# TOWARDS THE ORGANIZATION OF LEXICAL DEFINITIONS ON A DATABASE STRUCTURE

Nicoletta Calzolari

Istituto di Glottologia - Universitá di Pisa, Italy

Printed dictionaries are great repositories of information, and it is important that they can be exploited as fully as possible, with regard to all the different types of data they contain. This was one of the aims when organizing the Machine Dictionary of the Italian language on a database structure.

The design and organization of the lexical database for the first two relations implemented, i.e. the set of Lemmas (106, 091) and the set of Word-forms (1,016,320), has been described in other papers (see for example Calzolari and Ceccotti, 1980).

These two very large archives are maintained continuously on-line and are interactively invoked through a query language which permits to the user to access, in transparent mode, the data, and to have his particular "view" of the data. The database concept and methodology give rise, in fact, to a radical change in perspective when confronted with sequential organization of data. We have a dynamic rather than a static object which is flexible and easy to query, update, extend.

This lexical database is now being extended by the insertion of lexical definitions (185,899) and semantic data. The guiding principle behind this project is the conviction that the study of the defining vocabulary of an actual dictionary can provide a precious tool in the semantic analysis

of a language (see Noel, 1981).

The logical organization of this definitional information is not a trivial task, and must be performed bearing in mind the goals to be achieved. It must in fact be possible to have direct access to each and every piece of information contained in the definitions. The significance of "piece of information" in this context is in direct relationship to the eventual use to be made of it. By "piece of information" inside the definitions, we intend not only the single word-forms, as they are written in the definitions, but also the lemma, to which every word-form is connected; moreover, at a further stage of analysis, the specific sense of every polysemic lemma in the particular context (context=definition) must be considered.

The logical organization of the definitional part of the database must, therefore, be structured to provide, for each word in every definition, direct access to: a) the word-form itself, with the associated information (morphological, usage level, etc.); b) the lemma to which the word-form pertains, with the associated information (part-of speech, variants, usage level other word forms i.e. paradigm); c) the specific sense of the lemma. The implementation of a definitional archive thus requires an enormous task of disambiguation at all the three levels: word-forms, lemmas and senses, in order to produce material which can be used effectively to extract semantic information from the dictionary.

The first step in this direction is the lemmatization of the definitions themselves. For this task, the other two archives of the database (the word-form and lemmas archives) are being used, together with ad hoc procedures, to produce an automatic lemmatization of a large percentage of the words contained in the definitions. For the other words, those for which automatic lemmatization has not yet been achieved, a disambiguation strategy has been developed in which the human

operator works interactively with the computer, and the computer can memorize choices on homographic forms as they are made.

After lemmatization, each word is associated in the computer memory to the addresses of its word-form and of its lemma. Therefore, the definitions are organized in the memory not as actual strings of words, but as lists of addresses of word-forms and lemmas. In this way, a number of important results are achieved: a) a great reduction in storage size; b) data types (addresses i.e. binary numbers) which are easily handled by the computer; c) data which are strictly associated to the first two archives, aiming at the eventual construction of an integrated system; d) much more rapid data processing and direct accesses to each kind of data, in each position of the definition itself; e) the possibility of being able to immediately retranslate addresses into character strings, and list of addresses into phrases, i.e. definitions; f) the possibility of correcting, updating and inserting within the definitions.

Only once this preliminary stage has been completed is it possible to extract many kinds of semantic information from the dictionary. The memorized definitions have an internal logical structure which permits the construction of semantic chains (to evidence taxonomic relationships) and also of other types of semantic links (to evidence other types of semantic relationships, such as 'part of', 'set of', 'in the form of', 'apt to', etc.) between words in the lexicon. These chains and links, which can be not only displayed, but also handled by computer procedures in many different ways, surely provide a good starting point for the study of the semantic structure of the lexicon. In fact, it is hoped that the computerized dictionary will offer a model of the Italian lexical system in the various aspects which can be associated with a lexicon (phonology, morphology, syntax i.e. verbal frames,

lexical semantics). This approach is included in the general theoretical view which considers the lexicon as a central reference point both for language analysis and for many linguistic applications.

References:

Calzolari, N., M. L. Ceccotti, "A project for an exhaustive lexical database system", in Proceedings of the Second International Conference on Data Bases in the Humanities and Social Sciences, 1980, Madrid, in press.

Noel, J., "The Longman-Liege Dictionary project", Congres International Informatique et Sciences Humaines, Liege, 18-21 nov. 1981.

Procter, P., "Problems in dictionary making", Congres International Informatique et Sciences Humaines, Liege, 18-21 nov. 1981.