PERSPECTIVES FOR AN ITALIAN MULTIFUNCTIONAL LEXICAL DATABASE

ANTONIO ZAMPOLLI

1. SOME TERMINOLOGICAL REMARKS

At the beginning of the 60s it was evident that the pioneer work of R. Busa had resulted into a considerable increase in the field of literary and linguistic text processing, in particular for the creation of indexes and concordances. The problem therefore rapidly emerged of a classification of these activities. For example De Tollenaere (1963, *Introduction*) included these activities within the field of lexicology in opposition to lexicography[1].

As a matter of fact, a number of terminological uncertainties persist nowadays. In 1981 the European Science Foundation circulated a questionnaire for a «Survey of Lexicographical Projects» in Europe. The over 150 replies included a large typology of projects. In fact not only dictionary making, but also a variety of computer lexicons, indexes and concordances were considered as «lexicographical» (Cignoni *et alii*, 1983).

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 my opinion, indicate an important evolution within the field of computational lexicology and lexicography[2].

1.1 *Texts in machine readable form*

Textual Archive: some enterprises aim at collecting and storing in machine readable form (MRF) as many texts as possible,

1. «The collecting of the material to be made use by lexicographer, is done by lexicology» (DE TOLLENAERE 1963, p. 128).

2. It must be underlined that I do not intend to trace a classification of the sector, but only to clarify my use of terms which are used in different ways, and which however, as a complex, design the present «zeitgeist». I shall outline these words following Quemada (1983) and Zampolli (1983), even if with some changes.

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. our 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 Trésor de la Langue Française 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 fuer Deutsche Sprache, (Teubert, 1984), 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[3].

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

«Bases de données 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.

3. The creation of these data bases involves a number of problems at the methodological and operational level, described in Zampolli (1983).

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 données 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 by the lexicographic tags, automatically generates the text of the dictionary in a format directly suitable for photocomposition[4].

Printed dictionaries in machine readable form: the text of a printed dictionary recorded in MRF. Four 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, 1985);
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, (e.g., new updated editions of out-of-print dictionaries). Optical character readers are beginning to be used to transfer the printed version into MRF[5].
d) The dictionary is made interactively available to commercial users 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[6].

4. A typical example is the COMPULEXIS system (see Madsen, 1986), which is used in different lexicographical enterprises.

5. 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 to assess whether the typographical structure of the printed dictionary may be used to create, automatically or semiautomatically, explicit linguistic information. See Nistrup Madsen (1987).

6. See Benbow (1987).

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, Schreuder 1987).

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 kinds of dictionaries, both printed or 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.
After a brief overview of the development and the present situation of this field in Italy, we shall briefly discuss the problems and the perspectives for the creation of an Italian Multifunctional Lexical Data Base (IMLDB).

2. DEVELOPMENT OF MACHINE-DICTIONARIES IN ITALY

To my knowledge, the first (or one of the first) machine dictionaries for the automatic processing of literary and linguistic texts was created in 1960-1 for the lemmatization of the works of St. Thomas Aquinas[7].

In 1965-66 the Accademia della Crusca started the construction of a textual data bank for the preparation of the *Tesoro delle Origini* and of the *Dizionario storico della lingua italiana*.

7. R. Busa appointed Zampolli to head a large group of researchers working on the lemmatization of the 100 million words of the Index Thomisticus. Zampolli was immediately faced with the problem of optimizing execution times and of ensuring consistency among the different lemmatizers. He designed a procedure of semiautomatic lemmatization which utilized the technologies available at that time for projects in the field of the humanities: collating, reproducing, sorting, reproducing UR machines. The logical «schemata» of the procedure proposed at that time was essentially the same which is used today (with the computers) in several Centres for the literary and linguistic analysis of texts. It included:
A table of high frequency forms. The most frequent (50) forms appearing in the first subset of S. Thomas' texts were stored, together with their lemmas, in a table. The text-words were looked up in this table, and lemmatized in order of text.
A 'machine dictionary'. An alphabetical ordered set of forms (each stored together with its lemma). The words of a new text submitted for lemmatization were collated with the forms in the dictionary. The words found in the dictionary were lemmatized automatically if univocal. The homographic forms and the unfound words were submitted to the manual analysis of the lemmatizers. The new words were gradually added to the dictionary, which was thus progressively enriched with the analysis of new texts.

At that time the first computers were available for literary and linguistic researches at the CNUCE (IBM 7090 and, then, IBM 1401) and a first experience for semiautomatic lemmatization, which was conducted in 1967 (Duro, A., Zampolli, A., 1968), mainly consisted in using the manually lemmatized forms of the first chapters of the *Promessi Sposi* by Alessandro Manzoni to lemmatize the words of the following chapters.

These experiences, along with the contemporary evolution leading towards a growing interest of the linguistic theories for the role and the function of the lexicon in the linguistic models, led Zampolli to suggest the creation of a multifunctional *Italian Machine Dictionary* (DMI; Zampolli, 1968), which was to meet the following demands:

- to allow the semiautomatic lemmatization of contemporary Italian texts, and possibly their automatic phonological transcription;
- to provide the necessary information to parsers and generators of Italian sentences;
- to constitute a computational representation of the Italian lexical system, allowing a number of studies and researches on the quantitative and qualitative structures.

In its first version, the DMI was composed by a list of approximately 120,000 lemmas, obtained comparing the IX edition of the Zingarelli with other Italian dictionaries. About 1,000,000 forms were automatically generated by a flexional algorithm, which operated on the morphological codes semi-automatically associated to each lemma (Ferrari, 1973).

At a following stage, owing to a contribution given to the University of Pisa by the House of Deputies, 300,000 definitions were added in machine-readable form[8].

The creation of the Institute for Computational Linguistics made it possible to set up a project for the development of the DMI in the form of a relational data base (DMI-DB) (Calzolari, Ceccotti, Roventini, 1980).

The DMI-DB is actually used in a number of projects, which aim at representing in a formal way various types of grammatical and semantical relations (hyponymy, synonymy,

8. The DMI consists of three linked files: the File of Lemmas (DMIL); the File of Forms (DMIF); the File of Definitions (DMID).

derivation, etc.), implicitly or explicitly underlying the definitions (Calzolari, 1987).

The future development of the DMI-DB is now partly entrusted to the strategic project «natural language processing» of the CNR, coordinated by A. Zampolli. We shall describe here the main lines of development, which largely coincide with the trend towards the creation of *multifunctional lexical data bases*, which in the last two years seem to have emerged in various research groups of different countries.

The general organization of the DMI at the level of lemmas and word-forms, and the problems still to be solved (derivational morphology; idioms, collocations and, in general, multiword-entries; proper nouns, abbreviations, acronyms; etc.) have already been discussed in this journal (Zampolli, 1983), to which the reader may refer.

We shall now consider the following problem areas, which concern future research and development perspectives of the IMLDB.

a) *Feasibility of a linguistically neutral LDB.*
Is it possible to represent in the lexicon a subset of linguistic information (syntactical, semantical, etc.) in such a way that it is independent of the commitment to specific linguistic theories and computational framework?

b) *Reusability of lexicographical information*
How can the information presently available in printed MRDs be reused for the (semi)automatic enrichment of the IMLDB?

c) *The monolingual IMLDB in a multilingual environment*
How is it possible to use the information available in bilingual dictionaries in connection with monolingual Lexical Data Bases?

d) *Possible cooperation between linguists, computational linguists, and lexicographers.*

3. FEASIBILITY OF A «NEUTRAL» LDB

A large amount of work in computational linguistics (CL) is carried out on experimental lines, with consequently small-sized lexical prototype systems. Furthermore, emphasis is traditionally placed on the representation, organization and use of linguistic knowledge as encapsulated and expressed by linguis-

tic rules and procedures. Lexical data seem to be considered of secondary importance or, at least, easy to be handled.

At a recent workshop (UMIST, 1985) (McNaught, 1986) an informal poll – constituting a good representative sample of today's computational linguists – was conducted among the invited speakers, to establish the average size of the lexicons used by their systems. With the exception of a prototype MT system, the average size was about 25 words. This observation may be extended to a large number of systems within the computational linguistics paradigm, MT systems being the only eventual exceptions[9].

Today we are forced to consider the following facts:

a) Our CL community has been recently faced with the request of large scale NLP systems, owing to the recent advances in CL technology which make such applications feasible, and the interest expressed by supranational and national public and private organizations.

b) For real world applications, it is of fundamental importance that a CL system is able to deal with tens of thousands of lexical items. The projects presently underway must be accomplished within reasonably fixed time limits. The preparation of NLP dictionaries can be delayed no longer.

c) Various projects have been promoted for the same language. Up until now, it has been a fact that each system has its own ideas and conventions with regard to content, organization and structure of its lexicon. This makes it difficult or even impossible to share linguistically relevant information, across various NLP systems, for the same language.

d) Duplication of efforts may be a very «sad» fact. Building a comprehensive consistent NLPD is probably the most costly and time consuming task in every NLP project. In this situation, it is natural that not only researchers and developers, but also the promoting and financing authorities should put forward the question as to whether it is possible to design a rich, powerful and flexible LDB, where different linguistic theories and CL systems can find the relevant lexical information required.

9. The explanation is probably that Mt systems are the only NLP systems effectively and explicitly intended from the very beginning for real world applications. Spelling checkers are not considered here, because the function of the lexicons is limited, in general, to the recognition of orthographic words.

The problem of the feasibility of a neutral LDB, and the assessment of even partial solutions is obviously of primary importance for us, as we are only just starting to define the content and the representation of the syntactic and semantic information which we have to implement in the IMLDB.

In the following paragraphs we shall often refer to the papers and the discussions held at the workshop on «Automating the Lexicon», which we organized at Grosseto in May 1986[10].

3.1. *Methodological Issues*

It is a well recognized fact that different linguistic theories and different computational organizations may have important consequences on the grammar contruction. Less attention has

10. The workshop was sponsored by the European Community, the University of Pisa, and the Institute of Computational Linguistics of the Italian National Research Council. It was also under the auspices of the Association for Computational Linguistics, the Association for Literary and Linguistic Computing, EURALEX, and the AILA Commission on Computational Linguistics. The second in a series, it built on its predecessor, which was held in April 1983 at SRI International in Menlo Park, California.

Where the first workshop examined the machine-readable dictionary from the perspective of the research community, the publishers, and the emerging market intermediaries, the second was much broader. Its purpose was to explore research efforts, current practice, and potential developments in work on the lexicon, machine-readable dictionaries, and lexical knowledge bases with special consideration for the problems created by working with different languages. The intent was to identify the current state of affairs and to recommend directions for future activities.

To help in the realization of these objectives, a set of papers was solicited for the workshop under the following general headings: Research Areas, Core Problems, Application Areas, and Developing Research Resources. People were asked to prepare comprehensive surveys and evaluations of activities going on in the field. We also requested reports on national projects in related areas. At the end of the agenda a «Consolidation» session was scheduled to consider the following topics: the lexical entry as a basis for integration, cooperation and communication, priorities for research and development, and the next steps. The workshop papers are in print in D. WALKER *et alii* (1987).

The participants were chosen to bring together representatives from the different kinds of areas that we believed were relevant to the various problems associated with the lexicon. This led us to invite linguists, lexicographers, lexicologists, computational linguists, artificial intelligence specialists, cognitive scientists, publishers, lexical software marketers, translators, funding agency representatives, and professional society representatives. This text is extracted from Walker, *Introduction*, in Walker *et alii*, 1987.

been paid to the consequences on the lexicon. However, we have the intuition that lexicons designed for different linguistic theories may contain information which from a certain point of view is identical, as it describes the same linguistic facts. We have to assess the validity of this intuition before starting to implement in the IMLDB the information required by the NLP systems. A sound methodology for the evaluation of this intuition may consist in the following steps:

- to review the existing parsers and generators for various languages, and to examine the information contained in their lexicons and the way they are represented, in order to assess the possibility of convergence[11];
- to conduct a feasibility study on a representative subset of the Italian lexicon to assess the possibility of designing an IMLDB which is «neutral» with respect to the major linguistic theories.

Let us consider briefly these two problems.

3.2. *Comparison of existing lexicons*

On the occasion of the workshop «On Automating the Lexicon», organized in Grosseto in May 1986 (see D. Walker, 1987), we have requested a comparative study of the lexicons used in computational parsers and generators to B. Ingria (1987) and to S. Cumming (1987) respectively.

A preliminary question to be answered is obviously whether and to what extent the directionality of linguistic processing – i.e. analysis or generation – influences the content of the

11. This task is not very easy. A strange situation exists in CL today: «We hear about all sorts of parsing and generating systems, but we do not have a store of information about what is actually available and where, and principled way of comparing them» (Letter to ACL). A survey will be launched soon, within the framework of ACL, on parsing grammars.

In the present version (July 1986), the questionnaire requests detailed information: types of processing for which the grammar is used; purpose; size; language (general purpose; LISP; PROLOG; or special purpose grammar writing languages); how the analysis is performed (rules, procedures, arc sets; etc.); examples and typical size of rules; how the information is expressed in the output of the parser; examples of analysis structures; classification of the grammars; ordering of the rules; characteristics of the grammar (top-down – bottom-up; deterministic not-deterministic; etc.). But no information is requested concerning the lexical components.

lexicon. Some systems are explicitly planned to be bidirectional, i.e. to use the same lexicon for both analysis and generation but, in practice, the two types of lexicons tend to be rather different.

«The generation tasks set different priorities for the lexicon: roughly speaking, a generation lexicon has to put depth before breadth, while the reverse is true for understanding» (Cumming, 1987).

- Parsers must be able to accept a variety of inputs from the user: the grammar must be comprehensive at least with respect to the subset of the language covered; the dictionary must contain a large number of words, and support all the syntactic distinctions that the grammar can make.
- A generator does not need a full range of syntactic capabilities (nor does it need a very large lexicon, e.g. one word for everything it needs to say, and fewer syntactic distinctions). Instead it has to know more about the syntax and the lexicon: it needs a basis for choosing between syntactic alternatives and lexical items, so as to be not only conceptually appropriate and grammatical, but also cooperative, idiomatic, non-redundant, etc.

An analogy can be made with the experience of the human learning of a second language: typically the range of the language which the learner can produce appropriately is more limited than the range which he can comprehend[12].

The conclusion at the Grosseto workshop was that parsers and generators may largely share the same bulk of lexical information. A LDB may easily contain the union of the knowledge requested by both. Some of the information will eventually be used only in one direction. We shall adopt this point of view for the IMLDB.

12. Linguists have always made a difference between active and passive competence. A concrete example is represented by the difference between «collocations» and «idioms». Collocations are compositional expressions, which may seem semantically transparent to a listener, but which require specialized knowledge on the part of the speaker to be produced correctly (idioms of encoding). Collocations may be parsed compositionally, but a generator must know the special connection between these words. Idioms are instead non-compositional occurrence phenomena (idioms of encoding), which often violate also syntactic rules: both a parser and a generator must «know» the idioms.

From the point of view of the lexical information used by different parsers and generators, B. Ingria divides the NLP systems into two general sorts of orientation:

Syntactically oriented approaches

These systems typically categorize their lexical items in terms of traditional parts of speech and perform detailed syntactic analysis of input sentences or texts.

Semantically oriented approaches

The systems perform fairly idiosyncratic syntactic analysis, devoting most of their efforts to the detailed semantic analysis of their input.

Ingria decided to consider only the syntactically oriented approaches. While information might be shared, with varying degrees of success between the syntactically oriented systems, there is less likelihood of sharing information between different semantically oriented systems.

Furthermore, the information required by semantically oriented systems does not very often relate in a direct or obvious way to any particular linguistic theory, and pose specific requirements on processing configurations and lexicon content and structure.

Small's theory of word expert parsing, for instance, places a heavy demand on the lexicon (see Small, 1981): the lexical entries (word experts) are complicated programs with routine structures, the specification of which requires detailed knowledge of the architecture of the parser, judgement of what constitutes relevant information and how to translate that procedurally, and readiness to bring in an arbitrary amount of more general, common world, knowledge (see Boguraev, 1987).

Obviously, we are very far from a generalization that allows to include this type of information in a general-purpose and extensive LDB.

Several NLP systems, falling within the large class of 'knowledge-based systems', require a significant amount of structured knowledge about the real word, or at least about a particular domain of discourse. The way in which knowledge based systems organize and mantain their knowledge bases differ widely. There is no firm consensus on what kind of

structures are best suited for capturing the knowledge useful for NLP.

Nonetheless, it is possible to observe a common theme in a large number of NLP systems.

A part of their knowledge is often represented using a scheme based on the general notions of frame-like concepts with slot-like descriptions, organized in an inheritance hierarchy[13].

The availability of large hierarchically structured networks of concepts is certainly a very useful source of data for the construction of the knowledge of these systems, no matter whether they represent semantic knowledge in terms of decomposition into markers taken from a set of primitives, formulae constructed from semantic primitives, frame-based structures, etc.

We describe below (section 4) how we use the definitions existing in our present machine-readable dictionaries as an aid for the construction of various semantic relations of this kind in the IMLDB.

We summarize here the classification schemata of syntactical information suggested by Ingria (which largely coincides with the schemata of Cumming), because the next step of our project will probably follow these schemata widely, and will require choices among different possible competing representation systems.

13. Cumming distinguishes two basic methods by which systems notate semantic classification of lexical items: feature systems and taxonomies.

Explicit taxonomical concept hierarchies represent (at least) relations of inclusion among word meanings. Taxonomies also represent the inheritence of properties from more general to less general concepts. Taxonomies are composed of concepts, each of which may be associated with one or more lexical entries; the lexicon is generally the place where the correspondence between concepts and words is stated. Melcuk's (1985) systems contain a richer specification of paradigmatic relations. In addition to hyponymy, he has functions of different kinds of synonyms, antonyms, words which have the same basic meaning but with the syntactic roles of the arguments interchanged, and many others. This richness is vital in a system whose primary goal is paraphrase or translation, as it gives the system a great deal of knowledge about what expressions can be considered semantically equivalent, something not available from simple taxonomy. Taxonomy very often contains also information relevant to selectional restrictions.

3.3. *Types of information*

Ingria and Cumming consider the following types of information as generally present in the lexicons of the computational systems revised:

a) *Syntactic categories*

Most lexicons agree in their assignment of lexical entries to the major categories (N, V, Adj, Adv, Prep), though they may differ as to the exact names of their categories.

However, the treatment of other categories, and even of the subcategories of the major categories, differs from one lexicon to the other. A very interesting example is represented by the difference in the treatment of *quantifiers*[14]. However, the problem is limited because those categories constitute a «close» subset of the lexicon, and a normalization may easily be reached through manual intervention.

b) *Contextual features*

This type of information may be defined in terms of the contexts in which a given lexical entry may occur. Following Chomsky (1965), they may be divided into two types:

Subcategorization

This specifies the complement structures, e.g. transitive verbs that occur with an object NP, etc.

Selectional restriction

This specifies the nature of the items that can appear in the complements or in a subject position (e.g. transitive verbs that restrict their direct object to be animate).

Some systems regard selectional information as more syntactic in nature, others as more semantic.

c) *Inherent features*

These cannot, or cannot easily, be reduced to a contextual

14. For example, *each* is coded as ART, DET, ADJ, QUANT, DETERMINER respectively, in the various systems reviewed by Ingria.

definition, e.g.: countable/non-countable, abstract, animate, human, etc.

Some are treated as semantic (animate, human), others as syntactic (e.g. non-countable).

3.4. *Types of representations*

The authors also pointed out a set of diversities in the representations adopted by the reviewed systems.

a) *Syntactic categories*

Simple symbols: each configuration of categories and subcategories is represented by a single code (e.g. the Kuno system (1963) has 133 different syntactic codes).

Complex symbols: each category and subcategory is represented by an independent code. Each lexical entry is cross-classified with respect to an array of categories and subcategories.

b) *Contextual features*

Subcategorization

Two main types are recognized:

a) using features which assign the entry to a specific class, whose syntactic behaviour is described elsewhere in the system;
b) specifying the number of slots and the types of elements that may appear as complements in each slot.

The second type has some operational advantages:

1) All kinds of subcategorizations and selectional restrictions which need to be stated as properties of particular lexical items can be easily handled without any special mechanism. Only the allowed patterns are listed in the lexicon. Any combination of complement types may be represented without having to decide beforehand on a particular inventory of possibilities.
2) All kinds of idioms and collocational restrictions can be potentially handled by specifying the exact wording of the lexical phrase.

3) An indefinitely large syntactic range may be «simulated» by treating as idioms syntactic constructions that may not be generated by the grammar.

The principle may be extended to the point in which the lexicon «takes over» most of the grammar. If the principle is brought to its utmost effect, the grammatical patterns tend to be represented only in the specification of the lexical items to which they apply.

There are also some disadvantages:
- the lexicon becomes larger;
- fewer phenomena are treated in a general way;
- length and difficulty of updating and additioning;
- properties of lexical items that may in fact be predictable (on the basis of other lexical properties) must be specified anyway.

In some ways, the difference between the two systems may be reduced by automatic procedures. E.g., a case frame can be mapped onto a feature representation, in which a given feature corresponds to a particular case pattern, or viceversa. E.g. the feature «transitive» can be mapped onto a case frame representation containing a direct object slot. The case frame explicit representation seems to allow more freedom. But, on the other hand, since features can be thought of as an indication of the inclusion into classes of lexical items, a single lexical feature may efficiently encode a range of possible case frames that tend to co-occur with particular types of words.

In other words, all the subcategorizational possibilities of a particular sense of a verb are taken to be predictable from a single feature representing its word-class membership.

Of course, in order to be able to take advantage of this type of generalization, one must have a detailed theory of the word classes of a language; and it is clear that a reasonably complete grammar must make reference to a very large set of such word classes.

This observation is, in a certain sense, the starting point for our feasibility study described below.

Selectional restrictions

Two principle ways of representation are recognized:
a) Semantic restrictions are explicitly associated to each slot of a lexical entry;
b) The restrictions are not represented directly in the lexicon, but are captured in another part of the system. E.g., in lexicons organized as semantic networks, hierarchically ordered concepts can be related one to the other by relations, which specify the semantic roles a concept has, and the relations with other concepts that represent possible fillers of each role.

3.5 *A «neutral» scheme of classification*

Encouraged by the results obtained by B. Ingria and S. Cumming, and also by the discussions which followed the workshop, we have promoted a working group which will involve outstanding representatives of the major current «linguistic schools». The group will investigate in detail the possibility of representing the linguistic information frequently used in parsers and generators (e.g. the major syntactic categories, subcategorization and complementation, verb classes, nominal taxonomies, etc.), in such a way that they can be reutilized in the following theoretical frameworks: government and binding, generalized phrase structure grammar, lexical functional grammar, relational grammar, systemic grammar, categorial grammar. This group will work on various languages. We shall start by examining in detail the treatment which the foregoing theories will assign to a representative sample of English and Italian verbs.

We eagerly look forward to the results. Various scenaries may be envisaged. For example, let us suppose we are describing the Italian verbs by using the criteria, tests, formal apparatus of a given theory.

At the end, we shall subdivide the Italian verbs into classes, regrouping in a class all the verbs with the same description. We consider as members of a same class those verbs which have received the same description. The intuition we wish to prove is that the aforementioned theories will classify the Italian verbs substantially in the same way.

Each theory will of course describe the syntactic behaviour of a class using its own formal and explicative apparatus.

If this is true, it would be possible to label the verbs of the IMLDB by distributing them into classes. The interface between the IMLDB, a given theory and its relevant computational systems would thus contain the descriptions of the syntactic behaviour of the different classes according to this theory.

A possible objection is that many classes could have few members, possibly one, because very few verbs would have exactly the same behaviour. If this were true, we could agree on a common set of «linguistic properties» to be named in the same way, and we could describe the verbs by specifying which properties they have and which they have not.

This is of course only an abstract scheme, and we should envisage a number of strategies for its correct application. However, we feel that this intuition is the same as stating that the properties taken into account by the different theories are in large part the same, although differently described and explained.

In this framework, it should be possible to reutilize the partial description of the Italian verbs so far produced by the different schools. In particular, the descriptions performed following the model of M. Gross (see Elia, 1984) should result very useful.

4. REUSABILITY OF LEXICOGRAPHIC INFORMATION

The preparation of a machine dictionary, and in particular of a LDB – as we have said – is one of the longest and most costly enterprises in the construction of NLP systems.

The number of researchers, asking whether and how the various sources of lexical information already in existence are reausable, is increasing. The possible sources considered are printed dictionaries in MRF, terminological data banks, machine readable dictionaries for linguistic researches, computational lexicons for NLP systems.

The nature and extension of the reutilization process depend on several factors: nature of the source, its reliability, level of formalization and generality of the data. The transfer of the data from the original source to a LDB is never straightforward.

Several details must be taken into account, which may require 'ad hoc' decisions and manual interventions.

Even in the simple case of the reutilization of a list of lemmas and their part of speech codes, different criteria can come into play for the identification of the lexical units (homonymy versus polysemy; autonomy of derived words; expressions written as single or separate graphical units; etc.) and for the number and the denomination of the morphosyntactic categories.

A more interesting case is the utilization of the grammatical information of the type associated to the lexical units in advanced learner's dictionaries such as the LDCOE and the OALD.

Various groups (Boguraev 1987; Chodorov *et alii* 1985; Alshawi 1986; Mejis 1986) are designing semiautomatic procedures which extract from the LDCOE the information on the syntactic properties of the lexical units which is requested by their linguistic theories and/or NLP systems.

Unfortunately, owing to the different status of teaching/learning Italian as a second language, these types of dictionaries do not exist for Italian.

As far as we know, computational lexicons of adequate size for NLP systems do not exist either[15].

Besides, very little if nothing at all has been done for the reutilization of a computational lexicon in a system other than that for which it was originally created.

With regard to our IMLDB it is necessary to question ourselves about the reutilization of two different kinds of available sources:

- the 'lexique-grammaire' prepared in MRF at the Salerno University in the framework of the M. Gross school;
- the semantic information (synonyms, antonyms, definitions) available in Italian MRDS.

The utilization of the Salerno 'lexique-grammaire' should not present any particular problems. The information is already formalized, and represented very clearly. After having defined, as we hope, with the aforementioned feasibility study, the

15. The MT system SYSTRAN developed at the EEC DGXIII includes an English-Italian version. The bilingual dictionary contains a subset of «general» Italian plus specialized terminology of food and agriculture. The reusability of these lexical data is a topic of current interest at the DGXIII.

formal representation of the «neutral» syntactical classification of the lexical units, we shall be able to map on this classification the information of the «lexique-grammaire». Furthermore, we actually believe that in any case it is convenient to insert this information as it is now in our IMLDB, in order to use it as documentary data in the feasibility study.

Some research groups (Calzolari 1984, Amsler 1987) are adopting substantially similar methods for the reusability of the semantic information conveyed by the definitions of the MRDs they have available.

The researches carried out by these groups are based on the acknowledgement that the logic-linguistic structures of some types of definitions, already consolidated in the traditional lexicographical practice, implicitly represent some relevant semantic relations between the units of a lexical system.

Those semantic relations are expressed by a limited number of expressions.

The most studied relation is that of hyponymy, which structures large subsets of lexical units in conceptual taxonomies. The groups try to recognize, through pattern-matching techniques, the definitions structured in 'genus proximum' and 'differentia specifica'. By exploiting the highly specific linguistic patterns which are used in the lexicographic tradition, they try to locate the term representing the 'genus'. All the lemmas whose definitions contain the same 'genus' term, are connected to the latter in a hyponymical relation. Various types of thesaurus-like relations are simulated in the lexicon through a structure of pointers among the lexical entries. The resulting conceptual texonomy is directly relevant for several uses.

The human user (a linguist, a lexicographer, an every-day user) may «browse» the lexicon not only through single words, but also through concepts and 'families' of concepts.

If – as is the case in our Institute – a textual DB is accessible with the 'help' of a MRD, the query of the user can be expanded by searching in the texts not only the word he has used in the query, but all the related terms in the taxonomy.

By using appropriate procedures, linguistic information (selectional restrictions, semantic features, etc.) associated to a superordinated term can be inherited by its hyperonyms and used by parsers and generators.

The conceptual taxonomy seems to be accepted as a common semantic tool by several state-of-the-art NLP systems.

A group of our Institute coordinated by N. Calzolari (1987) is working to recognize also other types of semantic relations implicitly represented in the definitions of the MRDs at our disposal (part of, instrument, derivations, said of..., etc.).

Of particular interest seems the research which tries to identify the semantic relation which connects the derivative to its base, expressed by the definition of derived words.

One of the goals is to implement formal rules which express the syntactic and semantic modifications induced by the derivation on the syntactic-semantic properties of the base.

5. THE ITALIAN MLDB IN A MULTILINGUAL ENVIRONMENT

We have just started a study intended to assess the feasibility of constructing a bilingual MLDB.

We conceive a bilingual MLDB as a complex structure consisting, essentially, of the following components:

a) A MLDB for language L1, structured according to the aforementioned principles;
b) A MLDB for language L2, structured according to the same or similar principles;
c) A bilingual «bridge» connecting the two monolingual MLDBs: i.e., a set of relations and conditions connecting their elements;
d) A textual DB containing, along with a reference corpus for L1 and L2, also a set of (so-called) «contrasted bilingual texts». By this expression we intend a structure including a text in one language, its translation into another, plus a set of cross-references explicitly indicating the relations of (translational) equivalence between the corresponding elements of the two texts[16];
e) A set of procedure and software tools for the access to the data, both for programs and human users, which will allow

16. The nature of these elements can depend on the procedures available to identify the correspondences. It is possible to range from the correspondence between contextual units labeled by the same reference number in the text and in its translation (verses, paragraphs, etc.), to the correspondence, identified by looking up a bilingual dictionary, among a number of non-ambiguous words, to the correspondence between syntagmatic units or syntactic structures identified or generated by a system of automatic translation.

the retrieval operations to start from any pertinent point in the entire structure.

In our opinion, a bilingual MLDB must be considered as:

- a source of data for contrastive and comparative linguistic studies;
- a component in a workstation for assistance to translation[17];
- a source of information for the construction of lexical components for NLP systems which require some type of transfer between languages, such as machine translation, or for the searching, in contrasted corpora, of possible translation equivalents for bilingual lexicography;
- a tool for computer-assisted language teaching and acquisition.

Our project aims at the construcion of a series of bilingual MLDBs connecting our Italian MLDB to other languages, starting with English. The reason for this choice is the availability, at our Institute, of the bilingual English-Italian, Italian-English Collins dictionary in MRF.

The first phase of the project is the design and the experimentation of procedures which will make it possible to reuse the information provided in the bilingual dictionary as a starting-point.

As a first step, a program will parse the text of each entry of the two sides, trying to identify the various types of information: the graphical word or the syntagm representing the entry; the different subentries; parts of speech; senses; usage labels; subject fields; collocations; synonyms; translational equivalents; examples; etc.

Each type of information will be transferred to a tagged field, and the various fields will be connected in a provisional DB structure. The first experiments show that a large amount of information was recognized and treated correctly. The remaining ambiguous cases will have to be solved by interactive procedures.

A second step will check the reversibility of the two sides of the dictionary. It has already appeared clearly that they present various differences. If A = <a1, a2.....ai> is the set of the

17. It is possible to range from a bilingual component included within a word-processing system, to an integrated system for professional translators, such as ALPS.

translation equivalents suggested for the entry *a*, we expect that *a* will be offered as a translation equivalent of each of the members of *A* in the other side of the dictionary. Some omissions may be due to the fully aware decision of the lexicographers[18], but very often they are the result of forgetfulness or of inconsistencies, and must be repaired. A very peculiar case is represented by idioms, collocations and, in general, phraseology and surface syntactic information.

Let the sequence of words *a+b* have as a translational equivalent the sequence *a1+b1*. The complete set of cross-references (under *a*, *b*, *a1*, *b1*) is not always given in the dictionary. For example, the information may be given in the lexical entry of *a* but not in the lexical entry of *b*.

It is our intention to normalize this kind of fragmentary and dispersed information, by integrating the data found in different places, and also by making it possible to access this same data from every pertinent entry for retrieval operations (Calzolari, Picchi, 1986).

One of the major problems will be to establish the appropriate links among the various «meanings» across the entries of the components of the bilingual MLDB.

The decomposition in «meaning units» of a lexical entry may be very different in the monolingual and in the bilingual dictionary as a result of the use of different criteria. In the bilingual dictionary the decomposition is usually performed on the basis of appropriate mapping to the target language, whereas in the monolingual dictionary the criteria are «internal» to the lexical system of the language. Also the discrimination of the «meaning units» is often represented in a different way. For example, the bilingual dictionaries use very few definitions.

An interactive procedure will take into account at least the following operations:

a) Mapping the «meaning units» of an entry in the monolingual and bilingual dictionary. The bilingual entry may be more or less complex than the corresponding monolingual entry. As a result of the combination of the two decompositions, both the monolingual and the bilingual entry may be enriched.

If the monolingual Italian MLDB will be connected not only to English, but also to other languages, we shall have to decide

18. See the examples given by Byrd *et al.*, 1987.

whether or not to maintain separate the various bilingual MLDBs, or whether to combine them together in a multilingual structure.

In the latter case, we shall have to face the problem of the additioning, in the entries of the Italian monolingual MLDB, of the different decompositions required by the correspondences with the various languages. Up until now, we have not yet taken any decisions in this respect. As a matter of fact, the decision will have to be taken in the light of the experience gained by contrasting the IMLDB with other languages.

b) The information used to discriminate the different «meaning units» in the bilingual dictionaries, have the function of giving to the reader the elements needed for the transition to the second language. For various reasons, including the saving of space, this information is often expressed in an ambiguous or incomplete way. The effort made to interpret, disambiguate and translate it into formalized rules, by using interactive procedures, will not only make easier the mapping between the bilingual and the monolingual entries, but will also help to formulate conditions and constraints to be used in the lexical transfer of MT systems.

At the very end of this process, the information extracted from the bilingual MRD will be transformed into links mapping the «meaning units» of the two bilingual MLDBs, and into constraints formally expressing the conditions of application to the contexts of the texts to be translated.

As we have already said before, the «meaning units» of our Italian MLDBs are in the process of being connected within a network of various syntactic and semantic relations: synonymy, hyponymy, etc. Our interactive systems allow the user to «navigate» through the network starting from any entry and selecting the desired type of relations.

We hope that other groups, working with similar methodologies in their own languages, will be willing to collaborate with us. In this case, we could cooperate to connect our two similarly structured MLDBs.

In the resulting bilingual MLDB, the user will access, starting from an entry in one of the two monolingual MLDBs, or from an element in the translation conditions, a structure formed by three networks: the two monolingual networks of syntactic-semantic relations, and the «bilingual» network

linking, with «mapping relations» and «translation conditions», the two monolingual MLDBs.

The possibilities which are open to the human users seem to be obvious. For example, the thesaurus-type organization generated, in a monolingual MLDB, by representing the hyponymy relations, makes it possible to access the bilingual MLDB not only through a single word, but also through a concept (Calzolari, Picchi, 1987).

From a general point of view, the translator will be enabled to consider the correspondences not only from a word to its translation equivalents, but also from a «family of words» (e.g. a semantic field) in L1 to a «family of words» in L2 (see McArthur, 1987).

This organization can also be useful to expand the information concerning the translation conditions («semantic indicators» in the Collins dictionary) provided by a bilingual dictionary, to be reused in NLP systems involving two languages.

The «semantic indicator» is often a single word chosen to represent a «family» of words: e.g. a synonym, a hyponym, a typical subject or object of a verb, a typical noun of which an adjective can be predicated, etc. The monolingual MLDB can be used to expand this information, given as a single word, to the whole set of words to which it actually refers (Calzolari, Zampolli, 1987).

This expansion obviously implies the previous disambiguation of the relation between the word given as «semantic indicator» and the set of words it represents. It must be stressed that we are only at the beginning of this research[19].

6. POSSIBLE COOPERATIONS

A MLDB is called like this as it is addressed to different users. Therefore it is obvious that we should ask ourselves whether and which categories of users can cooperate to its creation. It is not only a matter of sharing, if possible, the effort and the cost of its construction, and to be able to use already available data

19. The same, as far as we know, is true for the only other groups working on bilingual MLDBs. The work of this group, which is coordinated by Roy Byrd at the IBM Scientific Center at Yorktown Heights, is described in Byrd *et al.*, 1987.

and information, but also to benefit from the know-how and the specific competences of each category. This complementarity is confirmed by a number of recent experiences which we have promoted. It is evident that great expertise in the sector of computer science is necessary if one is to design structure and programs for the access and handling of MLDBs. These systems must typically take into account a very large amount of data, the need for new types of information to be progressively added to the MLDB, the different types of access requested by human users and programs, the different supports and environments (main-frame, PC, CD, etc.).

It is evident that linguistic theories provide methods and criteria for the choice, the definition, the structuring, the representation of the information to be included at the different linguistic levels, and that computational linguistics must define the relations between this information and the different components of the NLP systems which use them: parsers, generators, transfers, retrieval, etc.

The role of both the traditional and the computational lexicographers continues to be much more controversial.

We have recently addressed particular attention to this problem, both from the point of view of the motivations which lexicographers may have in participating in the construction of LDBs, and of the competences and data by which they may give a contribution.

A LDB can be used as a starting point in the editing process of a lexicographic product.

Systems acting as an aid to editing, used particularly in England and in Denmark, are already available.

The system mainly consists in an editor, which enables the lexicographer to insert the data through the keyboard of a PC or a terminal, gradually «filling» in the tagged fields by which the structure of the lexical entry is formed.

The system provides the lexicographer with a defined «by default» structure, but the lexicographer can define a particular structure suitable to its lexicographic product, in the initialization phase of the system. This structure is presented as a «menu» on the screen, prompting the lexicographer to supply the information field by field.

The system organizes the information introduced by the lexicographer in the form of a database.

Thus the lexicographer can at any time trace decisions and data which have been treated previously, by using the research and «browsing» functions «supplied» by the system.

Furthermore, a program generates automatically, from the LDB, the text ready for photocomposition, taking care of the handling of the lexicographic metaformat, variation of fonts, abbreviations, cross-references, etc.[20]

The use of a LDB in the editing phase does not only allow a more economic and coherent insertion of the data (and up until now this has actually been the main motivation for the publishing houses), but it also has various consequences at the qualitative level.

The lexicographer is often limited by space and time constraints. Owing to its incremental character, a LDB, in a certain sense, can free lexicographers from these constraints.

The lexicographer can fully use his knowledge and competence at their best, inserting, also at different times, the treasure of lexical information he has available.

The reactions of the users can eventually by taken into account, as a kind of monitoring feedback.

A variety of information can be included in the LDB. The editing phase will have to select it appropriately and convert it into the form required by different types of traditional lexicographic products: let us consider, for example, the grammatical information typically present only in the learner's dictionary, or the difference between the coding and decoding type of information in bilingual dictionaries.

Furthermore, lexicographers are well aware that LDBs offer new markets to publishing houses, in particular with regard to the different activities of the so-called «language industry»: office automation, spelling checkers, machine-assisted translation, speech processing, etc.

20. The commercially available existing editorial systems are not directly intended to assist the lexicographer in the extraction of citations from a reference corpus and in their insertion in the structure of the lexical entry. The utilization of a reference corpus is essential, for example, for the preparation of historical dictionaries. Therefore, a more complete editorial tool is necessary for «scientific» lexicography. We are working at the ILC (Picchi, 1986) to construct a lexicographical workstation, making it possible for a lexicographer to interact with a corpus, preexisting dictionaries, lexical entries already prepared by his coworkers, etc.

Besides, the collaboration with linguists and computational linguists provides the lexicographer with knowledges which make it possible for him to improve his product in different ways.

For example, the procedures by which computational linguists process the definitions in order to extract different semantic relations, and the resulting subsets of structured lexical units, can improve considerably for the common user the access to the lexical information. The improvement is even more evident with regard to the consultation of a bilingual dictionary.

Furthermore, the possibility of retrieving information explicitly or implicitly present in a dictionary, using as multiple access points all the elements of the lexical entries, eventually combined in various ways, opens new kinds of dictionary usage to the academic world (linguists, philologists, historians, etc.), thus enlarging the potential market of some lexicographical products. The organization of a historical dictionary as a LDB offers the possibility of searching, for example, all the words which appear for the first time before a certain date, or of retrieving all the citations extracted from a given text, or of identifying all the words marked with a given field label and appearing in a given period, etc. Benbow (1987) provides interesting examples and a deep discussion of the relationships between the lexicographers, the publishing houses, the new technologies, the future developments of the electronic publishing market.

Byrd (1987) discusses the copyright problems, which are likely to arise between the publishing houses, which possess the lexicographical data and know-how, and the industries which aim at reusing the data for the development of new tools for the 'information society' and the language industries.

We also want to underline a new aspect of the relationships between theoretical linguists and lexicographers, which emerged as a result of the Grosseto workshop, and which encouraged a closer cooperation.

A study conducted by a lexicographer (B.T. Atkins, Collins Publishers) and two linguists (J. Kegl, Princeton University and B. Levin, MIT) has shown that the construction of dictionaries «should take advantage of theoretical linguistic work on the organization of the lexicon» (Atkins *et al.*, 1986), ensuring

explicitness, comprehensiveness and coherence across the whole lexicon. On the other hand, the examples collected in the dictionary entry can offer new evidence and new linguistic facts that the linguists must explain.

For example, the lexicographers recognize relevant linguistic properties of the lexical units, but do not often spell them out explicitly. They try to represent them to the users, either through examples or by variously combining definitions, examples, syntactic codes. In such cases, in other words, the lexicographers recognize that a general pattern is involved, but nowhere is this pattern explicitly indicated.

A typical case are the syntactical properties of a verb which reflect essential aspects of its meaning. It will be impossible for the lexicographer to take full advantage of his knowledge of the language, until this type of information is made explicit.

«Making the implicit knowledge encoded in a dictionary explicit is only possible in the context of a theory of the lexical organization. Linguists can contribute to work in lexicography by providing such a tehory. Theoretically guided linguistic investigations into lexical organization can be used as a starting point for this task». (Atkins *et al.*).

7. THE IMLDB IN THE CONTEXT OF INTERNATIONAL ACTIVITIES IN THE FIELD OF COMPUTATIONAL LEXICOLOGY AND LEXICOGRAPHY[21]

7.1. In the past two years, a number of initiatives at the international level have clearly shown that the creation of multifunctional lexical data bases, the reusability of existing lexicographical resources, and cooperation among research groups, industry, and publishing houses are considered to be issues of primary importance in an effective program of research and development on multilingual lexicology and lexicography[22].

21. In this chapter several passages are extracted from A. Zampolli, D. Walker (1986). I take this occasion to thank Don Walker for his pretious cooperation.

22. The following activities constitute examples:
a. Workshop on «Automating the Lexicon: Research and Practice in a Multilingual Environment», Marina di Grosseto, May, 1986.
b. Workshop on «The Lexical Entry», New York City, July, 1986.

This strong interest in lexicology and lexicography is due to the following factors, among others:

a. Theoretical developments within linguistics are placing increasing emphasis on the lexical component. It is proving to be a central source of semantic as well as syntactic information.
b. Demonstrations of the feasibility of applications of natural language processing are creating demands for large-scale systems in industry and in national and supranational organizations. For these systems to be practical they must deal with tens of thousands of lexical items.
c. The effort required to create comprehensive dictionaries for these purposes is substantial; it may prove to be the most costly and time consuming task in such developments. Currently, each system is building its own lexicon, and there is increasing recognition that the duplication of efforts is enormously costly. However, differences in the organization and content of these lexicons make it difficult or impossible to share linguistically relevant information across systems.
d. The computational linguistics community is becoming increasingly conscious of the extensive resources contained in published dictionaries and explorations are underway to determine how that information in machine-readable form can be exploited to expedite system development.
e. Publishers are beginning to realize the potential of their dictionaries for commercial purposes. They are recognizing the value of establishing lexical data bases from which a variety of dictionaries can be derived.

They are also becoming aware of the breakdown of the distinction between different reference works (dictionaries, fact books, encyclopedias).

f. Increasing communication among lexicologists, lexicographers, linguists, computational linguists, publishers, and commercial natural language processing software developers has led to a heightened awareness of common objectives and the complementarity of skills and knowledge.
g. Initial experiments have given support to the idea that it may be possible to construct «neutral lexicons» that can be shared, with different theories selecting relevant linguistic information through an appropriate interface.

7.2. The prospects for furthering work on multilingual lexicology and lexicography are good.

A basis for coordination has been established.

A network is being created to link interested people and organizations under the direction of Don Walker and Antonio Zampolli.

Working groups are being organized to address a large class of problem areas.

Meetings are being arranged under a variety of auspices (ALLC, EURALEX, ACL, AILA).

It is time now to consider how governmental organizations can further these efforts. The framework research program of the EEC 1987-1991 explicitly foresees as an objective the creation and reusability of lexical resources. The need for international cooperation has increased for the fact that projects for the creation of MLDBs have already started or are about to start in different countries, often organized at a national level.

In the United Kingdom, as a part of the Alvey Projects, a general purpose lexical system is being constructed within a larger project designed to provide an integrated system for morphological and syntactic analysis of a substantial subset of English (Boguraev *et al.*, 1987).

In Germany, at the University of Bonn, comparison of existing German MRDs has been conducted, and an activity is planned for converting existing LDBs into a lexical knowledge

c. Panel on «The Lexicon in a Multilingual Environment», COLING '86, Bonn, August, 1986.
d. Ad hoc Working Group on «Computational Lexicology and Lexicography», ESF.
e. Specialist Working Group on «Dictionaries and the Computer», EURALEX, Zurich, September, 1986.
f. Conference on «Standardization in Lexicography», ESF, Saabrücken, October, 1986.
g. Conference on «Advances in Lexicology», Centre for the New OED, Waterloo (Canada), November, 1986.
h. Session on «Words and World Representations», TINLAP 3, Theoretical Issues in Natural Language Processing, Las Cruces (New Mexico), January, 1987.
i. Special triple issue on the Lexicon, Computational Linguistics, 1987.
j. Workshop planned on «The Lexicon in Theoretical and Computational Perspectives», Summer Linguistic Institute, Stanford, July, 1987.
k. Summer School planned on «Computational Lexicography and Lexicology», Pisa, Summer 1988.

base, with particular attention to semantic dimension (Lenders, 1987).

In the Netherlands, a project for the realization of a very large repository of lexical data is planned at the Nijmegen Max Plank Institute, in connection with psycholinguistic research (Schreuder, 1987, CELEX n. 1).

The Government in Japan, in solicitation from and in cooperation with several major companies, has set up an Institute for the creation of a very large Japanese monolingual and Japanese-English bilingual LDBs (Nagao, 1987)[23].

Our project for a MILDB is connected in many ways with the aforementioned activities[24].

We cooperate in various forms to promote and improve the international cooperation, because we think that the dimension of the enterprise is such that no single research group can provide the necessary resources, know-how and data[25].

We consider the cooperation indispensable both for the design and the implementation of methods and tools, and for collecting and sharing data in a multilingual lexical network.

In particular, the following problems, in our opinion, urgently require international cooperation[26]:

23. A useful survey of automated lexical resources in Europe is provided by S. Warwick (1987).

24. We have to cope of course with the peculiarity of the Italian situation. For example: the existing lexical data in MRF are the result of activities carried on in the technical context of the late sixties (no MRDs available, no interactive methods); limited availability of MRDs, and in particular of learned dictionaries of Italian.

25. The research and development in the field of MLDBs require, as we have shown, interdisciplinary cooperation. They require competence and consideration in a variety of topics: linguistics, and in particular morphology, syntax, semantics; computational linguistics: parser, generation, transfer; computer science: data base management methodology and techniques; new technology: to identify and develop their potential in storing and accessing lexical information; lexicography: lexicographic practice must be considered as a source of know-how and data; artificial intelligence: relationship with knowledge bases and knowledge representation; cognitive science: reusing of LDB in psycholinguistic research; Language Industries Applications: office automation, information retrieval, machine (assisted) translation; speech processing; education of first and second language; development of reading and writing skills, communicative aids to the disabled; electronic publishing; etc.

26. The following recommendations were formulated at the Grosseto workshop and approved at the New York Workshop («The lexical Entry,

– Establishing procedures for creating MLDBs from the information contained both implicitly and explicitly in traditional dictionaries that are in machine-readable form.

July, 1986). «The Workshop has clearly identified, among a range of academic and industrial research and development groups, publishers, and commercial firms that market lexical products, a convergence of interests that would motivate the establishment of large computational lexical resources intended for shared activities. The complementarity of monolingual and multilingual concerns was consistently stressed.

1) Create and maintain registries of machine-readable dictionaries and related resources, lexical databases, text corpora, bibliographic references and the corresponding documents, and human resources; where appropriate, establish designated repositories for materials that are available for distribution.

2) Establish terminological conventions for working with lexical resources that can be shared by groups working with computers as well as those using more traditional approaches.

3) Clarify the copyright issues associated with the various lexical resources and establish a framework that supports the broadest distribution of these materials to groups of relevant users.

4) Organize a lexical data entry group with responsibilities for identifying lexical materials that should exist in machine-readable form, for determining a standard format or set of formats in which they should be represented, and for arranging to have them coded and made available through the repositories.

5) Establish a communication network, progressively computerized, that can link together the Workshop participants and other interested people and groups to allow sharing information about new and continuing developments and to provide a forum for examining critical issues.

6) Establish more general communication channels through professional societies, their journals and newsletters, and presentations at regular conferences (e.g., the Association for Computational Linguistics (ACL), EURALEX, the International Association for Applied Linguistics (AILA), the Association for Literary and Linguistic Computing (ALLC), and the Association for Computers and the Humanities (ACH)).

7) Arrange special meetings that promote further communication, coordination, and cooperation for both general and specialized interest groups focused on selected topics.

8) Study the work of lexicographers to model their behaviour, incorporating the results in knowledge-based systems that support lexicographic activities.

9) Study how people interact with standard and electronic dictionaries and lexical databases to determine the most effective procedure for human/machine interaction.

10) Develop lexical and lexicographic workstations embodying resources, data, and tools that directly support lexicological and lexicographic activities.

11) Investigate new technologies and products that could be incorporated into such workstations; correspondingly, identify design characteristics that would facilitate working with lexical materials and try to motivate their development as products.

- Developing computational tools for working more efficiently with lexical and lexicographic data, and providing «workstation» environments within which these tools can be used by lexicologists and lexicographers.

12) Support «internship» and «sabbatical» links that allow people in various disciplines to work closely with each other on project activities.
13) Develop curricula, courses, texts, and manuals for lexicology and lexicography that will further interdisciplinary understanding and that can be used in a variety of education and training contexts.
14) Compare and contrast lexical information, particularly in the form of «lexical entries», as reflected in logical and linguistic theories, computational linguistic systems, machine-readable dictionaries, translation activities, and lexicographic practice in order to identify dimensions of similarities and differences; based on those dimensions, create a metaformat that subsumes the structures of the various types of information to be included, and that can be used both as a reference frame for evaluation and exchanges and as a model of a computerized metalexicon from which lexicons for different research and applications may be derived.
15) Establish procedures for converting the contents of machine-readable dictionaries, text corpora, and other lexical data into formats appropriate for a range of computational needs.
16) Apply frequency measures to gather systematic and representative synchronic and diachonic data on a broad range of language variables in text corpora.
17) Determine whether dictionaries can be designed so that they can be used in both human and machine environments.
18) Convince publishers to begin saving the photocomposition tapes of books, journals, and other published materials and to make them available for research.
19) Establish project designs and patterns of cooperation that promote sharing of data, tools, and human resources (particularly scarce ones) among academic and industrial research and development groups, publishers, and commercial firms that market lexical products.
20) Create linguistic databases from existing and newly produced sources that embody machine-readable dictionaries and large text corpora (thousands of millions of words), and create tools that make it possible to explore their relationships systematically.
21) Create lexical databases and explore their utility for supporting the creation of general and specialized dictionaries, monolingual dictionaries, encoding and decoding dictionaries.
22) Establish procedures for deriving monolingual and bilingual lexical and lexicographic material from text corpora; of particular interest are strategies for identifying phrases, synonyms, hyponyms, and other classes of relationships automatically.
23) Establish large collections of «evaluated» paired and aggregated translations reflecting bilingual and multilingual sources, and develop procedures for exploring and exploiting their correspondences.

- Exploring the possibility of creating neutral LDBs that can be used independently of differences in linguistic theories and computational and applicational frameworks.
- Studying the interaction between LDBs and large text files in both monolingual and multilingual contexts to determine the most effective ways to utilize relationships among lexical elements.
- Furthering effective communication and collaboration among linguists, computational linguists, lexicologists, lexicographers, publishers, and commercial developers with sharing resources.
- Addressing the copyright problem as it relates to the research

24) Develop methodologies for interrelating monolingual and bilingual dictionaries; explore the possibility of combining technical monolingual dictionaries with bilingual general dictionaries to create technical bilingual dictionaries.
25) Establish lexical indices for determining and representing stylistic features, subject-matter codes, and other sociolinguistic parameters; create procedures to include them in machine-readable dictionaries and for using them for lexicological and lexicographic research.
26) Study the use of lexical information by children and conduct experiments to determine what kinds of lexical resources would be most effective for educational purposes.
27) Establish a range of dictionaries that reflect needs for specialized information or nonstandard modes of interaction (e.g., handicapped users), clarifying their similarities and differences as well as the feasibility of deriving them from standard dictionaries or from each other.
28) Develop new programming languages that support the coordinated manipulation of strings (text sequences) and structures (taxonomies, frames, and logical relationships).
29) Develop new database designs that allow storing, accessing and interrelating (at detailed feature levels) both the form and content of multibillion word text files.
30) Study pictures, tables, diagrams, and other illustrative material and develop workstation tools able to process and to relate them systematically to corresponding machine-readable dictionary entries and passages in text corpora.
31) Develop a theory of pictorial reference that will facilitate relating lexical and semantic information to images.
32) Extend and modify the typology of traditional dictionaries and lexical tools and resources so that it applies to materials that are now or will be in machine-readable form and to the emerging uses of these materials.
33) Determine how to incorporate the experiences of traditional academic, humanistic, and classical studies for lexical and lexicographic research». (Walker, 1987, *Introduction*).

community in order to increase the accessibility of critical data resources.

- Developing methodologies to connect «neutral» monolingual MLDBs of different languages, possibly using bilingual MRDs as a starting point; identifying and establishing a network of national DBs, possibly with the sponsorship of international, intergovernmental organizations.
- Establishing and developing relationships with other categories of «reference» works and knowledge DBs, such as encyclopedias, terminological DBs, etc.

REFERENCES

ALSHAWI H., Processing dictionary definitions with phrasal pattern hierarchies, University of Cambridge Computer Laboratory, Cambridge, 1986 (forthcoming).

ALSHAWI H., BOGURAEV B., BRISCOE T., Towards a dictionary support environment for real-time parsing, in *Proceedings of the 2nd European Conference of the Association for Computational Linguistics*, Geneva, 1985.

AMSLER R.A., The Structure of the Merriam-Webster Pocket Dictionary, Ph. D. Thesis, Department of Computer Sciences, University of Texas, Austin, 1980.

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

AMSLER R.A., Machine-readable dictionaries, in M. E. WILLIAMS (ed.) *Annual Review of Information Science and Technology (ARIST)*, ASIS Vol. 19, 1984, pp. 161-209.

AMSLER R.A., Deriving lexical knowledge base entries from existing machine-readable information sources, in D. WALKER et al. (eds.), 1987.

ATKINS B.T., KEGL J., LEVIN B., Explicit and Implicit Information in Dictionaries, in *Proceedings of the Conference on Advances in Lexicology*, Waterloo, 1986.

BENBOW T., WEINER E., Machine-readable Dictionaries for the General Public, in D. WALKER et al. (eds.), 1987.

BOGURAEV B.K., Machine-readable Dictionaries and Computational Linguistics Research, in D. WALKER et al. (eds.) 1987.

BOGURAEV B., BRISCOE T., Large lexicons for natural language processing: exploiting the grammar coding system of LDCOE. Unpublished manuscript, Cambridge, 1986.

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

BUSA R. (ed.), De Lexico Electronico Latino, *Calcolo*, V (1968) Suppl. 2.

BUSA R., ZAMPOLLI A., Centre pour l'automation de l'analyse linguistique (CAAL), Gallarate, in *Les machines dans la linguistique*, Prague, 1986, pp. 25-34.

BYRD R.J., Word formation in natural language processing systems, in *Proceedings of the 8th International Joint Conference on Artificial Intelligence*, Karlsruhe, 1983, pp. 704-706.

BYRD R., Dictionary Systems for Office Practice, in D. WALKER et al. (eds.) 1987.

BYRD R.J., CALZOLARI N., CHODOROW M.S., KLAVANS J.L., NEFF M.S., RIZK O.A., Tools and Methods for Computational Lexicology, IBM T.J. Watson Research Center, unpublished, 1986.

CALZOLARI N., Towards the organization of lexical definitions on a data base structure, in E. Hajicova (ed.), *Proceedings of COLING 82*, Prague, 1982, pp. 61-64.

CALZOLARI N., Semantic links and the dictionary, in *Proceedings of the Sixth International Conference on Computers and the Humanities*, Raleigh (North Carolina), 1983a, pp. 47-50.

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

CALZOLARI N., Detecting patterns in a lexical data base, in *Proceedings of Coling 84*, Stanford University (Calif.), 1984, pp. 170-173.

CALZOLARI N., CECCOTTI M.L., Organizing a Large-Scale Lexical Database, in *Actes du Congrès International Informatique et Sciences Humaines, Liège: L.A.S.L.A., 18-21 Novembre*, 1981, pp. 155-163.

CALZOLARI N., CECCOTTI M.L., ROVENTINI A., Documentazione sui tre nastri contenenti il DMI. Pisa, ILC-DMI (2) 1984.

CALZOLARI N., PICCHI E., The machine readable dictionary as a powerful tool for consulting large textual archives, in L. CORTI *Automatic Processing of Art History Data and Documents*, (ed.), Pisa, Scuola Normale Superiore, 1984, pp. 275-288.

CALZOLARI N., PICCHI E., A Project for a Bilingual Lexical Database System, in *Proceedings of the Conference on Advances in Lexicology*, Waterloo, 1986.

CALZOLARI N., ZAMPOLLI A., From monolingual to bilingual automated lexicons: is there a continuum?, 1987 (in print).

CHODOROV M., BYRD R., HEIDORN G., Extracting semantic hierarchies from a large on-line dictionary, in *Proceedings of the 23rd Annual Meeting of the ACL*, Chicago, 1985, pp. 299-304.

CHOMSKY N., *Aspects of the Theory of Syntax*, Cambridge (Mass.), 1965.

CIGNONI L., PETERS C., ROSSI S., *European Science Foundation: Survey of Lexicographical Projects*, Pisa, 1983.

COLLINS, *Concise Italian Dictionary*, Collins, London, 1985.

CUMMING S., Design of a Master Lexicon, ISI/RR-85-163, 1986.

CUMMING S., The Lexicon in Text Generation, in D. WALKER et al. (eds.) 1987.

DE TOLLENAERE F., *Nieuwe Wegen in de Lexicologie*, Amsterdam, 1973.

DURO A., ZAMPOLLI A., Analisi lessicali mediante elaboratori elettronici, in ACCADEMIA NAZIONALE DEI LINCEI, *Atti del Convegno «L'automazione elettronica e le sue implicazioni scientifiche, tecniche, sociali»*, Roma, 1968, pp. 121-739.

ELIA A., *Le verbe Italien*, Fasano di Puglia, 1984.

ENGEL G., MADSEN B.N., From dictionary to database, in R.R.K. HARTMANN (ed.), 1984, pp. 339-343.

EVENS M.W., LITOWITZ B.E., MARKOWITZ J.A., SMITH R.N., WERNER O., Lexical-Semantic Relations: A Comparative Survey, Edmonton, Alberta: Linguistic Research Inc., 1980.

FERRARI G., Procédés et méthodes pour la création d'un algorithme de flexion de la langue italienne, in A. ZAMPOLLI (ed.), 1973, pp. 97-100.

GOETSCHALCKX J., ROLLING L. (eds.), *Lexicography in the Electronic Age*, Amsterdam: North-Holland, 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 siècle (1789-1960)*, in A. ZAMPOLLI, A. CAPPELLI (eds.), 1983, pp. 119-144.

GROSS M., *Méthodes en Syntaxe*, Paris, 1975.

GRUPPO DI PISA, Il dizionario di Macchina dell'Italiano, in D. GAMBARARA, F. LO PIPARO, G. RUGGIERO, *Linguaggi e Formalizzazioni*, Roma, Bulzoni, 1979, pp. 683-707.

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

HARTMANN R.R.K. (ed.), *LEXETER '83 Proceedings*, Tübingen, 1984.

HULTIN N.C., LOGAN H.M., The New Oxford English Dictionary Project at Waterloo, *Dictionaries*, 6 (1984), pp. 182-198.

IANNUCCI E.J., Sense discriminations and translation complements in bilingual dictionaries, *Journal of the Dictionary Society of North America*, 7 (1985), pp. 57-65.

INGRIA R., Lexical Information for Parsing Systems: Points of Convergence and Divergence, in D. WALKER et al. (eds.), 1987.

JOSSELSON H.H., The lexicon: a system of matrices of lexical units and their properties, in *Preprints of COLING 1969* (Stockolm – Reprints), 1969.

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

LENDERS W., Data sources for a German lexical knowledge base, in D. WALKER et al. (eds.), 1987.

MADSEN H., Compulexis. A universal dictionary system, Abstracts of ZURILEX 86, 1986.

MADSEN B.N., Danish projects within the field of computational linguistics, in D. WALKER et al. (eds.), 1987.

MC ARTHUR T., The present and future use of machine-readable dictionaries in education, in D. WALKER et al. (eds.), 1987.

MC NAUGHT J., Computational lexicography and computational linguistics, unpublished paper, 1986.

MEIJS W., Lexical organization from three different angles, *Journal of the ALLC*, 13 (1986) 1.

MELC'HUK I., ZHOLKOVSKY A.K., *Explanatory Combinatorial Dictionary of Modern Russian*, Vienna, 1984.

MICHIELS A., NOEL J., Approaches to thesaurus production, in *Proceedings of COLING 82*, Amsterdam, North-Holland, 1982, pp. 227-232.

MOULIN A., JANSEN J., MICHIELS A., Computer exploitation of LDOCE's grammatical codes, in *Conference on Survey of English Usage*, London, 1985.

NAGAO M., Activities involving electronic lexicons in Japan, in D. WALKER et al. (ed.), 1987.

NEUHAUS J., The morphological structure of Shakespeare's vocabulary, Abstracts of ZURILEX 86, 1986.

OLNEY J., RAMSEY D., From machine-readable dictionaries to a lexicon tester: Progress, plans, and an offer, *Computer studies in the humanities and verbal behaviour*, III (1972), pp. 213-220.

PICCHI E., Workstation lessicografica. Unpublished paper, Pisa, 1986.

PICCHI E., CALZOLARI N., Textual perspectives through an automatized lexicon, in *Proceedings of the XII International ALLC Conference*, Nice, 1985.

QUEMADA B., Bases de données informatisées et dictionnaires, *Lexique*, 2 (1982), pp. 101-120.

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

SCHREUDER R., Using lexical databases in Psycholinguistics Research, in D. WALKER et al. (eds.), 1987.

SMALL S., Viewing word expert parsing as a linguistic theory, in *IJCAI 7*, (1981), pp. 70-76.

TEUBERT W., Setting up a lexicographical data-base for German, in R.R.K. HARTMANN (ed.), 1984, pp. 425-429.

URDANG, L., A lexicographer's adventures in computing, *Dictionaries*, 6 (1985), pp. 150-165.

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

WALKER D., ZAMPOLLI A., CALZOLARI N., (eds.), *Automating the Lexicon: Research and Practice in a Multilingual Environment*, 1987 (in print).

WIEGAND H., Metalexicographical reflections of the conception of a lexicographical data bank for contemporary German, (in print), 1986.

ZAMPOLLI A., Projet d'un dictionnaire de machine, in R. BUSA (ed.), 1968, pp. 109-126.

ZAMPOLLI A., (Ed.) *Linguistica Matematica e Calcolatori*, Firenze, 1973.

ZAMPOLLI A., L'Automatisation de la recherche lexicologique: état actuel et tendences nouvelles, *META*, XVIII (1973), 1-2, pp. 101-136.

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

ZAMPOLLI A., BROGNA D., Procedura elettronica di spoglio, in *Concordanza dei Grammatici Latini*, Torino, 1979, pp. 35-51.

ZAMPOLLI A., CALZOLARI N. (eds.), *Computational and Mathematical Linguistics*, Vol. 1, Firenze, 1977.

ZAMPOLLI A., CALZOLARI N. (eds.), *Computational and Mathematical Linguistics*, Vol. 2, Firenze, 1980.

ZAMPOLLI A., CALZOLARI N., Computational Lexicography and Lexicology, *AILA Bulletin* (1985) pp. 59-78.

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

ZAMPOLLI A., WALKER D., Multilingual lexicology and lexicography: new directions. Paper presented at the CG12-TSC of the EEC, 1986.

ZIMMERMANN H.H., Multifunctional Dictionaries, in A. ZAMPOLLI, A. CAPPELLI (eds.), 1983, pp. 279-288.