Organizing a large scale lexical database

by

Nicoletta CALZOLARI

Luigia CECCOTTI

Università di Pisa - ITALY

As the computer has now been generally accepted as an essential tool in dealing with large volumes of data of all types, there is no need to stress its great importance in handling linguistic data. The first applications of computational tools to linguistics were concentrated on text analysis where the manipulation of very large volumes of data (the texts) was involved. Other more recent and very active research work has been and is being conducted in the field of Artificial Intelligence where a limited part of the world is analyzed and decomposed in depth with the help of sophisticated mathematical, logical and computational means.

In the recent years, however, another field of research dealing with large amounts of linguistic data, but not directly with texts, has been developed. This is mainly concerned with the computational organization of lexicological data. Several projects (see /9/, /11/, /12/, /13/), including the DMI/DB (Italian Machine Dictionary as a Data Base) project of Pisa (see /3/), are now underway aimed at creating large lexical archives or lexical data banks which collect lexical information of different kinds on many levels (phonetical, morphological, syntactic, semantic, etc.), and can be used as a linguistic tool for many different researches or applications.

Within the specific domain of the organization of very large lexical archives, we should like to highlight, in this communication, the necessary contribution not only of generic computer tools, but more particularly of the database method and technology.

A database system is designed to memorize and organize a large quantity of data about a particular topic, so that it is possible to answer questions concerning this topic made by generic users. Many techniques have been developed and are now being developed in order to obtain a more efficient representation, storage and retrieval of a large amount of information. The adoption of the database concept gives rise to a radical change in perspective, as a very different and a much more flexible and complete utilization of the same data is permitted, and thus new applications are possible. In fact, it is now possible :

a) to have direct access to each type of data and to their attributes;
b) to effect a continual and rapid updating of the data, whenever required by new theoretical or applicative needs;
c) to extend the database into a logically more complex and complete structure.

The DMI/DB

The Italian Machine Dictionary organized on a database structure can be conceived as a nucleus around which language analyses at many levels, text processing, terminological data banks, automatic documentation, linguistic studies and researches which refer to the Italian lexicon, can be centred.

In the creation of two basic archives of the DMI/DB, we have found that the relational model (see /7/) provides the most suitable data structure. In designing our relational database, we have

defined a set of relations which describe the facts to be represented and which function as the user interface to the database. We have also determined the associations between entities or objects of different relations.

The logical structure of our lexical database is represented in Fig. 1. The boxes represent the primary files or relations; the arrows between the boxes represent the links between the relations, or associations; the names linked to the boxes represent the various secondary keys for direct access to the files.

The alternate indexes with the secondary keys permit direct access to the records through the value of other attributes, other than the lemma or the form, and these need not be unique keys.

In addition, we can use also a substring of each key to access or to select records. For example, one can select the subset of lemmas beginning with the Prefix RI- :

Initial.lem : RI ---->	RIABBAIARE	VI
	RIABBANDONARE	VTR
	RIABBARBICARSI	VIP
	RIABBASSAMENTO	SM
	RIABBASSARE	VTRI
	. . .	. . .

Similarly, by using a secondary key, all the lemmas belonging to a certain part-of-speech can be obtained (e.g. all the Substantive and Feminine, or Substantive and Feminine and in Locution) :

Gramm. Cat. : SF ---->	ABBREVIAZIONE
	AZIONE
	BATTAGLIA
	CALCOLATRICE
	. . .
Gramm. Cat. :SFL ---->	CRESTA
	DETTA
	RINFUSA
	. . .

Obviously, selections on the various keys, and on different relations can be combined. For example, only Verbs beginning with RI- can be requested (obtaining all the previous records but one), or subsets of lemmas with given Suffixes can be selected, e.g. with the Suffix -ZIONE :

Suffix. lem : ZIONE --->	AZIONE	SF
	AZIONE	I

SILLABAZIONE	SF
CIBAZIONE	SF
LIBAZIONE	SF
. . .	. . .

Furthermore, it is possible to connect each lemma with its definitions, to see, in this case, if particular endings select particular types of semantic markers or formats of definitions. This could be a useful tool in the analysis of derivatives in the lexicon.

These alternate indexes make it possible, therefore, to generate multiple views of the same data. The data can, in fact, be ordered so that only those occurrences which satisfy one or more conditions can be accessed.

Other examples, also obtained by querying the database in interactive mode, concern the relationship between a lemma and its variants (which may be graphic, inflexional, morphological . . .) :

Variants : GIOCO ---->	GIUOCO	SM
	IOCO	SM 1
Variants : IO ------>	EO	PQ 1
	ME	PQ
	MECO	PQ 3
	MEVE	PQ 1
	MI	PQ
	NOI	PQ

Viceversa, we can enter a variant and ask whether it points to a basic lemma :

Pointer : LA R ---->	IL	R
LA PQ ---->	ESSA	PQ
LA SM ---->	without pointer	

The relations which have been implemented so far are the two basic relations of the entire archive, i.e. LEMMARIO (a set of 106,091 lemmas and associated relevant information) and FORMARIO (a set of 1,016,320 word-forms and associated information). Fig. 2 gives some information about the dimensions of the primary relations (or clusters), and of the inverted files (or alternate indexes) automatically built on them to provide direct access to all the various types of information recorded (e.g. morphological codes, graphic or morphological variants, etc.). Obviously, a bidirectional relationship (l, m) between lemmas and respective word-forms is also possible.

Despite the large dimensions of the two archives, the query response times are very quick. We feel that this confirms the validity of the kind of approach we have chosen to memorize the word-forms. After having taken into consideration and subsequently discarding the possibility of storing a set of roots or stems and associating them to the corresponding suffixes, we decided to memorize the complete list of more than 1,000,000 word-forms. We have found that the great reduction in the response times, obtained by listing all the forms and thus doing away with the need for recognition and segmentation algorithms, largely compensates for the greater volumes of space occupied on secondary storage supports. The amount of secondary storage occupied has become even less of a problem recently as a mass memory storage system has become available to us. We have used mass memory storage to allocate our database. Mass memory is in fact very similar to disk storage without some defects of disks, and we can have the entire database, despite its great size, always on-line and interactively accessible from our virtual machines and video terminals.

In addition to the obvious advantages which a database system offers for updating (inserting, correcting, deleting) data, we have designed this lexical database to be a very flexible tool for studies on the Italian lexical system and also for various linguistic applications. For example, these facilities are very useful in the improvement of text processing, in order to give the possibility of lemmatization at different levels and with different criteria. Moreover, only with a system of this kind, is it possible to conceive a lemmatization process in text order rather than alphabetical order, so that a parsing algorithm for the disambiguation of homographs through analysis of the immediate left and right contexts can be associated to the dictionary.

The very next step in our project will be to reorganize the definitions associated to the lemmas, in order to render more easily feasible a systematic study of the semantic or definitional aspects of the Italian lexical system applying automatic research methods. Some previous attempts to restructure definitional data (see /2/, /5/, /6/) led to the conclusion that a completely different organization of the set of definitions from that actually implemented is essential in order to achieve this type of study.

It is important to make a final consideration. In order to be able to work on a project of this type, it is essential to have experts available who are competent at the same time in the linguistic and the computer science fields and this, unfortunately, is not always easy.

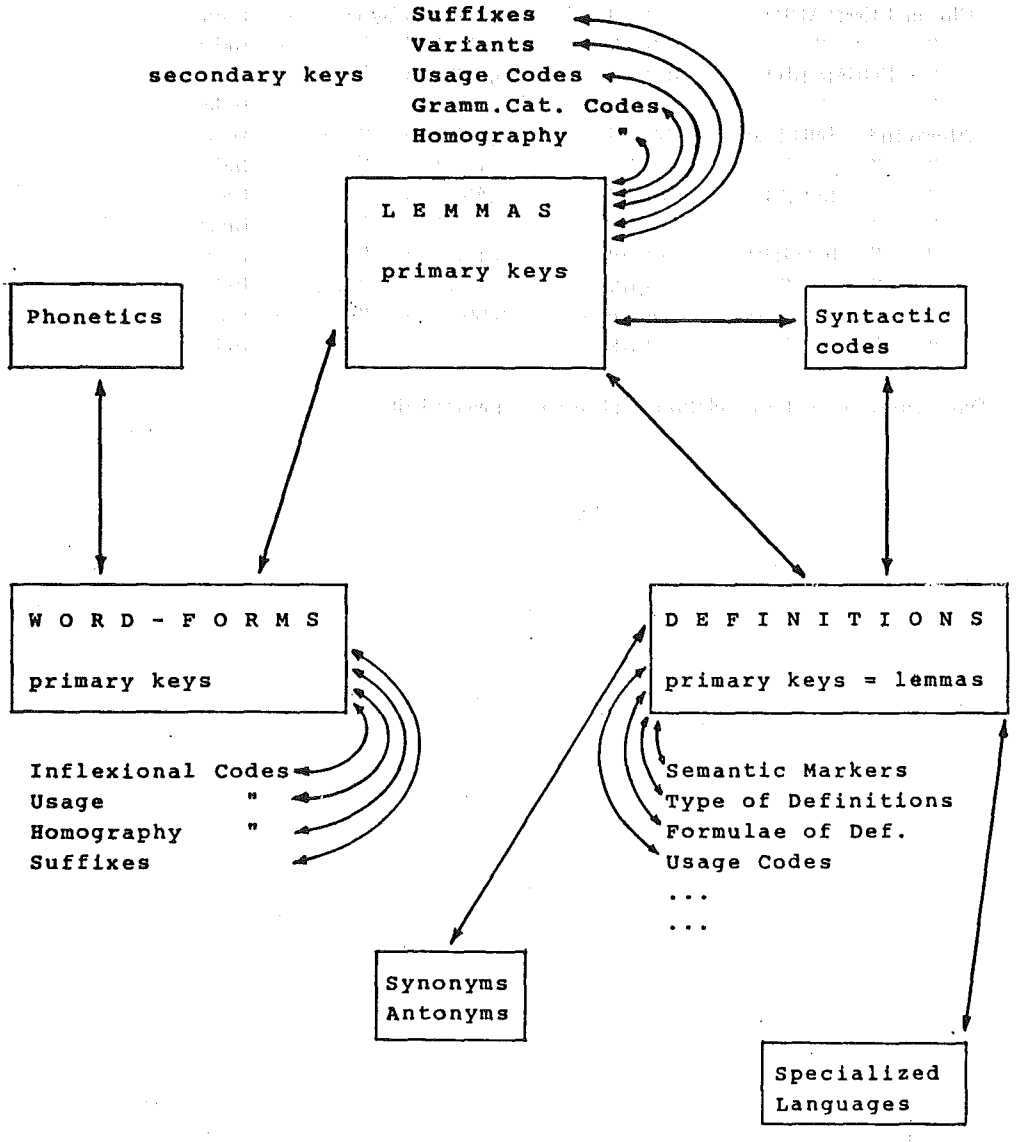

Fig. 1. The logical structure of the lexical database.

Cluster LEMMARIO	53 cyl	12,374,016 bytes	—→	Data
" "	5 trk	56,320 "	—→	Index
" FORMARIO	199 cyl	46,460,928 "	—→	Data
" "	19 trk	214,016 "	—→	Index
Altern.Index IND.FOR	135 cyl	31,518,720 "	—→	Data
" " "	13 trk	146,432 "	—→	Index
" " IND.FLEX	19 cyl	4,435,968 "	—→	Data
" " "	2 trk	22,528 "	—→	Index
" " IND.INVF	22 cyl	5,136,384 "	—→	Data
" " "	3 trk	33,792 "	—→	Index
" " IND.LEM	24 cyl	5,603,328 "	—→	Data
" " "	3 trk	33,792 "	—→	Index

Fig. 2. Dimensions of the basic relations and of some inverted files.

REFERENCES

/1/ Bahr, J., 1978, Reflections on the project of a lexical data bank, *Cahiers de Lexicologie*, 32 (I), pp. 55-64.

/2/ Calzolari, N., 1977, An empirical approach to circularity in dictionary definitions, *Cahiers de Lexicologie*, 31 (II), pp. 118-128.

/3/ Calzolari, N., Ceccotti, M.-L., (forthcoming), A project for an exhaustive lexical data base system, in *Proceedings of the Second International Conference on Data Bases in the Humanities and Social Sciences*, 16-19 June 1980, Madrid.

/4/ Calzolari, N., Ceccotti, M.-L., 1980, Una base di dati lessicale, *Proceedings of AICA Annual Congress*, 1980, Bologna, pp. 359-362.

/5/ Calzolari, N., Moretti, L., 1976, A method for a normalization and a possible algorithmic treatment of definitions in the Italian dictionary, Coling 76, Ottawa.

/6/ Calzolari, N., Pecchia, L., Zampolli, A., 1980, Working on the Italian Machine Dictionary : a semantic approach, in A. Zampolli, N. Calzolari (eds.), *Computational and Mathematical Linguistics*, Firenze, Olschki, vol. II, pp. 49-69.

/7/ Codd, E.F., 1970, A relational model of data for large shared data banks, *Comm. ACM.*, XIII (6).

/8/ Date, C.J., 1977,*An Introduction to Database Systems*, Reading (Mass.), Addison-Wesley.

/9/ Engels, L., 1980, Verslag IPEK-Werking 1978-1980, K.U. Leuven, Subgroep Engelse taalbeheersing (woordenschat), K.U.L., Leuven, Rapport interne.

/10/ Gruppo di Pisa, 1979, Il Dizionario di Macchina dell'Italiano, *Linguaggi e Formalizzazioni*, Atti del Convegno Internazionale di Studi, Catania 1976, a cura di D. Gambarara, F.Lo Piparo, G. Ruggiero, Roma, Bulzoni, pp. 683-707.

/11/ Michiels A., Mullenders J., Noel J., 1980, Exploiting a large data base by Longman, *Coling 1980*, Tokyo, pp. 374-382.

/12/ Nagao, M., Tsujii, Y., Mitamura, K., Hirakawa H., Kume M., 1980, A machine translation system from Japanese into English, *Coling 1980*, Tokyo, pp. 414-423.

/13/ Svartivk, J., Quirk, R., 1980, *A Corpus of Spoken English*, Lund.

/14/ Tsichritzis, D.C., Lochovsky, F.H., 1977, *Data Base Management System*, New York, Academic Press.

/15/ Zampolli, A., 1975, L'elaborazione elettronica dei dati linguistici : stato delle ricerche e prospettive, Colloquio sul tema : Le tecniche di classificazione e loro applicazione linguistica, Roma, Accademia Nazionale dei Lincei.