COMPUTATIONAL LEXICOGRAPHY AND LEXICOLOGY

ANTONIO ZAMPOLLI
Istituto di Linguistica Computazionale
del CNR, Pisa

NICOLETTA CALZOLARI
University of Pisa
Pisa

Other articles which appeared in the AILA Bulletin (Rondeau in 1979, N.2) were concerned with the application of computers in terminology. This article is a brief overview of the uses of data processing in lexicology and lexicography. These uses may be classified in different ways. The following scheme will be followed:

- Text processing for the production of indexes, concordances, quantitative data, etc.;
- The use of the computer in dictionary making;
- Machine dictionaries;
- Lexical/linguistical data bases.

1. INDEXES, CONCORDANCES, QUANTITATIVE LINGUISTIC DATA

The first applications of the computer in the production of indexes and concordances was by R. Busa S.J. and began in 1949 (Busa, 1951). The diffusion of these text processing methods increased rapidly between the fifties and the sixties; nowadays the use of the computer in this sector is accepted as routine.
For a history of the development of these applications see Zampolli (1973a), Les machines dans la linguistique (1968), various numbers of Cahiers de Lexicologie, of the Bulletin of the Association for Literary and Linguistic Computing, and of Computers and the Humanities, and in particular, in this journal, the series of four articles by D. Burton on this subject.
Although the procedures implemented in the different specialized centres differ from each other, they have a common logical scheme.

1.1 Acquisition of Texts in Machine-Readable Form (MRF)

There are several alternatives when converting textual material into machine-readable form: key-punch; key to paper tape; on-line type-writer; key to disk; selectric type-writer; key-boarding on visual display terminals; etc. All these require that someone types the text on a key-board, and until recently this manual operation was a major obstacle, owing to its cost.
Nowadays, the need to manually type a text is limited 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 by publishing-houses in MRF, so that they may be printed using photocomposition techniques;
- 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 resolved:
- organizational: how to exchange information and data between different textual archives;
- juridical: 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 may guarantee that the representation of a text in TRF contains all the minimal information requested by linguistic, lexicographical, philological processing.

1.2 Indexes and concordances

Each specialized centre has its own procedures to produce the results usually requested by linguistic and literary text-processing: 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 allow their users to specify a number of "processing options". As an example, let us discuss the options for concordance preparation.

Nature and Organization of Entries

Units of different linguistic levels may function as concordance entries: graphic forms, lexical forms, lemmas and forms (1), syntagms, codes representing syntactic, semantic and thematic structures. In some cases, the internal organization of the entries is similar to the microstructure of a dictionary entry.

Selectivity versus Completeness of Documentation

The decision to print the complete concordance of a text, or to select particular items of interest concerns:
- The entries: for example, some units may be excluded by a stop list, by frequency criteria, or by manual intervention during the lemmatizing stage.
- The occurrences: for example, only the global frequency or only the index locorum is given for extremely frequent words. Mixed documentation is often produced, e.g. selected contexts and/or selected references, followed by the indication of frequencies for the remaining occurrences.

Construction of Contexts

The user can specify the maximum length of the context, which can be constructed using different algorithmic rules:
- The word for which the context is being constructed (CW: contextualized word) is always placed at the center of the context.
- The text is segmented in a sequence of contextual units, identified by delimiting factors chosen by the user: changes of references (new line, verse, paragraph, etc.) punctuation marks (colon, exclamation mark, question mark), marks which have been manually inserted during the pre-editing stage. Each contextual unit functions as a context for all the CWs it contains. The so-called contrastive concordances constitute a very interesting case. If different versions of a text are to be compared (translations or different editions), any word of any version will receive as its context not only its contextual units but also the corresponding contextual units of the other versions.
- The context depends on the grammatical nature of the word. For example, a preposition may be given its prepositional group as context. Obviously, this algorithm presupposes previous grammatical tagging of the text.
- The computer gives an over-abundant context, which is then manually reduced by the exclusion of interactively selected syntagms.
- The position of the CW in its context is "adjusted" according to the

presence on both sides of predefined, hierarchically ordered, textual phenomena (weak punctuation marks: comma, brackets, semicolon; strong punctuation: colon, question mark, exclamation mark; reference level: verse, chapter, etc.). This system is by large the most frequently chosen, as it seems to provide the best compromise between two opposing requirements: space saving and "significance".

Ordering of Elements

Ordering may affect both the sequence of the entries (e.g. the forms may be listed under their lemma in alphabetical or in a particular morphological order) and the sequence of the contexts: text order, chronological order, alphabetical order of the word preceding and/or following the CW. This last arrangement, in fact, seems to be preferred by classical philologists, in particular when studying formulas or the influence of an author on successive authors.

1.3 Lemmatization

The dilemma "to lemmatize or not to lemmatize" the concordances and the indexes dates back to the very beginning of the use of computers in data collection 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 versus philological or literary applications.

Two of the main obstacles to lemmatization are:

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

Although lemmatization is in most cases still completely manual (2), various semi-automatic systems have been experimented. The most common of these consists in the use of the so-called machine-dictionary (MD).

Machine-dictionaries are discussed in detail in section 3. Here we discuss only their role in semiautomatic lemmatization. From this point of view, MDs may be subdivided in two main groups:

1.3.1 MDs of word-forms

An alphabetically ordered list of word-forms is recorded on a magnetic support. Each form is accompanied by an "analysis", i.e. a set of linguistic information which includes the lemma (or the lemmas, if the form is homographic in the lexical system), and the grammatical categories of the form and the lemma(s).

The look-up algorithm receives as input the DM and the concordances of the text to be lemmatized, both in alphabetical order of forms.

The Algorithm:

- searches the word-form in the DM, i.e. identifies the form in the DM which is "graphically identical" to the given word-form.
- associates the "analysis" suggested by the DM to each word-form "found" in the DM;
- prints the concordances with each form followed by its associated "analysis".

When examining the concordances, the lemmatizer:

- checks (if necessary) that the lemma associated to unambiguous forms is correct and, in particular, verifies that these forms are truly unambiguous in the text being analysed;

- checks the lemmas proposed for homographical forms and associates the appropriate lemma to each single occurrence of these forms;
- assigns a lemma (or lemmas for homographic forms) to forms not found in the DM. These "new" forms are then considered for the 'updating' of the DM.

1.3.2 MDs of stems

From the very beginning, considerable efforts have been made by machine translation projects to reduce the size of their MDs in order to optimize look-up times and to save storage space. Techniques were already available in the early '60s based on the decomposition of the word-form into substrings of different types (prefixes, stems, suffixes and endings) which were then searched in different sections of the MD.

The MD usually includes a stem dictionary, a table of endings (subdivided into different inflectional paradigms), a table of possible enclitics, and eventually affix (suffixes, prefixes) and enclitic tables.

All these elements are given morphological codes specifying their relative compatibility: e.g. which inflectional paradigm is compatible with which given stem.

An algorithm tries all the possible decompositions of a form according to the following sequence: (PREFIX)* "ROOT" (INFIX) (SUFFIXES) * (ENDINGS) (ENCLITICS)*, and accepts the decomposition (or decompositions for homographic forms) into elements with compatible morphological codes. The lexicographer can interactively check the "analysis" proposed by the analyzer. New forms (i.e. forms for which the analyzer has not found any corresponding analysis) and homographic forms are signalled to the lexicographer, for manual lemmatization.

1.3.3 Homograph disambiguation

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

a) Local disambiguation

The algorithm usually attempts to resolve 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. 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 data, obtained by examining previously lemmatized texts, to provide the frequency distribution of words and grammatical categories in the immediate context of 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.

b) Disambiguation via syntactic-semantic parsers

Automatic parsing of natural language texts has always been the main interest of computational linguists. On the contrary, up until now lexicographers have never shown any particular interest in parsing systems. In computational 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 stress the possible use

of a parser as a homograph disambiguator. If a parser succeeds (i.e. if the sentence is not in itself ambiguous and its overall syntactic structure is recognized), it will obviously accept only one of the possible "analyses" of a homographic form.
The problem is that the presently existing parsers are unable - to our knowledge - to exhaustively treat the large variety and quantity of phenomena present in the types of corpora which usually constitute the basic documentation of the lexicographer.

1.4 Quantitative data

Textual archives must also be regarded as a source for statistical research. Although a certain amount of studies have already been carried out at the phonological level, and others are now beginning at the syntactic level, the majority of data are available at the lexical level. In fact, the statistical unit "word" in a certain sense (a string of letters between two blanks) is almost immediately given in a text in machine-readable form, whereas the representation of syntactic structures must be inserted using an extremely time-consuming procedure.
As is known, the diffusion of text processing has partly falsified the models and 'laws' formerly established (Zampolli, 1975). Within both the structuralist and transformalist frameworks, some linguists have tried to draw up a research program leading to the construction of a new explicative model of frequency stability and variation in the texts. Both approaches agree with the necessity of building such methods inductively by identifying, through frequency distribution analyses in sufficiently extended and stratified corpora, the relationship between subsets of texts (literary genres, different authors, diachronical distribution, etc.) and the stability or instability of the frequencies of the linguistic units (Dolezel, 1969).
However, the major obstacle to the development of such research plans is the scarcity of texts analyzed at different linguistic levels. We hope that the cooperative effort required for the construction of a linguistic data base (see section 4) will contribute towards extending the subset of analyzed texts. Until now, a major source of data is constituted by frequency dictionaries, whose elaboration is known to be an enterprise demanding the involvement of a computer (Bortolini et alii, 1981; Juilland et alii, 1964, 1973).

2. DICTIONARY MAKING

Dictionary creation using the computer is usually divided in three stages:
- Acquisition of data
- Preparation and editing of dictionary entries
- Printing/Diffusion of data

2.1 Data acquisition

The editing of a dictionary, and in particular of large historical dictionaries, traditionally first requires the creation of a collection of documentary data. These can be of different types (monographs, grammatical information, etc.), but can be divided into two main groups:
- Collection of existing dictionaries in the same language
  If these dictionaries are recorded in machine readable form they can be consulted by using multiple-access tecniques as described later in 3.7.
- Collection of textual citations excerpted from a representative text corpus
  Traditionally, human excerptors were set to read through the texts,

selecting what they apprehended as unusual or particularly informative examples and copying each textual citation onto a paper citation slip.
For the bulk of the corpus 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, p.34)
Even so, those collections may contain several hundred thousands of quotation-slips (for ex., the OED in its present form contains about 1,820,000 quotations, selected from nearly six million quotation slips).
Computers are nowadays increasingly successfully used for collecting textual data. Computer generated word indexes, concordances, and quotation slips are now widely used by lexicographers. They are produced using the same methodologies described in section 1. The main historical/academic dictionaries which collect and process large corpora of texts in machine-readable form are: the Trésor de la Langue Française (Nancy), the Dictionary of Old English (Toronto), the Historical Dictionary of the Hebrew Language (Jerusalem), the Tesoro della Lingua Italiana delle Origini (Florence), the Dictionary of the Old Spanish Language (Madison, Wisconsin), the Dictionary of the Older Scottish Tongue (Edinburgh).
A number of "commercial" lexicographical projects have also started to use computers to collect or to store quotations, e.g. the New York Times Everyday Dictionary, Merriam-Webster, Houghton-Mufflin, etc.

2.2 Preparation and editing of Dictionary Entries

The lexicographer wanting to compile the article for a dictionary entry has to read, in principle, all the contexts of its word-forms which appear in the archive of concordances or slips collected manually or with the aid of a computer (see Howlett, 1983, p.156 and Aitken, 1983, p.145).
The editorial analysis consists of a "rapid shuffling and re-shuffling of examples" (Aitken, 1983, p.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 linguistics know-how and methodology.

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 examples 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 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 ways of establishing habitual collocations statistically. One of these is to "start from the observed frequency between node and collocate (3), 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, p.85) (4).
It is obviously important to be able to systematically explore the archive so that particular constructions may be emphasized.
"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, p. 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 dispose of a parser capable of automatically producing the description of the syntactic-semantic structure of the text. In this case it would be possible to identify quotations which exemplify particular syntactic and semantic patterns. Unfortunately, the existing parsers are not yet capable of exhaustively treating the variety and quantity of phenomena present in the corpora which constitutes the typical basic documentation of the lexicographer. However something new is now moving in this sector. Examples are the DEREDEC system (Montréal) (Plante, 1983) and the researches of the Institute for Computational Linguistics of Pisa.

2.2.2 Towards a lexicographical workstation

A major challenge is to develop a "lexicographic workstation", by which the lexicographer editing the dictionary entries can interact with a lexical data base, conceived as a set of different knowledge sources (text corpus, old archives of quotations, pre-existing dictionaries, bibliographical references, etc.) available on-line and accessible by means of appropriately designed software tools.
According to some scholars, the publication of a dictionary in 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 perspective 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 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 if 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 essentially aim at facilitating access to textual samples and dictionary entries stored in a data base, and the storing of preliminary versions of lexical entries for further processing (Lentz, 1981; Zampolli, 1983).
A first component ensures quick access to the data of the data base, thus enabling the lexicographer to use the corpora interactively via terminal. For example, the lexicographer can search specific word forms, word forms matching (beginning, containing or ending with) a specified string of

graphemes, cooccurrences of word forms and/or grapheme strings in a given span of text. If the texts are already lemmatized, the lexicographer may operate on both lemmas and/or word forms. As output, the component provides the lexicographer with information on the relative and absolute frequencies of distribution in different sections of the corpus. The lexicographer may then request the contexts of selected occurrences to be displayed on the video screen or to be output in printed form. The context, which is algorithmically "cut out" by the computer, may be interactively modified by the addition or exclusion of selected syntagms (Picchi, 1983).

A second component enables the lexicographer to consult existing dictionaries in the data base. The lexicographer will obviously benefit from the multiple dictionary access techniques described in 3.7, which enable him to search different information fields in the entries of both existing dictionaries and the dictionary which is being constructed. Summarising, it could be said that the lexicographer can construct the "microstructure" of a lexical entry by interactively examining lexical entries in existing dictionaries and comparing their descriptions with contexts in the corpus.

Specific functions permit the insertion and cyclic reordering of selected contexts under 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 used immediately by multiple access procedures for consultation and comparison within the new dictionary, in order to help homogeneity and coherence.

We feel that a greater cooperative effort between lexicographers and computational linguists is needed if a complete experimental procedure is to be constructed. In particular, the operations which the lexicographer performs when preparing a dictionary entry must be analysed 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). We could thus have to continue to print concordances or, even better, citation slips which can be used by the editor in the traditional manner. In order to avoid retyping the selected citations, all the citations stored in the computer's memory can 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 on this problem was probably held during a round-table between lexicographers from more that ten countries held in Pisa, 1972. The conclusions were very uncertain. The situation today is probably somewhat different due to the evolution of data base methodologies and of workstation technology (5), which seem to offer today the chance of "simulating" on the video the games of "solitaire" which the lexicographer has always played, ordering and reordering the traditional slips.

2.3 Printing

Photocomposition tecniques are now commonly used by most publishing houses for the third stage, i.e. the printing of dictionaries.

In order to be photocomposed, the text of a dictionary must be registered in machine-readable form.

A variety of editorial controls and readjustments are thus possible before the final printing.
Quemada (1983, p. 27) provides some examples:
"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 text du dictionnaire; pour la gestion des exemples et des citations retenus, leur analyse, séléction et classement et les aménagements textuels, etc. qui en découlent; pour le traitement des définitions, la normalisation des définisseurs, l'homogéneisation 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."
Other examples are given by Knowles (1983, p. 136-87), by Howlett (1983, p. 157) Petersen (1983), Pfister (1983).

3. MACHINE DICTIONARIES

Over the past few years there has been a growing interest towards the interplay between "lexical data" and the "computational approach". Many research groups are acknowledged to frequently refer to that relationship, and there is a widely-shared awareness among scholars that the experience of Machine Readable Dictionaries (MRD) is fundamental not only to lexicological research or lexicographical practice, but to a large number of applications, ranging from Natural Language Processing systems, to dictionary printing, spelling verification, knowledge base systems, etc.
Many researchers are actually engaged in the creation and/or utilization of different types of MRD, according to their different purposes. Almost every dictionary has its own approach, methodology, data structure, computational formats, etc.
We shall now examine and make some considerations about MRDs from the point of view of sources, aims and purposes, access, potential users, evolution, lexicological theory, possible applications, and so on.

3.1 Typology

The expression "Machine Readable Dictionary" - widely used nowadays in Computational Linguistics - is employed in various frameworks and assumes different meanings in correspondence with different objects. It is applied either to the different approaches towards a common underlying generic notion or to actually distinct concepts.
In order to clarify the terminology, we must therefore begin by drawing a tentative typology of MRDs, and by classifying the main types usually intended by this term. It must however be kept in mind that the following categories are by no means rigid and definite.

a) Transcribed Printed Dictionary

The text of a printed dictionary is recorded in machine-readable form without any supplementary information, apart from (eventual) photocomposition codes. The TDs will probably increase in the near future because of the diffusion of photocomposition.

b) Classified Transcribed Dictionary

Codes explicitly classifying different linguistic information are inserted into the machine-readable text of a dictionary. 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. Unfortunately, there are only a very few cases of this so far. The Longman's Dictionary of Contemporary English is probably the best known example.

c) Sometimes a "Transcribed Printed Dictionary" is progressively transformed into a "Classified Transcribed Dictionary". J. Olney's work on Webster's Seventh New Collegiate Dictionary and on the New Merriam-Webster Pocket Dictionary is well known.

Optical character readers may be used to put printed dictionaries into machine readable form, for example with the aim of publishing new, and updated, editions. A recent interesting project intends to create a lexical data base recording the (out-of-print) French-Danish Blinkberg and Høybye dictionary. Experiments were conducted in order to assess whether the structure of the printed dictionary (special printed symbols, defining formulae, etc.) may be used to create, fully or semi-automatically, an explicit linguistic classification of the dictionary information.

d) The Lemmatizing Dictionary:

The lemmas which appear progressively in a corpus of electronically processed texts are recorded in MRF. The lemmas are supplied with information which will be useful for the lemmatization of other texts: parts of speech, inflectional paradigms, homographical relationships, brief explanations of the meanings of lexical homographs, etc. Several centres for literary, philological and lexicographical text-processing have created dictionaries of this type.

e) The Machine Dictionary

The lemmas of a printed dictionary, or of a set of dictionaries, are recorded in MRF (mainly for the text-processing lemmatization stage) together with selected information, extracted from the printed dictionary but directly classified and encoded in a format suitable for EDP.

f) The Parsing Machine Dictionary

Until now, it was all too common for each separate natural language processing project to construct its own dictionary.

These dictionaries are almost never related to any printed dictionary and the number of lexical items is very limited, usually restricted to a technical domain. The dictionary is focused on syntactic and semantic (if not pragmatic) information, highly formalized and strongly finalized to the parsers (or the generation programs) that use it.

g) Lexical Data Base (LxDB)

The lexical data are organized in the form of a multifunctional database, which offers multiple access points to a variety of "users", both human and computer programs, each of which will have a particular "view" on the data (Zimmermann, 1983).

This type, which will be the specific subject of the follow-up of this paragraph, is a task for the near future.

3.2 Scopes

There is no doubt that almost whatever task is to be undertaken in the field of Computational Linguistics, in one way or another, one has always to cope with "words". It is thus necessary to be able to access to a lexicon in which different types of information on words are stored.

The range of scope of MRDs can be very wide indeed, going from the more applicative to the theoretical area.

On such a complex and differently organized map of applications, the

researches and the work which still have to be done in this field are considerable.

3.3 Sources

One of the most suitable starting points for a MRD is the traditional standard printed dictionary. Such dictionaries provide in fact a large amount of data which can become an invaluable source of information when organized according to proper linguistic and computational structures.
Many research groups are now well aware of this fact, see for example the groups in Bonn (Brustkern et alii, 1983), and Liège (Michiels et alii, 1981).
Nowadays almost all new printed dictionaries are prepared in machine-readable form if only for the printing, i.e. for photocomposition. This fact involves the publishing "world" in the more general issue of MRDs.
We can observe that there are an increasing number of dictionary projects relying on computer techniques, and it is also worth noting that a large effort is currently being made to convert the Oxford English Dictionary into machine readable form and into a lexical data base (LxDB).
The conversion of a simple machine readable dictionary into a complex and structured LxDB is really a major undertaking both from the linguistic and the computational point of view. A dictionary in its surface is an elaborately formatted object, and its computerization is a complicated and difficult matter, even more so if one wants to go beyond the surface data in order to discover the underlying information.

3.4 Types of information

An important point which must be stressed is that the lexicon can be considered somewhere at the crossroads between the traditional levels of linguistic analysis: graphical, phonetical, morphological, syntactic, semantic, but also cognitive, pragmatic, psychological, sociological issues are not only of relevance when speaking about the lexicon, but are often strongly interrelated.
Information on each of these levels can and must be codified (according to different theories) in a lexicon, especially in the case of a computerized lexicon.
Thus we are dealing with a domain of research shared by various distinct sectors, which can be approached according to different perspectives.
In this framework the following fundamental question arises: is it possible to reach a lexical description - crossing the specific boundaries of each area - which is sufficiently general and neutral enough to allow the different theories to select and pick up from the stock of shared knowledge only those elements relevant to the specific application or research.
We think this is the direction towards which LxDBs should move in the near future, and efforts should be made, on a theoretical basis, to establish up to which point a stock of structures and formats compatible with different applications or also with different languages can be envisaged, defined, and implemented.

3.5 Aids in word processing

One of the major applications of computerized dictionaries in the commercial environment is the checking of texts in order to detect spelling errors. Some systems are also able to suggest the corrections to be made, by displaying in a window on the screen some possible alternatives.
Although a simple application, and certainly not the most interesting from a

theoretical point of view, spelling checking requires research to obtain sophisticated techniques for word compression (organization in a tree structure, hashing functions, etc.), so that continually growing dictionaries can still be stored on small personal computers. It also uses statistics regarding the most common typos, determined for example by keyboard layout, or by pronunciation.
Satisfactory results have been obtained in this area, and some systems now available for certain languages are invaluable tools in detecting typing errors, and extremely helpful and time-saving when combined with one of the many word processors which can be found on the market.
More ambitious projects are emerging from this first stage: their aim is to provide more complex aids for writing and composition, by means of several functions accomplished by different modules. We can mention the Epistle project, developed at IBM Yorktown, and the Writer's Workbench, designed by Bell Communications Research.
Other components envisaged for systems of this kind are: a syntactic parser to detect errors in grammar, a synonym dictionary offering the choice of more suitable or alternative ways to express a concept, a general dictionary in computerized form to control the exact meaning of a word, a stylistic analyzer to eliminate simple errors in style, a statistic component to calculate the number of rather vague words, or of difficult or rare ones, etc.

3.6 Lemmatization

For many years MRDs have been used in the so-called (semi)automatic lemmatization phase of text processing. We have already described these applications in 1.3.

3.7 The Access issue

What is of crucial importance in the design of a LxDB is the "word access issue". A LxDB can obviously overcome all the limitations imposed by a sequential alphabetical organization of the lexical entries. What appears to be less obvious is instead how new and linguistically relevant search methods can be developed. In addition to the possibility of access to the dictionary by pronunciation, or the use of grammatical classes and categories as access points (e.g. all the coordinating conjunctions, or all the transitive verbs which can govern a that-clause), far greater advantages and significant gains can come from using definitional data to construct 'semantic' equivalence classes. It is possible to construct virtual IS-A hierarchies, or to implement other taxonomies based for example on the "Part-Whole", or "Set-of" Relations, which help us to discover in the lexicon new dimensions of "relatedness" between words. For each different Relation, a different data structure and a different index to the dictionary are required. It will be possible to see the original data as if through different semantic relational grids, which structure and re-structure the data for the dictionary user according to the different semantic relations requested at the moment of the query to the LxDB.
One can query a LxDB, ordered by ideas, not only to find a word, but also to a certain extent a particular meaning, and this is possible efficiently and interactively by exploring all possible paths and relating this general meaning to all the correlates nearby, according to the different kinds of selected Relations. The dictionary provides us with a range of perspectives. One can start with the notion of 'sound , and look for all the instruments producing sounds, or limit the search to animal sounds, or

search for verbs meaning "to produce a sound", or see which adjectives can be associated to names of sounds, and so on.
While working with different procedures on natural language definitions of standard dictionaries, in order to obtain a tool of the type described above, we are moving towards the transformation of the metalanguage of definitions into a first nucleus of a large Knowledge Base, where techniques developed in Artificial Intelligence (as e.g. semantic network structures) or in Data Base Management Systems must be taken into serious consideration. In designing a logical structure for the lexical data, with many access points, (i.e. with many kinds of correlations), we are also tending towards a simulation of memory organization. Thus, the theories, methods, and techniques involved in psycholinguistics are of great importance.

3.8 Bilingual or Multilingual Computerized Dictionaries.

Little research seems to have been carried out so far in the field of bilingual MRDs which goes beyond the storing of bilingual dictionaries on tapes ready for photocomposition, or the compilation of lists of corresponding words for Machine Translation.
This could be a very promising field for exploration in the next years, both for theoretical and for applicative reasons. It is not difficult to imagine application areas which can demand new types of access to bilingual data. Moreover, research in the logical design of a bilingual LxDB stimulates work in related fields, and interesting issues are raised in epistemology, knowledge representation, sociolinguistics, contrastive linguistics, etc.
In this respect, another issue to be considered is the following: are printed bilingual dictionaries a good or useful source of information for the implementation of bilingual LDBs, as acknowledged for monolingual printed dictionaries, or is there a need for other sources of information?, is their structure computationally exploitable for the generation of correct and valuable links not between simple lexical words, but between lexical concepts? This is important for example in order to avoid what C. Miller (in this Bulletin) calls "kidrule" errors in a bilingual context, i.e. semantically anomalous sentences. Another question could be the following: can bilingual dictionaries be used to connect two monolingual LxDBs?

3.9 The future: the commercial "teledictionary".

We can conclude by imagining a multi-access dictionary with all the properties so far described, and perhaps others, recorded on a diskette and accessible to the normal user in his own home. We are here envisaging the "dictionary of the future" or "teledictionary" for general consultation, which will become the "dictionary of the present" within a few years. Modern technology can certainly support this prediction.
Within this framework, it is a task of the computational linguist to try to invent new search strategies for the look-up process. We should be able to guess the possible needs of both the specialist and the normal user, and their expectations when looking up a dictionary. In this way, this relatively new discipline, namely "computational lexicology", can also move to explore an old familiar object, the dictionary, so that completely new requirements may be satisfied.

4. LINGUISTIC DATA BASES

Bernard Quemada has called "néo-dictionnaires" the possible developments described at the end of the previous paragraph.

The "terminological data bases", which have been available for several years in both public and private organizations, may suggest similar ways to construct and handle very large lexical/linguistic data bases (LDB).
We use LDBs to define a set of linguistic data of different types (not only texts and dictionaries, but also, for example, the result of socio-linguistic studies, linguistic models, bibliographical data, etc.) finalized for the interactive utilization by multiple categories of potential users, stored, structured and linked for this purpose, and associated to specialized software modules for access, interrogation and on-line processing.
The design of LDBs of this type is still under development, but several projects in different countries are converging in this direction.
A LDB can fulfill a variety of tasks and objectives. First of all, it is able to increase the potential use of continually growing sets of linguistic data in machine readable form.
Bearing in mind the widespread penetration of computing technologies in our "information society" (terminals, personal computers, telematic networks), which leads to a greater familiarity with computational tools, we should like to suggest working with the following objectives:

a) To "release" the researcher from the static ordering of indexes and concordances (and, even more so, of the traditional dictionaries), profiting from data base methodologies and distributed access technologies, in favour of multiple dynamic interactive access to the data.
b) To "enrich" the archives, not only with regard to the quality of the primary data (texts, dictionaries, etc.), but also and above all by the explicit representation of the units and relations of the different linguistic levels.

With this aim in mind, we must attempt to:
- Reduce the cost of linguistic analyses of texts, using automatic or computer-assisted procedures (dictionary look-up, parsers, etc.);
- Propagate minimal norms and computational tools which make it possible to compare analyses performed by different researchers;
- Encourage collaborations not only for the acquisition of data in MRF but also for analyses. Whoever uses a test from an archive must "deposit" the analyses he has made, even if only partial, so that they can be exploited by other users.

c) To "assist" the researcher in identifying and processing pertinent data.
It seems necessary to:
- identify the "profiles" of potential user categories, describing the operations and the procedures which constitute their research activities;
- create interrogation, processing and access functions to assist the performance of these operations;
- make available in the same "interactive environment" possible different "knowledge sources": dictionaries, bibliographies, rules, parsers, statistical analyses, etc.

In the preceding paragraph we have described the different possible uses of a machine-dictionary organized on a data base structure.
A LDB also offers new possibilities to publishing houses. In fact, they may well find it of great advantage, on a practical and economic level, to have a database of this type, plus advanced software, in order to be able to obtain a number of commercial versions and variations, revisions and updates of a dictionary.
Under the best conditions, it will be possible to produce from a LDB a variety of different lexicographical products.

"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, 83, p. 163).
If appropriately coded, the information structured in a LDB may allow the editors to produce, (semi)automatically, different kinds of dictionaries (printed or sold on magnetic support).

NOTES

(1) We shall use the following terminology. A lexical unit is the basic unit of the lexical system of a language, often coinciding with a simple "lexeme" (see Cowie, 1983, pp.100-101). A lexical unit usually corresponds to a dictionary entry, or article. A lexical unit is represented by a set of related (word)-forms (inflectional or graphical variants), one of which (e.g., the present infinitive ,for the verbs) is usually chosen as the headword, or lemma, for the corresponding dictionary entry.
If the same written word-form (i.e. the same string of letters) represents word-forms of different lexical units, or different inflections of the same lexical unit, we call this written form a homograph. A written form may appear in different places of a text. Each appearance of a written word-form is an occurrence (of this word-form), and its frequency is the number of its occurrences.

(2) In our experience, lemmatization performed for lexicographic projects consists of the basic operations specified below (usually performed on the concordances).

1. Lemmatization of unambiguous forms: the lemma is entered at the side of the first occurrence of the form. The program then automatically assigns this lemma to all occurrences of the same form.
2. Lemmatization of homographic forms: the different lemmas which can be assigned to a homographic form in the text are written at the side of the form. Each of these lemmas is given a number. The number of the lemma which must be associated to that particular occurrence is entered at the side of its context.
3. Composition and Decomposition of Forms: the procedure must permit:
a) the reconstruction of lexical forms which are divided in current spelling e.g. locutions (head over heels), compound noun forms (station wagon), compound verb forms (I have seen);
b) the decomposition of orthographically united forms e.g. enclitic and verb forms, (dirtelo: dire+ti+lo), articles and prepositions (allo = a + lo), etc.

4. Assignment of words into different processing categories: frequently, at least for some applications, there is a strongly felt desire to reduce the mass of data by excluding part of the lexical material from the processing work which may follow lemmatization. Normal procedures usually permit the material to be assigned to different types of treatment. For some lemmas or forms in concordances it is possible to print: a) only the frequency of occurrence; b) only the index locorum; c) only some contexts, chosen by the lemmatizer; d) all the contexts.

(3) Collocations may be defined as "the co-occurrency of two lexical items within a specific context (...). A 'node' is the lexical item whose collocational pattern we are looking at. A 'collocate' is any lexical item which co-occurs with the node within the specific co-text. A 'span' is the co-text within which the collocates are said to occur" (Martin et alii, 1983, p.84)

(4) "Collocational studies, (...) should go far in exploring the techniques of meaning discrimination as established by classical lexicography. If one could draw a distinction between the several readings of the 5,000 most frequent words (many of which are polysemous) on the basis of their different collocational behaviour, ... not only the problem of homography could be solved, but the meaning of many polysemous words in a particular co-text could be established as well" (Martin et alii, 1983, p.87), thus offering new perspectives for quotation selection.

(5) Size of the screen, programmability of the characters, subdivision into "windows"; possibility of displaying a citation-slip like file from which selected citations can be extracted, added to, reduced or overlain by moving a "mouse" or simply by touching the screen; possibility of writing directly on the screen; possibility of recalling and restoring at any moment previous states of organization of the material referring to an entry, by a trial and error process; immediate access and display, through temporarily opened windows, of entries in other dictionaries or in the same dictionary (eventually edited by other members of the editorial staff), of bibliographies, of monographs, etc.; possibility of linking to video discs, voice synthetizers, etc., in the near future.

REFERENCES

AITKEN, A.J., "DOST and the Computer: A Hopeless Case?", ZAMPOLLI, CAPPELLI (eds.) 1983, pp. 51-63.

AMSLER, R.A., Challenge Paper, Stanford, 1983.

AMSLER, R.A., "Computational Lexicology: A Research Program", Proceedings of the American Federation for Information Processing Societies, AFIPS, 1982, pp. 657-663.

BAHR, J., "Reflections on the Project of a Lexical Data Bank", Cahiers de Lexicologie, 1978, 32(1), pp.55-64.

BORTOLINI, U., TAGLIAVINI, C., ZAMPOLLI, A., Lessico di frequenza della lingua italiana contemporanea, IBM Italia, 1971.

BRATLEY, P., "Computers and Lexicography: Advances and Trends", ZAMPOLLI, CAPPELLI (eds.) 1983, pp. 83-95.

BRUSTKERN, J., HESS, K.D., "Machine Readable German Dictionaries - From a Comparative Study to an Integration", Linguistica Computazionale, III (1983) Supplement, pp. 77-93.

BURTON, D.M., "Automated concordances and word indexes: the fifties", Computers and the Humanities, 15 (1981)1, pp. 1-14.

BURTON, D.M., "Automated concordances and word indexes: the early sixties and the early centers", Computers and the Humanities, 15 (1981)2, pp. 83-100.

BURTON, D.M., "Automated concordances and word indexes: the process, the programs and the products", Computers and the Humanities, 15(1981)3, pp. 139-154.

BURTON, D.M., "Automated concordances and word indexes: machine decisions and editorial revisions", Computers and the Humanities, 16(1982)4, pp. 195-218.

BUSA, R., Sancti Thomae Aquinatis Hymnorum Ritualium. Varia Specimina Concordantiarum. Primo saggio di indici di parole automaticamente composti e stampati da macchine IBM a schede perforate, Milano, 1951.

CALZOLARI, N., "On the Treatment of Derivatives in a Lexical Database", Linguistica Computazionale, III (1983a) Supplement, pp.103-113.

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

CALZOLARI, N., PECCHIA, L., ZAMPOLLI, A., "Working on the Italian Machine Dictionary: a Semantic Approach", ZAMPOLLI, CALZOLARI (eds.), 1980, pp.50-69.

CAPPELLI, A., FERRARI, G., MORETTI, L., PRODANOF, I., STOCK, O., Parsing an Italian text with an ATN parser, NT-NLU 1; ILC-CNR, Pisa, 1978

CATARSI, N., RATTI, D., SABA, A., SASSI, M. (eds.), Ordenadores y Lengua Española, (1981), Pisa, 1982.

CIGNONI, L., PETERS, C. (eds.), "Computers in Literary and Linguistic Research, Proceedings of the VII International Symposium of the ALLC", Linguistica Computazionale, III(1983) Supplement.

CIGNONI, L., PETERS, C., ROSSI, S., (eds.) European Science Foundation: Survey of Lexicographic Projects, Pisa, Istituto di Linguistica Computazionale, 1983.

COWIE, A.P., "On specifying grammar", HARTMANN (1983), pp.98-107.

DOLEŽEL, L., BAILEY, R., (eds.) Statistics and Style, New York, 1969.

DOLEŽEL, L., A "Framework for the Statistical Analysis of Style", DOLEŽEL, BAILEY (eds.), (1969), pp.10-25.

EVENS, M.W., SMITH, R.N., "A Lexicon for a Computer Question-Answering System", American Journal of Computational Linguistics, XV(1978)4, pp.1-101.

FARR, G., "Lexicography and the Computer in the United States: Projects supported by the National Endowment for the Humanities", ZAMPOLLI, CAPPELLI (eds.) 1983, pp. 107-118.

GOETSCHALCKX, J., ROLLING, L., (eds.) Lexicography in the Electronic Age, Amsterdam, 1982.

GORCY, G., "L'informatique et la mise en oeuvre du Trésor de la Langue française, Dictionnaire de la Langue du 19e et du 20e siecle (1789-1960)", ZAMPOLLI, CAPPELLI (eds.) 1983, pp. 119-144.

HARTMANN, R.R.K. (ed.), Lexicography: Principles and Practice, Academic Press, 1983.

HEIDORN, G.E., JENSEN, K., MILLER, L.A., BYRD, R.J., CHODOROW, M.S., "The EPISTLE Text-Critiquing System", IBM System Journal, 21(1982)3, pp.305-326.

HOWLETT, D.R., "The Use of Traditional and Computer Techniques in Compiling and Printing a Dictionary of Medieval Latin from British Sources", ZAMPOLLI, CAPPELLI (eds.) 1983, pp. 153-160.

JUILLAND, A., CHANG-RODRIGUEZ, E., Frequency Dictionary of Spanish Words, The Hague, 1964.

JUILLAND, A., TRAVERSA, V., Frequency Dictionary of Italian Words, The Hague, 1973.

KAY, M., "The Dictionary of the Future and the Future of the Dictionary", ZAMPOLLI, CAPPELLI (eds.) 1983, pp. 161-174.

KIPFER, B., Computer applications in lexicography, Summary of the state of the

art, 1983.

KIPFER, B., "Bibliography of Computer Applications in Lexicography and Lexicology", Dictionaries, 1983.

KNOWLES, F., "Towards the Machine Dictionary", HARTMANN (ed.), 1983, pp.181-193.

LENTZ, L.T., "An Integrated Computer-based system for creating the Dictionary of the Old Spanish Language", Linguistica Computazionale, I(1981) pp.19-42.

Les Machines dans la Linguistique, Praga, 1968.

LYONS, J., Semantics, Cambridge, 1977.

MACDONALD, N.H., FRASE, L.T., GINGRICH, P.S., KEENAN, S.A., "The Writer's Workbench: Computer Aids for Text Analysis", IEEE Transactions on Communication, 1982.

MARTIN, W.J.R., AL, B.P.F., van STERKENBURG, P.J.G., "On the processing of a text corpus", HARTMANN (ed.), 1983, pp.77-87.

MICHIELS, A., 1981. Exploiting a Large Dictionary Data Base, University of Liège, 1981. (Ph.D. Dissertation).

NAGAO, M., TSUJII, U.Y., TAKIYAMA, M., "An Attempt to Computerize Dictionary Data Bases", GOETSCHALCKX, J., ROLLING, L., (eds.), 1982, pp.534-542.

OLNEY, J., RAMSEY, D., "From Machine-Readable Dictionaries to a Lexicon Tester: Progress, Plans, and an Offer", Computer Studies in the Humanities and Verbal Behavior, 3(1972)4, pp.213-220.

PETERSEN, P.R., "New Words in Danish 1955-75. A Dictionary compiled and worked out in a traditional way and managed and typed via Computer", ZAMPOLLI, CAPPELLI (eds.) 1983, pp. 179-186.

PFISTER, M., "Présentation du LEI (Lessico Etimologico Italiano): Possibilités d'Établir des Index Lexicaux et Morphologiques par Ordinateur", ZAMPOLLI, CAPPELLI (eds.) 1983, pp. 187-200.

PICCHI, E., "Problemi di documentazione linguistica. Archivio di testi e nuove tecnologie". Studi di Lessicografia Italiana, V(1983), pp. 243-252.

PLANTE, P., "Le Système de programmation Deredec", HOTS, VI(1983).

QUEMADA, B., "L'automatisation de la recherche lexicologique: état actuel et tendences nouvelles", META, XVIII(1973)1-2.

QUEMADA, B., "Présentation du Programme", ZAMPOLLI, CAPPELLI (eds.) 1983, pp.13-31.

RATTI, D., SABA, A., CATARSI, M.E., CAPPELLI, G., Analizador Morfosintactico de textos en lengua española, Pisa, 1982.

ROSENGREN, I., "The quantitative concept of language and its relation to the structure of frequency dictionaries", Études de linguistique appliquée, I(1971)1, pp.103-127.

TOMBEUR, P., "Propositions Nouvelles pour une Lemmatisation Unifiée du Latin", ZAMPOLLI, CAPPELLI (eds.) 1983, pp. 207-228.

VENEZKY, R.L., "User Aids in a Lexical Processing System", LUSIGNAN, S., NORTH, J., (eds.), Computing in the Humanities, Waterloo, 1977, pp. 317-325.

VENEZKY, R.L., RELLES, N., PRICE, L., "LEXICO: A System for Lexicographic Processing", Computers and the Humanities, 1977, pp.127-137.

WALKER, D.E., AMSLER, R.A., "The Use of Machine-Readable Dictionaries in Sublanguage Analysis", KITTREDGE, R.I., (ed.), Workshop on Sublanguage Analysis, New York, 1984.

ZAMPOLLI, A.,(ed.), Linguistica Matematica e Calcolatori. Proceedings of the First International Summer School, Firenze, 1973a.

ZAMPOLLI, A., "L'automatisation de la recherche lexicologique: état actuel et tendances nouvelles", META, XVIII(1973b)1-2, pp.101-136.

ZAMPOLLI, A., "L' elaborazione elettronica dei dati linguistici: stato delle ricerche e prospettive", Roma, Accademia dei Lincei, 1975.

ZAMPOLLI, A. (ed.), Linguistic Structures Processing, Amsterdam, 1977.

ZAMPOLLI, A., "Lexicological and Lexicographical Activities at the Istituto di Linguistica Computazionale", ZAMPOLLI, CAPPELLI (eds.) 1983, pp. 237-278.

ZAMPOLLI, A., CALZOLARI, N. (eds.), Computational and Mathematical Linguistics, I, Olschki, Firenze, 1977.

ZAMPOLLI, A., CALZOLARI, N. (eds.), Computational and Mathematical Linguistics, II, Firenze, 1980.

ZAMPOLLI, A., CAPPELLI, A. (eds.), The Possibilities and Limits of the Computer in producing and publishing Dictionaries, Linguistica Computazionale, III (1983).

ZIEBERMAN, H.L., "Multifunctional Dictionaries", ZAMPOLLI, CAPPELLI (eds.) 1983, pp. 279-288.