingua italiana, Bologna: Zanichelli, 1971.