Papers presented at the Council of Europe

meeting on 'Language Industries'

Istituto di Linguistica Computazionale - CNR

Pisa - 1987

Lexical Data Bases and Lexicographical Activities*

A. Zampolli, L. Cignoni, S. Rossi

A. For a better definition of the object of this paper, the multifunctional lexical database, we shall now consider a number of concepts which have emerged recently and which, in our opinion, indicate an important evolution within the field of computational lexicology and lexicography.

1.1 Texts in machine readable form (MRF)

Textual Archive: some enterprises aim at collecting and storing in machine readable form (MRF) as many texts as possible, from different (internal and external) sources. The main goal is to maintain the texts in MRF available for researchers other than those who first recorded them (e.g. the archive at the Oxford Computing Centre). Some enterprises, in addition to maintenance, convert all the texts to the same encoding standard, e.g. the Institute for Computational Linguistics (ILC) of CNR - Pisa (Zampolli, Brogna, 1979).

Textual Data Bank: a homogeneous set of texts is recorded in MRF as a source of textual documentation for a specific task or for a well defined language subset. Well known examples are the Tresor de la Langue Francaise of the INALF-CNRS (Gorcy, 1983), and the Thesaurus Linguae Graecae of Irvine (TLG).

Textual Data Base: one or more textual corpora are structured for interactive consultation and analysis and associated with specialized software for management, maintenance, access, interrogation, on-line or batch processing. The number of projects aimed at restructuring textual archives or banks in the form of data bases is now increasing: ILC (Zampolli 1983, Picchi 1983), INALF (Dendien 1986), Institut fur Deutsche Sprache, Shakespeare Corpus (Neuhaus 1985), TLG (TLG Newsletters)

Lexicographical Textual Data Bases: a particular sub-type of textual DB, conceived as the documentary basis for the editing of a dictionary, aimed at providing the lexicographer with the textual material which is suitable to his purposes. This generally - but not always - consists in a lemmatized collection of quotations selected from the basic textual bank/archive.

1.2 Machine Dictionaries: a set of lexical units recorded in machine readable form.

"Bases de donnees dictionnairiques": the set of different elements progressively collected and organized by the editor (or the editorial staff) in the computer-assisted process of constructing and structuring the basic material for the dictionary entries. This material often has a very evolutive character, as requested by this highly creative stage of the "dictionnairique" work: modifications and readjustments are permanently possible. (At the end of the process, the "base de donnees dictionnairiques" in its final consolidated form, may be regarded as a lexicographical DB. (See Quemada 1983, p. 28).

Lexicographic Data Base

The lexicographer inputs the lexical data into a database-like structure of (hierarchically ordered) tagged fields. A program, using information conveyed by this structure and the lexicographic tags, automatically generates the text of the dictionary in a format directly suitable for photocomposition.

Printed dictionaries in machine readable form: the text of a printed dictionary recorded in MRF. Three main source types may be distinguished:

a) The text of a dictionary is directly typeset in a format and with the typographical commands suitable for photocomposition (e.g. the Garzanti and the Collins Concise Italian-English bilingual dictionary (1986);

b) The text of the dictionary is automatically generated for photocomposition from a lexicographical data base.

c) A dictionary already printed with traditional methods is (progressively) recorded in MRF, both for research (Webster's Pocket Dictionary, Olney 1972) and/or editorial purposes, (for example new updated editions of out-of-print dictionaries). Optical character readers are beginning to be used to transfer the printed version into MRF.

d) The dictionary is made interactively available on an on-line 'mainframe', or on an electronic support CD-ROM, etc. connected to a PC. The first projects have just been started. The best known example is probably the new computerized OED project.

Dictionaries in MRF for computer program use (NLPD): a set of lexical units is stored in MRF with linguistic information encoded in a formal representation suitable for direct use of programs embedded in systems of natural language processing (lemmatization in literary text-processing, parsers, generation,

bilingual transfer, etc.). These programs usually include (or have access to) a look-up component which searches in the dictionary the lexical entry corresponding to the text-form to be processed.

MRD for linguistic researches: a set of lexical units is recorded, with more or less formalized descriptions of their properties at different linguistic levels, structured as computer processable representations of a subset of a lexical system, explicitly designed for linguistic and psycholinguistic researches (Josselson 1969; Zampolli 1968; Gross 1975, Schroeder 1986).

Multifunctional Lexical Data Bases (MLDB)

Various projects have tried to produce MRDs with one or more functions. For example, our Italian Machine Dictionary has been conceived from the very beginning as a tool for semiautomatic lemmatization, a component for parsing and generating systems, a computational representation of the Italian system for linguistic research (Zampolli 1968).

Nowadays, owing to a number of circumstances and reasons, we must ask ourselves whether it is correct to conceive and/or whether it is feasible to realize in practice a multifunctional lexical data base: i.e., a collection of lexical units in MRF, structured and accessible using DB methods, able to supply, directly or by specialized interface components, the information requested by most of and possibly all the different types of utilizations: production of various types of dictionaries, both printed and on electronic support for human users; (semi)automatic text lemmatization; spelling checkers; parsers and generators; lexical knowledge-based full-text retrieval; linguistic and psycholinguistic researches, etc.

1.3 Integrated Linguistic Data Bases

The next step could possibly consist in designing an integrated linguistic DB: a structured set of interrelated DBs of different nature and level (textual, lexical, grammatical, bibliographical, sociolinguistic, etc.), associated to an open-ended set of computational linguistics modular software components performing different tasks (dictionary look-up, quantitative investigation and statistical calculus, data access, etc.), located within an organizational framework which provides maintenance, distribution, updating, specialized working stations, international exchanges, copyright protection, user advice and training, etc.

B. The concept of a MLDB is only just emerging (useful discussions have taken place at two recent workshops: at Grosseto in May, and in New York in July) and must be carefully verified at the scientific, organizational and practical level.

We think that the goal of a MLDB must be pursued progressively, the first step consisting in assessing if a unified LDB is feasible and desirable within the different applications of a specific domain, e.g. the production of different types of dictionaries; different activities in the field of NLP; etc.

Given the specific institutional goals and the nature of our Institute, at present we are working in the following directions:

- possibility of a "neutral" LDB for NLP, which can be used from parsers and generators working in different computational and theoretical linguistic frameworks;
- possibility of reusing information explicitly or implicitly available in MRDs, and extraction methodologies;
- structure of the linguistic information to be inserted in our existing Italian data base (DMI-DB), and priorities;
- methods for constructing new types of information from the data already present in our DMI-DB: e.g., a conceptual taxonomy (and other semantical relations) from the definitions;
- possibilities and methods for the connection of a monolingual Italian LDB to the LDB of another language, via a bilingual DB, constructed through the (semi)automatic exploitation of a bilingual MRD;
- methods to connect the LDBs to textual corpora, monolingual and, possibly, multilingual (both of independent or contrasted corpora)
- general structure and access methods for our integrated linguistic DB (monolingual LDB, plus bilingual LDBs, plus corpora)
- different categories of possible partners in the construction of the LDB: linguists, publishers, etc.
- international cooperation and normalization: not only for the construction of bilingual DBs, but also for the normalization of the linguistic content, representation formalism, definition of the information units, access methods, etc.

We shall examine in more detail the following points:

1. Normalization of syntactical information
2. Relations in the lexicon
3. Bilingual components
4. Lexicographic Workstation

1. Normalization of syntactical information

A large amount of work in computational linguistics is carried out on experimental lines, with consequently small-sized lexical prototype systems. Emphasis is traditionally placed on linguistic knowledge expressed by linguistic rules and procedures, and lexical data seem to be considered of less "scientific" relevance, and easy to deal with.

At a recent workshop at UMIST, the average size of the lexicons used by the systems of the invited speakers was about 25 words.

Our computational linguistic community has been recently faced with the request of large scale NLP systems, by supranational and national, public and private organizations. For real word applications, it is necessary to deal with tens of thousands of lexical items. The preparation of adequate large automatic lexicons can be no longer delayed. Up until now, it has been a fact that each NLP system has itw own ideas and conventions with regard to the content, organization and structure of its lexicon. This will inevitably involve the duplication of efforts in an extremely costly and time-consuming task.

In the present situation, not only researchers and developers, but also the promoting and financing authorities put forward the question as to whether it is possible to design a large flexible LDB, where different linguistic theories and CL systems can find the relevant lexical information required.

A sound methodology to answer this question consists in starting to review existing parsers and generators and in examining the information contained in the lexicons and the way they are represented.

A first review was presented at the Grosseto workshop by B. Ingria (1986). His conclusion was that the information of the lexicons in semantically oriented systems, or of systems which include a large-scale syntactic component, cannot be compared easily. The lexical entries are very often complicated programs with routine structures, the specification of which requires detailed knowledge of the architecture of the parser, judgement of what constitutes linguistically relevant information and how to translate that procedurally, and readiness to bring in an arbitrary amount of more general, common world knowledge.

There is no firm consensus on what types of structure are best suited to capture the knowledge useful for a particular NLP system, which is often bound to a very restricted specific domain.

The situation seems to be different for the lexicons of syntactically oriented systems.

Here it is possible to recognize a convergence towards the inclusion of the same kind of linguistic information.

Present systems represent this information in very different ways. However a careful examination leads to the intuition that this information can be compared: in a certain sense, it describes, with different representations, the same 'linguistic facts'. If this intuition is true, it should be possible to represent the syntactical properties of the words in a neutral way, and to create interfaces which automatically transform this representation into the formalism required by each linguistic theory.

We have promoted a feasibility study, which will first have to recognize and describe the different linguistic theories for which it is possible to realize this task, describe the different grades of feasibility, and suggest a working methodology.

This task is complicated by the diversity among syntactic frameworks that can be used in parsers and generation systems. Nevertheless, the key observation is that, although linguistic frameworks differ in the way they analyze syntactic structures, they do not basically differ in the extension of the classes they propose.

The feasibility study, involving representatives of the major linguistic schools of five countries, will examine in particular the following topics:

- major syntactic categories
- subcategorization and complementation
- verb classes
- nominal taxonomies

as used in the following frameworks:

- Government and binding
- Generalized phrase structure grammar
- Lexical functional grammars (in different variants)
- Systemic grammar

2. Relations in the Lexicon

Computational techniques traditionally apply to the "lexicon" in at least two main directions:

- in the various phases of "dictionary making" (computational lexicography);
- in the analysis and organization of the "lexicon" proper (computational lexicology).

In a certain sense, the combination of the two disciplines with new technologies and methodologies in computer sciences has recently given rise to a new trend which seems to be very promising, namely the creation of new types of dictionaries on electronic devices, or lexical data bases where, compared to traditional printed dictionaries and machine readable dictionaries, the structure, access methods and searching procedures are completely different. Furthermore, the information resulting from the original data is richer and much more complex. This means that as a consequence of the application of computational techniques to the lexical data also the "content" becomes in a certain sense different.

This is the direction in which we have been working over the past few years in Pisa, where we are creating a large lexical data base of the Italian language.

The aim of the project is to create a large repository of lexical data organized in data base form, where lexical units are stored together with many kinds of lexical properties and lexical relations, and where access is provided at the various levels of lexical units, properties, and relations.

Furthermore, part of the information (i.e. definitions) - which in usual dictionaries is provided in natural language - in our linguistic data base is processed by definition parsing procedures and is transformed either into properties (for inherent features), or into attribute value pairs, or into qualified relations and pointers (for morphological and semantic relations).

Thus the entry has a formal part at all linguistic levels, and also at the semantic level.

Our programs parsing existing machine readable dictionaries can extract the different types of information by decoding the typesetting codes, and can distribute this information to its appropriate location in the above model entry-scheme.

We tend towards a dictionary representing lexical information by relations from words to words, or from words to metalinguistic codes, using existing dictionaries as one of the sources for the raw data.

The lexicon will appear as virtually divided into as many subsets as the relations which have been determined and formalized. The values of some relations will range over restricted sets (e.g. of syntactic categories, usage labels, inflectional codes, etc.), while the values of other relations will range over considerably larger and less determined sets (i.e. the very words of the lexicon differently grouped and picked up according to the different relationships, e.g. by synonyms, antonyms, derivatives, thematic roles, etc.). By representing the lexicon as the set of all these relations, we can access the dictionary either by lexical items, or by features, or by relations, search the network to see where it matches with the query, and retrieve different parts of the lexical content on the basis both of the access point and of the options activated on this point.

Thus a suitable structure enables us to provide wide access to lexical information both in breadth and in depth for a number of foreseeable applications.

In the Italian LDB, a number of systematic lexical semantic relations can actually be detected or retrieved or activated by the different modes of access already implemented. The relations which are now implemented or are at an advanced stage of implementation are the following:

a) Hierarchical relations, based on hyponymy. These have already been implemented all over the lexicon by a number of programs which process existing natural language definitions in the MRD-source. So we can interactively ask for example for all the "kinds of" or "names of" vehicles, or of colours, sounds, instruments, etc.

b) Synonymical relations, in both directions.

c) Derivational relations, for dealing with word-formation.

d) Other taxonomies, based on relations such as 'Part-Whole', 'Set-of', etc. are implemented too.

e) Co-occurrence or collocational relations can be detected.

f) Terminological sub-lexicons can be easily set up.

g) Restriction or modification relations are being implemented (see Calzolari, 1984a). These are very important to characterize the internal structure of words, and once formalized they will constitute a first step towards the development of a knowledge base (KB). I refer here to relations such as 'in the form of', 'apt to', 'used for', 'provided with', 'consisting of', etc.

h) Case-type or argument relations can also be inquired.

i) Semantic fields, i.e. generically-related words can obviously be retrieved, even though not exhaustively (if the word "exhaustive" can be given any precise meaning in this context).

j) Selection restrictions: we can gather significant data in particular for a number of Adjectives. For instance, together with an adjective, we can often retrieve the information of which types of Nouns it can be predicated, or viceversa.

When taking into account all these functions, it is absolutely necessary to make it possible for a LDB to provide the values (which are words) that fill these types of lexical-semantic functions in relation to other words. In this respect it is important to work on the entire lexicon. As a matter of fact these relations are being implemented in the Italian LDB over the whole set of almost 200,000 natural language definitions of about 106,000 lemmas. We can for example already retrieve on-line from the LDB those lemmas with a given grammatical/syntactic code, or only dialectal lemmas, or we can interactively ask for words with a given ending, or synonyms, hyponyms, and so on along a number of different dimensions. These different types of searches evidence that different modes of access give rise to lexical activation of differently related

groups of entries. One of the purposes of implementing these relations is to transform a particular natural language text, i.e. definitions (in a certain sense a sublanguage, with lexico-grammatical restrictions which are very useful to exploit) into a knowledge base.

The formal information present at the semantic level in a monolingual dictionary - which serves to discriminate among the different word-senses - should be in principle of the same type that is given in bilingual dictionaries in the form of "semantic indicators" or "selective information" to guide the choice of a particular translation.

The problem of mapping between word-senses in monolingual dictionaries and different translations in a bilingual dictionary is one of the most interesting among those concerning the connection of these different types of dictionaries.

As one of the main problems in translation is the capability of choosing among the various meanings of lexically ambiguous words, we feel that it is absolutely necessary also for a Machine Translation or a Machine Assisted Translation system to be linked to a linguistic data base, i.e. a source of lexical information organized in the form of a thesaurus by multi-dimensional taxonomies, where the possibility of disambiguating lexical items is at least semiautomatized.

3. Bilingual Components

A new direction towards which computational lexicography and lexicology are moving is the organization of bilingual lexical data base systems.

We are now working on a project which uses bilingual machine readable dictionaries as a "bridge" connecting two otherwise independent monolingual lexical data bases.

One of the objectives is to integrate the different types of information traditionally contained in monolingual and bilingual dictionaries, so as to expand the informational content of the single components in the new integrated system.

The entries of the new system should therefore be of a composite nature, and perhaps organized at different levels according to the different possibilities of access.

We can envisage the original monolingual lexical entries, augmented with the different types of information coming from the corresponding bilingual entry: different sense discriminations, other examples, syntactic information, collocations, idioms, etc.

We can also reverse the perspective, and look at the bilingual entries augmented with the information traditionally contained in monolingual entries: mostly definitions.

One of the two different viewpoints, both virtually present in the integrated bilingual system, will be simply activated and made available to the user by the first way of accessing to the on-line bilingual lexical data base.

We would like to maintain in a unique structure both the independent feautures of the source monolingual and bilingual dictionaries and the integration of the two with different views on the data.

Moreover, we would like to introduce within the integrated system:

- the possibility of a standard look-up to the information given in natural language in traditional dictionaries;
- more sophisticated searching procedures for information retrieval operations on the data of the mono- and bilingual lexical data bases where the ″natural language″ data have been, where possible, transformed into ″formalized″ or ″coded″ data (features or relations). For example, the information which appears in the form of examples can also appear in a coded form giving the surface syntactic structure.

An obvious case where a monolingual lexical data base organized also as a thesaurus proves to be very useful in extending the information provided by a bilingual dictionary concerns what in the Collins English-Italian Dictionary are called ″semantic indicators″. These may be field labels, synonyms, hyponyms, or contextual indicators as typical subjects or objects of verbs, typical nouns of which an adjective can be predicated, etc.

Now, the monolingual lexical data base can be used to expand this information given as a single word to the whole set of words to which it actually refers.

For example, the entry ″guaire″ has different translations according to the following contextual indicators for the subject (in brackets):

guaire... (dog) to yelp, to whine

(person) to whine

The generic semantic restrictions on the subject can be taken in a certain sense as a semantic feature, and can be procedurally expanded by the monolingual thesaurus to all the possible hyponyms (at the query moment) so that the appropriate translation can be chosen in any context where a specific name of ″dog″ or a specific ″human″ noun is found.

The method we are adopting is that of analyzing and transforming the information already contained in usual dictionaries.

Some phases that we are envisaging as first steps are the following:

- reversing of the two sides of the dictionary;
- unification of identical information, now present in the two parts. This unified information will obviously be addressed from all the different pertinent access points. A result is the elimination of redundancy at the level of the unified system.
- integration of information which is only available in one part, and "missing" in the other. Obviously, it is our intention to normalize this kind of fragmentary and dispersed information, by integrating the data found in different places, and also making it possible to access this same data from every pertinent entry for retrieval operations. At this point, it makes no difference which of the two languages are taken as a starting point. In a certain sense, we would no longer have a source language and a target language, since the look-up and access procedures are independent and neutral with respect to direction (the object becomes bidirectional). Bidirectional cross-references will also be automatically generated for the information contained at each sense level as semantic indicators, i.e. synonyms/hyperonyms or contextual indicators.

One of the main uses of the system should be that of machine-aided translation (MAT), as a powerful aid for translators.

The end result may in fact be viewed as a 'translator workstation', where access is provided to many types of dictionaries and other lexical resources, and where the power and the functions of lexical data bases and of textual data bases is exploited at best.

4. Lexicographical workstation

It is our intention to develop a lexicographical workstation. We conceive a tool of this type as a means which the lexicographer can use when preparing his dictionary entries. In this way he can interact with a lexical database consisting of a set of different types of information (corpora of texts, lexicographical archives, already existing dictionaries, bibliographical references, etc.) available on-line and accessible using appropriately designed software tools.

A system of this type can present all the lexical information not only in an attractive but also in a functional form, which can be seen as the basis for the dictionary of the future in which components of an underlying lexical database are merged and welded together and can also be accessed dynamically by the system.

One of the basic needs of the lexicographer is to update the existing dictionary. Using the lexical database and analysing new texts to retrieve new citations to be inserted, he can thus adapt or create ex-novo the lexical entry.

This aspect of the work can be seen as part of the more general integrated system for the creation of new dictionaries, in which all the functions from data acquisition to the printing of the results also using photocomposition systems, are processed and handled using computerized procedures.

Text acquisition which in the past had to be manually input into the system can now be facilitated by the use of optical readers, by the possibility of using texts prepared for photocomposition by publishing houses and also by the increase in lexicographic studies which use computing resources.

A fundamental software component of the system is that which permits the memorization and quick access to the data 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). The component provides the lexicographer with information on the frequencies of distribution of the searched elements, for different sections of the corpus. 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.

Exploiting the multiple dictionary access tecniques described above which make it possible to query the dictionaries stored in the lexical database from different access points and at various levels, another component of the system permits the consulting of both existing dictionaries and those in the compilation stage.

Specific functions select the information searched using the previous components and storing it with links to the entries being edited.

Other functions make it possible to define the structure as the dictionary entry, permitting the integration of information retrieved during searches on existing dictionaries. The system also permits the cyclic classification of the previously selected contexts according to the different sections of the structure defined and the reordering of these contexts in relation to the structure itself.

The lexicographers using recursive functions compile the entry and classify the context thus obtaining increasingly updated versions of his new dictionary entry. All stored information can be altered, expanded

and corrected at any time and consulted immediately within the new dictionary, in order to help homogeneity and coherence.

All the stages of the lexicographer's work must be analyzed and collaboration between the lexicographer and the computational linguist is essential in order to construct a tool which is both complete and efficient.

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 continuing in the production of 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 can be numbered. The editor then keys in only the microstructure (headword, etymologies, grammatical information, definitions, etc.) and then, 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 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.

* Parts of this text are extracted from the articles: A. Zampolli, 'Perspectives for an Italian Multifunctional Lexical Data Base' and A. Zampolli, N. Calzolari, E. Picchi, 'Italian Multifunctional Data Base'.