The Dictionary and the Language Learner

Papers from the EURALEX Seminar
at the University of Leeds, 1–3 April 1985

edited by Anthony Cowie

Offprint

Max Niemeyer Verlag
Tübingen 1987

Nicoletta Calzolari, Eugenio Picchi and Antonio Zampolli

THE USE OF COMPUTERS IN LEXICOGRAPHY AND LEXICOLOGY

A commonly accepted classification of the uses of computers in lexicology and lexicography does not yet exist. For practical reasons, we shall divide these uses into two main groups:

(A) The use of computers in dictionary making.

(B) Lexicographical data bases.

A. Dictionary making

The uses of computers as an aid in dictionary making are usually subdivided into three main groups, which correspond to three operational stages:

(1) data collection;
(2) lexical entry preparation;
(3) editing and printing.

A.1. Data Collection

Two main types of data are collected as documentary sources:

A.1.1. Pre-existing dictionaries

Some lexicographical enterprises have decided to convert into machine-readable form appropriately chosen printed lexicographical resources (general or technical dictionaries, lexicons, etc.), as an aid for the creation of new dictionaries. Once in machine-readable form (MRF) these resources, as well as dictionaries already independently available in MRF, may be consulted and exploited, during the editing of a new dictionary, by using multiple-access techniques, as described later in section B.

A.1.2. Citations excerpted from a corpus

Traditionally, human excerptors were set to read through various texts, selecting what they noted as unusual or particularly informative examples and copying each textual citation onto a paper citation slip.

As far as historical dictionaries are concerned, the "excerption density" was usually very low (1%). Only a small fraction of the corpus, which is often called the "basic archive", "could be read at a much higher excerption density, say 25 or 30%, so as to take in not only the more unusual and special - if you like, lexicographically interesting - examples, but also a large representation of the more common-place uses of common words" (Aitken 1983:34).

Even so, those collections may contain several hundred thousands of quotation-slips (for example, the OED in its present form contains about 1,820,000 quotations, selected from nearly six million quotation slips).

The use of computers to produce different kinds of indexes, concordances and quotation-slips from a text or a corpus is today a routine task, at least in those cases in which the researcher has access to already available software. Relevant software packages at present exist both for mainframes and for personal computers. Although different, the various procedures prepared by academic or commercial specialized centres follow a common logical scheme, which consists of three main steps:

- acquisition of texts in machine-readable form;
- lemmatization;
- production of textual documentation.

A.1.2.1. Acquisition of texts in machine-readable form

There are several alternatives when converting textual material into MRF: key-punch; key to paper tape; on-line typewriter; key to disk; selectric typewriter; keyboarding on visual display terminals; etc. All these require that someone types the text on a keyboard, and until recently this manual operation was a major obstacle, owing to its cost.

Nowadays, the need to type a text manually is obviated by the existence of four major sources of texts in MRF:

omni-font scanners: devices able to recognize graphic characters and to record them on magnetic support;

textual archives: in many countries specialized centres are collecting texts in MRF, and copies are often distributed on request, under particular conditions;

photocomposition; many texts (books, newspapers, etc.) are now recorded for photocomposition by publishing-houses in MRF, as input for photocomposition;

word-processing: "the airwaves and cables of the information society are already filled with electronic digital texts" (Amsler 1983), and in the office automation framework the majority of texts are electronically digited for word processing purposes.

If these sources are to be utilized, a number of problems must be solved:

organizational: how to arrange the exchange of information and data between different textual archives;

legal: how to deal with the copyright of the different "owners": authors, publishing houses, those responsible for textual archives, etc.;

scientific: some minimal norms are needed, the adoption of which will guarantee that the representation of a text in MRF contains all the minimal information requested by linguistic, lexicographical or philological processing.

The availability of corpora in MRF varies with different languages. Some academic/historical dictionary projects are collecting and processing large textual corpora. Well-known examples are the "Trésor de la langue française" (Nancy), "Tesoro italiano delle origini" (Firenze, Pisa), "The Dictionary of Hebrew" (Jerusalem), "Old Spanish" (Madison, Illinois), "Old English" (Toronto), etc.

Corpora collected to provide statistical information for "frequency dictionaries" may also be considered. However, they are usually rather small and sometimes selected according to sampling criteria rather unsuitable for lexicographical use.[1]

It seems worthwhile to consider the feasibility of promoting action, at a national[2] or international level, for the creation for each language of corpora in MRF, which may serve as a reference

[1] See for example the Brown Corpus, the LIF Corpus, etc.

[2] The Swedish Språkdata may be cited as a model (Allen 1983).

corpus for a variety of research tasks, including lexicographical projects. Two major problems are obviously to be taken into account:

- responsibility for the maintenance and updating of a corpus, once it has been created;
- the copyright problem for different uses, at an academic and commercial level.

A.1.2.2. Lemmatization

After a text has been converted into MRF, it is possible to ask the computer to produce immediately the documentation (concordances, frequencies, citation slips, etc.) regarding graphic forms, which are the only linguistic units explicitly represented in the printed text. If this documentation is instead to take account of other linguistic units (lemmas, collocations, syntagms, etc.), it is necessary to introduce the explicit representation of these units in the text, before the production of concordances, indices, etc., can begin.

The dilemma "to lemmatize or not to lemmatize" the concordances and indexes dates back to the very beginnings of the use of computers in linguistic and philological text-processing.

The answer depends on various factors. The decision to lemmatize is usually taken for highly inflected languages, and for linguistic or lexicographic analysis rather than for philological or literary applications. Two of the main obstacles to lemmatization are:

- the lack of precise, widely accepted, linguistic norms, ensuring comparability and reutilization of lemmatizations performed by different researchers (Tombeur 1983);
- the high cost of lemmatization, especially if performed entirely manually.

Although lemmatization is in most cases still completely manual, various semi-automatic systems have been experimented with. They are usually based on the distinction between word-forms which are considered univocal (i.e. pertaining to only one lemma) in a given lexical system, and forms which are considered homographic. The former are directly lemmatized by the program. The latter are submitted to the analysis of the lemmatizer. (For a description, see Zampolli 1983).

Disambiguation of homographs is a very time-consuming task. However, so far only few attempts have been made with regard to (semi-)automatic disambiguation. These can be subdivided into two basic approaches:

(1) Local disambiguation

Algorithms usually try to solve homography between forms belonging to different parts of speech. For each possible pair of parts of speech, rules are formulated which examine the context immediately surrounding the homograph, up to a certain number of contiguous words. Specific elements (words or grammatical categories) or sequences of elements are searched for. The formulation of these rules is essentially based on the concept of "impossibility" or "possibility" of co-occurrence of a pair of given words and/or word-classes in a specified "span" of context. These rules are often aided by statistical algorithms which use quantitative information, obtained by examining previously lemmatized texts, on the frequency distribution of words and grammatical categories in the immediate context of the homographs in a sample corpus. At present, an estimated 80-90% of words in a text can be successfully analyzed in certain systems (Ratti 1982). The remaining 10-20% are submitted to an interactive manual disambiguation process.

(2) Disambiguation via syntactic-semantic parsers

Automatic parsing of natural language texts has always been the main interest of computational linguists. By contrast, until now lexicographers have never shown any particular interest in parsing systems. In computational lexicology and lexicography, parsers could be used for different purposes: e.g. to study collocations, to identify particular syntactic-semantic patterns for the choice of "lexicographically interesting" quotations, to analyze the definitions given in a dictionary, etc. Here we shall focus on the possible use of a parser as a homograph disambiguator. If the sentence is not in itself ambiguous and its overall syntactic structure is recognized, the parser will obviously accept only one of the possible grammatical "analyses" of a homographic form.

The problem is that the presently existing parsers are unable - to our knowledge - to treat exhaustively the large variety and quantity of phenomena present in the types of corpora which usually constitute the basic documentation of the lexicographer. It would be interesting to explore the feasibility of creating less ambitious parsers able only to identify surface constituents, and to quantatively evaluate their efficiency for the disambiguation of homographs in the process of lemmatization.

A.1.2.3. Production of textual documentation

Each specialized Centre has its own procedures to produce the results usually requested:[3] direct and reverse indexes, various types of frequency distribution, rhyme-indexes, index locorum, concordances, quotation slips, etc. Some software packages are parametrized, i.e. they enable their users to specify a number of "processing options", which may look at the nature and ordering of the entries, selectivity versus completeness, different contextualization criteria, etc.

With regard to data collection for dictionary making, it is not yet clear which type of documentation is more convenient. Some are of the opinion that the best solution is to continue to produce contexts printed separately onto citation-slips, or in the form of concordances. Others prefer the possibility of displaying selected contexts on the screen on request.

Some academic projects, which accurately lemmatize the texts during the data collection phase, pre-select particularly significant examples, so that the editor, when compiling a lexical entry, works on contexts which have already been strongly reduced in number, and are grouped under their own lemma. In other projects, the editor is presented with all the contexts of the corpus. He has to accomplish at the same time both the task of selecting the examples and of grouping together the different graphic forms, separating the

[3] Very well known are for example COCOA of the Oxford Computing Centre. In Italy, almost all ongoing projects use the "procedure di spoglio" of the Institute for Computational Linguistics of CNR (National Research Council).

homographs, etc.[4] The experimental data presently available still seem to be insufficient as a basis for a final choice between the two alternatives.

A.2. Preparation of Dictionary Entries

It is useful, in our opinion, to distinguish between two different frameworks:

A.2.1. First framework

The dictionary is centred around the descriptive content: definitions, syntactic information, etc. Some methodology and software tools are already being used, mainly in a commercial context.

In particular, they assist the lexicographer by:

- reducing the work connected with the handling of the graphic conventions, usually so bulky in dictionaries;
- making it easier to retrieve previously stored information;
- ensuring automatically the formal coherence of the information.

As a typical example we may quote the *Compulexis* dictionary system.[5]

A common interesting feature consists in allowing the data to be entered in typographically neutral form. A system of tags is used to store the data-type of each lexicographical entity of the dictionary: examples, idioms, translations, parts of speech, syntactic information, etc. Those systems automatically assign a specific typographical representation to each data-type, generate separators (commas, semicolons, etc.) between the various elements, and generate fixed repetitive text elements.

[4] The arguments involved are both economic (time and cost: does the time gained by the editor compensate for the lemmatization pre-editing work?) and scientific (is it possible for the "lemmatizing" researcher, often performing text-by-text work, to select the quotations to be retained compatibly with the interests of the editor when writing the lexical entry?)

[5] "The system is designed as a complete set of tools for the development and handling of monolingual as well as bilingual dictionaries. It serves three different but related purposes, namely:
- Editorial tools for the compiler/editor
- Data bases for further lexicographical development
- Type-setting tools

(...) The principle of the system is to allow the user to input, output and retrieve the exact data in the dictionary."

Furthermore, the tags may direct the system to form a data base from which other products may be retrieved, either as complete products or as the basis for further (strongly reduced) editing. A system of this type requires the development of a fairly generalized taxonomy of the type of information that can appear in dictionaries. Within this framework, however, the computer assists the compiler only in the formal part of his work.

A.2.2. Second framework

The second framework concerns dictionaries strongly based on the classification and ordering of a very large number of quotations from a corpus, as for example large historical academic dictionaries.

The editorial analysis consists of a "rapid shuffling and reshuffling of examples" (Aitken 1983:41), in an iterative process in which the editor selects from the archive "significant" quotations and tentatively arranges them into senses with provisional definitions, thus progressively constructing the microstructure of the entry. Furthermore, he may wish to consult different types of complementary sources of information, ranging from bibliographic references to relevant entries of pre-existing dictionaries.

Computing facilities are not yet widely used in this editing stage, and much work is still necessary to take full advantage of the potential benefits offered by computational linguistic know-how and methodology.

A.2.2.1. Selection of quotations

Given the quantity of contexts produced by computational text processing, the problem of reducing the number of quotations to be treated editorially, by a process of representative selection, appears in certain cases to be very urgent. In other words, there are often far too many computer-produced quotations for the lexicographer to go through manually, and strategies are needed to (semi-)automatically screen the material.

Among the solutions suggested for words of high frequency, the simplest is to instruct the computer to select only one context in every *n* (where *n* grows progressively (10, 20 ... 100) with the frequency of the word).

A more refined system consists in making the computer select significant collocations.

There are several possible statistical aids. One of these is to "start from the observed frequency between node and collocate, compare it to an expected one (based on the frequency of both elements in the corpus) and then to evaluate the possible difference between observed and expected values by means of a standard deviation" (Martin et al 1983:85).

Another possibility is to search systematically in the archive for a predetermined construction for a given word.

> "Par exemple, pour la rédaction de *debout* disposer des exemples de *debout* en emploi interjectif *debout!*; pour la rédaction d'*homme* explorer les constructions du type *homme à* + infinitif" (Gorcy 1983:122).

This type of interactive question-answering obviously requires a well constructed interrogation language, operating both on word forms and on grammatical taggings. The ultimate goal would obviously be to have at one's disposal a parser capable of producing automatically a description of the syntactic-semantic structure of the text. In this case it would be possible to identify quotations which exemplify in the corpus occurrences of particular syntactic and semantic patterns. Unfortunately, as already noted above, the existing parsers are not yet capable of exhaustively treating the variety and quantity of phenomena present in a corpus. However something new is now moving in this sector. Examples are the DEREDEC system (Montréal) (Plante 1983) and research at the Institute for Computational Linguistics in Pisa.

A.2.2.2. Towards a lexicographical workstation

A major challenge is to develop a "lexicographic workstation", by which the lexicographer "preparing" the description of dictionary entries can interact with a lexical data base, conceived as a set of different knowledge sources (text corpus, old lexicographical archives/deposits, pre-existing dictionaries, bibliographical references, etc.) made available on-line and accessible by means of appropriately designed software tools.

According to some scholars, the publication of a dictionary of the future will represent a design decision made at an editor's workstation, in which components of an underlying lexical data base are "sculpted" together in an attractive visual form, without changing any of the underlying computer data.

This prospect may not be so far away for the updating of existing traditional dictionaries. The lexicographer could effectively examine and characterize newly found citations automatically extracted from an incoming text stream, modifying and creating lexical entries in an existing data base, thus continually updating the dictionary to remain contemporary with the use of the language (Amsler 1983).

Far more complex is the situation where one devises an integrated system for the compilation of new dictionaries, in which the gap between data collection via electronic text processing and final photocomposition is filled by computer-assisted editing of the entries. Some experimental projects aim essentially at facilitating access to a corpus and to a dictionary, and the storing of preliminary versions of lexical entries for further processing (Lentz 1981, Zampolli 1983).

A software component ensures quick access to the data, thus enabling the lexicographer to use the corpora interactively via the terminal. For example, the lexicographer can search for specific word forms, word forms matching (beginning, containing or ending with) a specified string of graphemes, co-occurrences of word forms and/or grapheme strings in a given span of text (if the texts are already lemmatized, the lexicographer may operate on lemmas and/or word forms). The component provides the lexicographer with information on the frequencies of distribution, in different sections of the corpus, of the searched elements. The lexicographer may then request the contexts to be displayed on the video screen or to be output in printed form. Each context, which is algorithmically "cut out" by the computer, may be interactively modified by the addition or exclusion of selected syntagms (Picchi 1983).

Another component enables the lexicographer to consult existing dictionaries in the database. The lexicographer will obviously benefit from the multiple dictionary access techniques described in B, may search for different types of information in the entries both of existing dictionaries and of the dictionary which is being constructed.

Specific functions permit the insertion and cyclic reordering of the selected contexts in the different sections of the microstructure, thus producing a preliminary version of the new dictionary entry. All stored information can be altered, expanded and corrected at any time and consulted immediately for comparison within the new dictionary, in order to ensure homogeneity and coherence.

We feel that a greater cooperative effort between lexicographers and computational linguists is needed if a complete procedure is to be constructed. In particular, the operations which the lexicographer performs when preparing a dictionary entry must be analysed and described accurately.

The objection of many lexicographers is that "no computer system offers a way for the editor to shuffle and re-shuffle examples, of which the editor's work so largely consists" (Kipfer 1982). They think that the "traditional dictionary-slips-on-the-table method" is still the best because "the computer is limited in the number of slips one can see on one video-screen" (Kipfer 1982). They suggest that editors continue to produce concordances or, even better, citation slips, which can be used in the traditional manner. In order to avoid retyping the selected citations, they suggest that the citations stored in the computer's memory should be numbered. The editor then keys in only the microstructure (headword, etymologies, grammatical information, definitions, etc.) and for each section keys in the code numbers of the citations he wants.

The first explicit and general discussion of this problem was probably held during a round-table meeting between computational linguists and lexicographers from more than ten countries, held in Pisa in 1972.

The situation today is probably somewhat different due to the evolution of data-base methodologies and of workstation technology, which seem to offer the opportunity of "simulating" on the video the games of "solitaire" which the lexicographer has always played, ordering and reordering the traditional slips.

A.3 Editing and Printing

Photocomposition techniques are now commonly used by most publishing houses. A variety of editorial controls and readjustments to the text of the dictionary prepared in MRF for photocomposition are thus possible before the final printing. Quémada (1983:27) provides some examples.[6] Other examples are given by Knowles (1983:186-87), Howlett (1983:157) Petersen (1983) and Pfister (1983).

Even more effective controls are obviously possible if the dictionary is prepared directly in MRF (as in the frameworks described in A.2.), following a tagging scheme. This might cover: control of different types of cross-references; print-out of lists of words and expressions with specialized meaning to be submitted to experts; automatic verification of the coherence of the typographical conventions, etc. In historical dictionaries, it is possible to retrieve exhaustive lists of the citations of a given author from a work, so as to re-control them in the original text, or to replace them in the case of availability of new critical editions.

The advantages of working on a dictionary in MRF are obvious when revising or updating a pre-existing dictionary.

B. Lexical data bases

B.1. Typology of MRDs

The expression "Machine Readable Dictionary" (MRD) is increasingly employed nowadays in the field of Computational Linguistics. The expression is however employed within different frameworks, and with different meanings according to different objects, and is applied either to different approaches to an identical underlying generic notion or to notions which are distinct one from the other.

In order to clarify the terminology, we first wish to draw up a tentative typology of MRDs, listing at least the principal types which are usually denoted by this expression. We should however

[6] "Citons en particulier, pour la mise au point de la nomenclature: les inventaires cumulatifs des entrées figurant dans de nombreux dictionnaires, en parallèle à l'index de formes dans les corpus; les résolutions des variantes graphiques; l'élimination des mots cachés (oubliés) dans le texte du dictionnaire; pour la gestion des exemples et des citations retenus, leur analyse, sélection et classement et les aménagements textuels, etc., qui en découlent; pour le traitement des définitions, la normalisation des définisseurs, l'homogénéisation du métalangage, etc....; pour les corrections et les contrôles divers du texte en cours d'élaboration et, avant son achèvement, les renvois, l'équilibrage des exemples, la normalisation et homogénéisation des informations, etc."

bear in mind that these categories are by no means rigid and separate.

(1) *Machine Readable Lexicons*, which are extracted from a corpus of electronically processed texts and which refer to single authors;

(2) *Machine Readable Dictionaries*, prepared for photocomposition, simply with typesetting codes and without supplementary information;

(3) *Machine Readable Dictionaries plus codes* explicitly classifying linguistic information. In other words, information on the nature and structure of the data is recorded not only implicitly via changes in type-faces (as a side-effect of photocomposition commands) but also via codes explicitly intended for future access and retrieval.

(4) *Machine Dictionaries*, classified, encoded, and with selected information (the Italian Machine Dictionary on tape can be considered a prototype);

(5) *Lexical Data Bases (LDB)*, with structured and formalized information, both at the entry level and particularly at the level of relations between entries, finalized for interactive utilization by many categories of potential users, and associated with specialized software modules for access, interrogation and on-line processing.

The last type of MRD is the specific subject which will be discussed below.

B.2. Sources and types of information

Traditional standard printed dictionaries are certainly one of the most suitable starting points for MDs since they provide a large quantity of important data. Furthermore, they are an invaluable source of information if appropriately structured both from a linguistic and computational point of view. Practically all the new printed dictionaries nowadays are prepared in machine-readable form for simple printing, i.e. for photocomposition. This is a sector in which the publishing "world" is strongly involved.

Thus an increasing number of dictionary projects are now relying on computer techniques, and the *New Oxford English Dictionary* is certainly an example worth mentioning of the considerable effort presently being made in the area of conversion into machine readable form (Hultin and Logan 1984).

However, the conversion of a simple dictionary in machine readable form into a true MD or even more into a complex and structured LDB is really a major undertaking both from the linguistic and the computational point of view. A dictionary on its surface is an elaborately formatted object, and its computerization is a complicated and difficult matter, specially if one wishes to discover the information underlying the surface data.

An important aspect we wish to underline is that the lexicon can be considered as lying at the crossroads between the traditional levels of linguistic analysis: graphic, phonetic, morphological, syntactic and semantic. However, cognitive, pragmatic, psychological and sociological issues are also important with regard to the lexicon, and are often strongly connected.

Information concerning each of these levels can and must be codified (according to different theories) in a lexicon, especially in the case of computerized lexicons. The following are some examples of the range of information which a computerized lexicon can contain:

lemma-word, written according to its usual orthography, plus phonetic codes;

morphosyntactic labels: parts of speech, gender, etc.;

homograph codes, with a distinction between "lexical" and "grammatical" homography;

semantic explanation: a very brief definition (synonyms, paraphrases) to distinguish between homographic lemmas which pertain to the same part of speech;

usage status: archaic, dialectal, popular, literary, etc.;

paradigms: grammatical codes specifying the type of morphological inflection;

forms, written in their usual orthography;

morphosyntactic labels of those forms: number, gender, tense, etc.;

definitions : the definitions of the reference dictionaries;

taxonomy: a numbering system is used for classifying the different meanings of a polysemous lemma;

semantic procedure codes, e.g. metaphor, metonymy, extension, etc.

A particular domain of research can thus be shared by various separate sectors, which can be approached within different perspectives.

One of the main goals to be pursued is to reach a lexical description crossing the specific boundaries of each area, and which is sufficiently general and neutral to allow the different theories to select and pick up only the basic elements from the amount of shared knowledge which are relevant to the specific application or research.

Again this is the direction in which LDBs should move in the near future, and efforts should be made, on a theoretical basis, to establish up to which point a set of structures and formats compatible with different applications, or with different theories, can be envisaged, defined, and implemented.

B.3. Uses and users

Within linguistics, interest over the past few years has gradually moved from syntax to the lexicon, so that an increasing number of well structured and comprehensive lexicons have been developed and created for users. In this respect, a large number of systems, which range from parsing to machine-translation, lexicon-drivers, and large computerized lexicons, are being employed in a wide variety of natural language processing applications.

The aim of LDBs is also to achieve structured and finalized information at many descriptive levels by extending and developing the scope of early machine dictionary projects intended mainly for lemmatizing purposes. They employ increasingly sophisticated computational technology, and the variety of application areas is such that they may be considered as one of the most promising fields of research.

The potential users of LDBs may be classified as follows:

(1) human users (specialists and lexicographers, lexicologists, linguists, or normal users for everyday dictionary look-up);

(2) "procedural users" (i.e. other programs or complex systems for which the LDB is one of the components).

Therefore, a LDB must be as flexible as possible, both from a computational and a linguistic point of view.

LDBs may be used in a wide variety of cases ranging from lemmatization to spelling verification, and from lexicological research and lexicographical practice (e.g. to improve coherence and consistency in dictionary-making) to a number of computational linguistic applications, such as parsers, question-answering systems, man-machine communication, machine (aided) translation, language teaching, etc. They are used within the field of the "language industry" for all applications requiring the use of a lexicon, i.e. in practically all cases, since the problem of lexical access arises whenever we are dealing with words, and wherever the issue of natural language is involved. Consequently LDBs should be considered as the repositories of all the information to which any natural language processing system must have access: morphological, syntactic, semantic, pragmatic, conceptual.

B.4. A prototype of the lexicographical workstation

At Pisa, we are currently working on the design and development of a prototype system of a multifunctional lexical database. We wish to underline the following aspects:

- the source of the data;
- the computational and logical organization of the data in a database structure;
- the results achieved or achievable using this database organization in terms of new information obtained from the original data of a machine-readable dictionary;
- the characteristics, considered from the end-user's perspective, of a lexical database of the type envisaged;
- the link between the lexical and the textual database;
- the relevance of this new concept of integrated linguistic database (lexical plus textual database) when the end-user is the lexicographer himself.

B.4.1. Source of our data

A number of Italian standard printed dictionaries, either transcribed in machine-readable form or already available from the publishers for photocomposition were used as data sources for our lexical data base.

Machine-readable dictionaries are nowadays acknowledged to be invaluable sources of information on the lexical system of a language. However, if left in simple sequential alphabetical form, with only the codes necessary for photocomposition, and in text or string format (as provided by publishing houses) they are of little interest. They must instead undergo a complex process of transformation so as to exploit their enormous information potential. Dictionaries are in fact a relatively structured type of text, and this facilitates their organization into a database. However, there is in particular one type of information which is not explicitly structured for itself, namely definitions and examples or citations, but which is nonetheless of great value when trying to extract new types of information from a machine-readable dictionary.

B.4.2. The database organization

It was decided to re-structure the dictionary data according to the methods of database structuring, and in particular to select the relational data model in which relations are used to describe connections between data items.

Our dictionary database comprises a number of relations. Each relation is a table in which each column corresponds to a different attribute (e.g. the morphological codes), and each row to a distinct entity tuple (e.g. a lemma). The lemma relation was obviously the first relation to be implemented.

The new database organization gives us direct access to all those information categories which in a normal dictionary are already present in coded form. Lemmas can obviously be used for a normal search in the automatic as well as in the printed dictionary, but other new means of consultation are also available. It is thus possible to consult the dictionary not only by lemmas, but also by grammatical category, or by usage code, or by inflectional codes, etc. The creation of inverted files on the fields corresponding to these attributes makes it possible to obtain immediately the entire set of lemmas with a common value for a specific attribute.

Thus a list, for example, of all the adverbs, or all the intransitive verbs, or all the dialectal words recorded in the dictionary, or all the words with a given ending, etc., can be obtained interactively.

Furthermore, the possibility of using natural language definitions to extract semantic information on the lexical entries is particularly interesting. With this in view, a number of linguistically relevant relations has been set up with regard to which it is possible to obtain significant data by running appropriate procedures on the dictionary definitions. It is moreover important that these relationships are defined over the entire lexicon. Information can be obtained, for example, with regard to the relations of synonymy, hyponymy and hyperonymy, antonymy, morphological derivation, co-occurrence, case-frames, etc. These relations can be recognized and stated on the basis of several patterns which occur repeatedly in the definitions and which can often be directly connected with certain types of semantic features or semantic relations or functions.

B.4.3. Relations among words

Once relationships of these types have been defined in the whole lexicon, structures of the lexicon are easily traced along a number of different paths, depending on the relation chosen. For instance, one can ask for all the hyponyms of a word, thus obtaining a semantically coherent cluster with certain properties in common. One can query all the verbs of movement, or all the names of sounds, or all the types of furniture, and so on. Within the lexicon a number of hierarchical structures are thus created and we can work on them in order to formalize, for example, the property of inheritance of relevant features.

It is possible to ask for all the lemmas ending with a certain substring, and connect them with all their definitions. Queries of this type are very useful when analyzing the phenomenon of word-formation. In fact we can obtain very interesting data concerning the interaction between morphology and semantics, and we can evaluate extensively the meaning changes effected by the addition of given suffixes to bases, with the aim of coding regular morphological and semantic behaviour.

Our dictionary is completely cross-indexed according to the abovelisted relations, and it is evident that a computerized dictionary organized according to these semantic relations becomes a first nucleus of a knowledge base, from which a great deal of encyclopaedic information is also retrievable.

One of the effects of this restructuring will also be that of reducing the amount of explicit information on each lexical entry by handling with rules all the information which is predictable on the basis of what is already present (e.g. by inheriting properties from superordinates in the hierarchic arrangement of entries which are available in our dictionary).

B.4.4. The end-user's perspectives

In order to meet the needs of the user, one of our principle aims was to create a system particularly suitable for the "linguist" user, who need not be obliged to learn special computer techniques.

If the term "data model" is used to indicate the entire lexical data universe, i.e. the complete set of relations stored in the system, and if a "schema" is a set of declarations describing the data model, then the set of the relations available for particular users is known as the "data sub-model" and the set of declarations for the data sub-model is called a "subschema".

Tables are temporarily created to give the meaning of, for example, "the superlatives of (Italian) adjectives ending in *-o*, *-a*", or of "archaic adverbs", etc., as required by the user. This is very similar to the database interrogation process, as each query operates on the resident relations to build or to define new relations. A requirement may be a subset formed by only one relation, aimed for example at supplying the meaning of "all the lemmas ending in *-ità* which have archaic graphic variants". Or one's view may extend over more than one relation as with "join" operations, to obtain information of the following type: "all the word-forms of certain irregular verbs of the 3rd conjugation".

Until now we have implemented a query language in an interactive environment. This language - which is very useful at this stage of the project - enables the user to access the primary and secondary keys in transparent mode. The resulting information is essential if the present structure is to be extended into a logically more complex structure.

User/database interaction must be possible on various levels:

(1) *Standard queries*, e.g. all the word-forms which belong to a lemma;

(2) *Complex queries* which were not always envisaged when the database was created, e.g. a phonetician could be interested in selecting all or some of the words in which a voiced consonant is followed by the vowel *a;*

(3) *Complex processing* of the information; e.g. statistical surveys of the distribution of the forms in the various inflexional classes, or of occurrences of homography in the various parts-of-speech;

(4) *Modifications to the schema* with the definition of additional relations.

A more complex non-procedural language is now being designed. With this language the database can also be accessed by a number of users concurrently.

B.4.5. Integration of the LDB with a textual DB

The dictionary can function in a stand-alone mode for human users, or as a module within larger systems, thus providing different possibilities of lexical access for a number of other applications. The application which is of interest here is the connection with an Information Retrieval System for large textual corpora. The two systems are directly compatible and they work as an integrated system to query texts by means of the database dictionary.

The modes of querying texts are defined by the user, and the dictionary is allowed interface to the texts.

Each lemma is expanded into its full inflectional paradigm, and the original input word is replaced in the query by a cluster composed of all the members of the inflectional paradigm. This cluster is produced without any intervention by the user, in a perfectly transparent way, and its members are used by the query program as access keys to the textual corpus.

Obviously every possible query that can be put to the dictionary alone can also be used as a "filter" to make enquiries to textual corpora. We can therefore make very "precise" searches of texts, where the search-key is no longer a simple word-form or a lemma, but for example a word acting as a "semantic marker". The dictionary in fact extends this "semantic marker" to all the lexical items which are coded as its hyponyms. It becomes possible, for instance, to ask the dictionary for all the registered names of colours, and then to go to the texts with this semantically homogeneous subset to find all the contexts where a colour name is used.

The same is obviously possible for synonyms, derivatives, grammatical categories, and so on. It is clear that the dictionary is used in connection with an information retrieval system on texts as a powerful tool for making linguistic generalizations on the lexical level, and for using the detected regularities with a filtering function in the retrieval operation. It can be conceived as an automatic guide to the human user in the investigation, in texts, of the use of particular sets of lexical elements interrelated according to one or the other of certain dimensions of relatedness.

This prototype workstation can therefore be conceived as a central nucleus of basic data (lexical and textual), organized according to suitable structures, plus a set of software mdoules which render these data available at different levels for different users.

B.5. General characteristics of a LDB

The LDBs we have described above should have two important properties, which are those of "multi-functionalism" and "multi-dimensionalism".

By the term "multi-functional" we mean the possibility of using LDBs in various applications, by different categories of users. The availability of a single central repository of "neutral" dictionary data would enable access by many different interfaces, according to the needs of the whole range of possible applications (dictionary server). It should be possible for different external procedures to use different parts of the dictionary content in specific applications. The user (human or procedural) virtually ignores the internal physical structure of the LDB, considering only the data which are useful to his/its own purposes.

The concept of multi-dimensionalism is strongly linked to that of multiple access. By following different paths within the DB it is possible to search different word aspects. When the original data can be viewed within a variety of different perspectives, the important effect of "multiplying" the information offered by the same set of source data is obtained.

Moreover, it is possible to create, as by-products, many virtual secondary sub-lexicons, containing specifically selected parts of the dictionary, such as terminological sub-dictionaries, synonym dictionaries, thesauri, etc. In a well-structured and comprehensive LDB, they only differ in the way the original data are selected, sorted and interrelated.

We can imagine a multi-access dictionary with all the properties described so far, and possibly more, as recorded on a diskette or a compact disc and accessible to the ordinary user in his own home. We are here envisaging the "dictionary of the future" or "tele-dictionary" for general consultation, which will become the "dictionary of the present" in a few years. Modern technology will certainly be able to produce the kinds of facilities we have described so far.

C. Conclusions

A lexical data base also offers new possibilities to publishing houses, since they will be able to produce from a LDB a variety of different lexicographical products in the most favourable circumstances.

> "From the very numbers of dictionaries of varying shapes and sizes that follow in the wake of the major ones, it is clear that different levels of detail are appropriate to different people and to different kinds of use (...) The amount of information that these editions contain is clearly chosen for largely economic reasons and from the point of view of any dictionary user" (Kay 1983:163).

If appropriately coded, however, the information structured in a LDB could allow editors to produce, (semi-)automatically, different kinds of dictionaries (printed or sold on magnetic support).

In this framework, the relevance of a standard taxonomy of lexicographical data must be stressed. This would facilitate not only the non-ambiguous description of the content of dictionaries, but also the totally or partially automatic exchange of lexicographical data between different dictionaries.

The creation of a lexical DB, in our opinion, should directly involve dictionary publishers. These are still, nowadays, the "owners" of the major repositories of lexicographical information.

Because their collection is very expensive and their economic value very high, there is certainly good reason to protect the data, and to exploit its potential value for new products required by the so-called information society: spelling checkers, translation aids, text-editing aids, automatic indexing, etc.

These utilizations require that the present lexicographical collections be transformed into data bases structured according to models which take into account the linguistic nature of the lexicographical information independently of its various possible applications, according to the principles indicated above at B.

We can consider these data bases as intermediate products which are necessary - among other reasons - to optimize the production costs of printed dictionaries; to integrate the lexicographical data within the information systems previously mentioned; and to create new lexicographical products, as for example mono- and plurilingual dictionaries on CD-ROM.